Advancing surrogate modelling for sustainable building design

by

Paul W. Westermann
M.Sc. MEng, ETH Zurich, 2017
B.Sc. MEng, ETH Zurich, 2015

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Civil Engineering

University of Victoria

Advancing surrogate modelling for sustainable building design

by

Paul W. Westermann
M.Sc. MEng, ETH Zurich, 2017
B.Sc. MEng, ETH Zurich, 2015

Supervisory Committee

Dr. Ralph Evins, Supervisor
(Department of Civil Engineering)

Dr. David Bristow, Departmental Member
(Department of Civil Engineering)

Dr. Nishant Mehta, Outside Member
(Department of Computer Science)

External Examiner
Dr. Bryony DuPont
(College of Engineering, Oregon State University)

Abstract

Building design processes are dynamic and complex. The context of a building project is manifold and depends on the cultural context, climatic conditions and personal design preferences. Many stakeholders may be involved in deciding between a large space of possible designs defined by a set of influential design parameters.
Building performance simulation is the state-of-the-art way to provide estimates of the energy and environmental performance of various design alternatives. However, setting up a simulation model can be labour intensive and evaluating it can be computationally costly. As a consequence, building simulations often occur towards the end of the design process instead of being an active component in design processes.
This observation and the growing availability of machine learning algorithms as an aid to exploring analytical problems has lead to the development of surrogate models. The idea of surrogate models is to learn from a high-fidelity counterpart, here a building simulation model, by emulating the simulation outputs given the simulation inputs. The key advantage is their computational efficiency. They can produce performance estimates for hundreds of thousands of building designs within seconds. This has great potential to innovate the field. Instead of only being able to assess a few specific designs, entire regions of the design space can be explored, or instantaneous feedback on the sustainability of building can be given to architects during design sessions.

This PhD thesis aims to advance the young field of building energy simulation surrogate models. It contributes by: (a) deriving Bayesian surrogate models that are aware of their uncertainties and can warn of large approximation errors; (b) deriving surrogate models that can process large weather data (≈150'000 inputs) and estimate the associated impact on building performance; (c) calibrating a simulation model via fast iterations of surrogate models, and (d) benchmarking the use of surrogate-based calibration against other approaches.

Acknowledgements

I would like to express my thank to my supervisor, Dr. Ralph Evins, for giving me the opportunity to join him and the young Energy and Cities group in beautiful Victoria, for his guidance, and for his support to accommodate any of my plans. A special thanks goes to the rapidly growing team, which always had an open ear for my research ideas and brought in valuable input for my work. Especially I would like to thank Gaby Baasch, David Rulff, Matthias Welzel, David Fritzsche, Kevin Cant, Theo Christiaanse, and Gaëlle Faure. I also owe many thanks to Professor Arno Schlüter and the A/S research group at the Institute for Technology in Architecture, ETH Zurich, for hosting me during my visits in Zurich. Finally, I would like to express great gratitude to limitless support off-campus. Thanks to Chris Wood, Miguel Alvarez, Toby Cotton, Aurélien Liné, Claire Remington, the UVIC Field Hockey team and all the others. Thanks to my sisters and parents. Thank you, Fredi.

Paul W. Westermann
University of Victoria, July 2020

Table of Contents

List of Publications

The research conducted throughout the course of my PhD studies has been published in high-ranked, international scientific journals or conference proceedings. In total I have contributed with five journal papers, of which three were accepted and two are submitted or ready for submission, and five conference papers, of which three have been published and one awaiting publication in the proceedings of the eSIM 2020 conference, which has been postponed due to the COVID-19 crisis.
The papers are sorted into two groups based on their relevance to the core research objectives of this thesis. They are listed in order of their appearance in the thesis; secondary publications are included in the appendix.

Primary publications

P1: *Westermann, Paul; and Evins, Ralph. "Surrogate modelling for sustainable building design - A review." Energy and Buildings 198 (2019): 170-186.*

PW conducted the data collection, analysed and compiled the findings and wrote the paper. RE revised the manuscript.

P2: *Westermann, Paul; Rulff, David; Cant, Kevin; Faure, Gaelle; and Evins, Ralph. "Net-Zero Navigator: A platform for interactive net-zero building design using surrogate modelling. Submitted to eSIM 2020 (2020).*

PW conducted the surrogate modelling, analysed the results and wrote the majority of the paper. DR developed the building simulation model. KC developed the building simulation model. GF wrote and revised parts of the manuscript. RE leads the NZN project, contributed to the concepts and revised the manuscript.

P3: *Westermann, Paul; and Evins, Ralph. "Bayesian modelling for uncertainty-aware surrogate models." Submitted to Journal of Advanced Engineering Informatics.*

PW conducted the data collection, analysed and compiled the findings and wrote the paper. RE revised the manuscript.

P4: *Westermann, Paul; and Evins, Ralph. Adaptive Sampling For Building Simulation Surrogate Model Derivation Using The LOLA - Voronoi Algorithm. Proceedings of the BS Rome 2019, (2019).*

PW conducted the data collection, analysed and compiled the findings and wrote the paper. RE revised the manuscript.

P5: *Westermann, Paul; Welzel, Matthias; and Evins, Ralph. "Using a deep temporal convolutional network as a building energy surrogate model that spans multiple climate zones." Accepted to Journal of Applied Energy.*

PW conducted the data collection, analysed and compiled the findings and wrote the paper. MW conducted data collection, analysed and compiled the findings. RE revised the manuscript.

P6: *Westermann, Paul; Deb, Chirag; Schlueter, Arno; and Evins, Ralph. "Unsupervised learning of energy signatures to identify the heating system and building type using smart meter data." Applied Energy 264 (2020): 114715.*

PW conducted the data collection, analysed and compiled the findings and wrote the paper. CD supervised and revised the manuscript. AS provided resources and revised the manuscript. RE revised the manuscript.

P7: *Baasch, Gaby; Westermann, Paul; and Evins, Ralph. "Advanced Techniques for Learning Quantitative Building Properties from Sensor Data: An Empirical Perspective on Competing Paradigms." Draft ready for submission to Energy and Buildings (2020).*

GB generated the synthetic data set, conducted the calibration of lumped parameter models, trained the black-box models and wrote the manuscript. PW supported the data generation, conducted the surrogate-based calibration approaches, and wrote the manuscript. RE revised the manuscript.

Secondary publications

P1: *Westermann, Paul; David, Nigel; and Evins, Ralph. "Machine Learning Recommendations for Control of Complex Building Systems Using Weather Forecasts."*

Proceedings of eSim 2018 (2018).

PW conducted the data analysis and compiled the findings and wrote the paper. ND provided measurement data. RE revised the manuscript.

P2: *Bowley, Wesley; Westermann, Paul; and Evins, Ralph. "Using Multiple Linear Regression to Estimate Building Retrofit Energy Reductions." Proceedings of eSim 2018 (2018).*

WB collected all data and wrote the majority of the paper. PW ran the regression analysis, made the figures and wrote parts of the paper. RE revised the manuscript.

P3: *Westermann, Paul; Braun, Johanna; Murphy, Eamon; Grieco, Joel; and Evins, Ralph. " Insight Into Predictive Models: On The Joint Use Of Clustering And Classification By Association (CBA) On Building Time Series." Proceedings of the BS Rome 2019, (2019).*

PW analysed the data, and wrote the paper. JB, EM, JG collected the data and analysed the data. RE revised the manuscript.

Key contributions

The key contribution of this thesis is the advancement of fast machine learning surrogate models to become a second pillar in sustainable building design alongside common physics-based performance simulation. We lay the technical foundations to robust, uncertainty-aware surrogate models that generalize over a large scope of design tasks that architect and building designers may face.
The thesis is divided into two parts. First, we focus on deriving more robust surrogate models where we integrate powerful methods from machine learning literature into our domain. In the second part, we take advantage of computational efficiency of surrogate models to efficiently calibrate building performance models to measured sensor data. This is an essential prior step to well-informed retrofit design for existing buildings.
The main contributions are listed below:

Part I

Collection of relevant literature [P1]: The field of surrogate modelling is young. As a first contribution we provided the first collection of relevant studies that used surrogate modelling to facilitate building design. We extracted major achievements and research trends, and conceptualized surrogate models augmenting simulation tools to form a two-system-based building performance assessment tool. Similar to a human brain, a fast, intuitive surrogate model (System 1) can be used to analyse frequently occurring design problems, and a high-fidelity, physics-based model can be used to assess more complex designs which integrate new technologies (System 2). The following research was grounded on that literature review.

Surrogate models in use [P2]: A tool is being developed that hosts surrogate models on a web server, such that it can be actively used by building designers and

architects for fast, interactive design of net-zero energy buildings. In the study, we train a surrogate model that covers a large number of design parameters (inputs) and performance metrics (outputs), which pushes the current state of research.

Uncertainty aware surrogate models [P3]: Surrogate models are a statistical approximation of a high-fidelity model. Although they achieve high emulation accuracy on average, large errors can occur. We transfer novel findings from the machine learning literature, i.e. Bayesian deep learning approaches, to our domain. As a result, our surrogate models are capable of quantifying the uncertainty associated with the approximation process. This may be crucial for a robust use of surrogates in the future, and can also be used to train them more efficiently, by actively picking training samples in regions of the design space where high uncertainty was observed [P4].

Generalization of surrogate models [P5]: One fundamental criticism of surrogate models is that they are only valid to the narrow scope of design problems that they have been trained for. Expensive retraining of the surrogate model is necessary if the design task slightly changes. Until this study, a generalized surrogate model that is trained to cover different climate impacts was lacking in the literature. The climate is directly linked to a specific location so, a surrogate model was location-bound. We derived a deep temporal convolutional network that can process the exact same weather inputs as the high-fidelity simulation model, such that we could significantly improve the generalizability of a trained surrogate model to multiple design problems.

Part II

Energy signatures for building characterization [P6]: The inputs to a calibration process are measured building sensor data and a raw, uncalibrated model. Smart meter data is the most prevalent source of measured building data, in particular in Canada [11], and it is suitable to calibrate a large stock of buildings. Automatically determining a suitable structure of an uncalibrated model for a large number of buildings remains challenging.

We developed a method that integrates building domain knowledge with data driven algorithms. It extracts qualitative building properties from the same

smart meter data, which subsequently are used to set up the uncalibrated model. We use the concept of energy signatures, a scatter plot with outside air temperature on the x-axis and electricity consumption on the y-axis, which condenses each building's electricity use into one highly informative graph. They allow us to automatically infer the installed heating system type and building type without requiring any additional data. This was shown on two smart meter data sets covering 889 buildings. Afterwards, the calibration process can begin.

Surrogate-based calibration benchmarking [P7]: In this study, surrogate modelling was compared to other calibration approaches. To allow detailed analysis of the performance and to design informative experiments, synthetic building measurement data was generated using parametric building simulation runs. We showed that surrogate model-based calibration outperforms many other approaches in estimating the building's heat loss coefficient, a metric that quantifies whole building energy efficiency. Future work will inform how well surrogate-calibration works in the real world environment.

Chapter 1

Introduction

1.1 Sustainable building design for the clean energy transition

According to the International Energy Agency (IEA), the building sector accounted for 28% of global carbon emissions in 2019, reaching an all-time high of 10 $GtCO_{2,e}$ [12]. Current efforts decrease energy use per floor area (0.5% - 1% per year since 2010) but are not enough to outweigh the ever growing building stock (2.5% per year since 2010). The IEA recommends significantly increasing quality and coverage of building energy codes, fostering retrofits, ramping up heat pump installations, and improving air conditioning efficiency.

Architects and building designers are responsible for transferring these high level paradigms to the level of individual projects. This is a challenging endeavour as each real estate project is unique, differing in climate, built environment, occupant behaviour and design preferences of the owners. An optimal sustainability strategy for one building is not necessarily suitable for another. Furthermore, the preferences of the many stakeholders involved in a project can differ strongly.

1.2 Building performance simulation

Given the large set of variables in a sustainable building design task, the design process is often supported by building performance simulation (BPS) software to predict and assess the performance of a building design [10]. BPS software is based on a steadily growing knowledge of building physics and used to model the thermal loads of a building given material properties, the setup of heating, cooling, ventilation and air-conditioning (HVAC) systems, the occupant behaviour and comfort preferences, the external climate conditions, the indoor daylight conditions, hygrothermal effects and other influences. EnergyPlus is the BPS program used throughout this thesis [3].

While accuracy in the outputs is desirable, the major goal of BPS is to increase problem understanding, where design parameter sensitivity analysis and performance uncertainty analysis are fundamental aspects. It is widely known that there is an expected *performance gap* between simulated and measured buildling performance, caused by mistakes by the modellers, by mistakes in the construction phase, and by the probabilistic nature of building loads (e.g. occupant behaviour) [4].
While this thesis focusses on the use of BPS for architects and building designers to design better buildings or assess retrofit options, it can also be applied for high-level policy design, or by HVAC engineers to optimize the operation of a building.

1.2.1 Towards an *exploration* of sustainable building designs

In the last two decades, a large set of computational methods have been developed to augment stand-alone BPS. In particular, the use of heuristic or gradient-based optimization approaches which operate over the BPS software have received a lot of

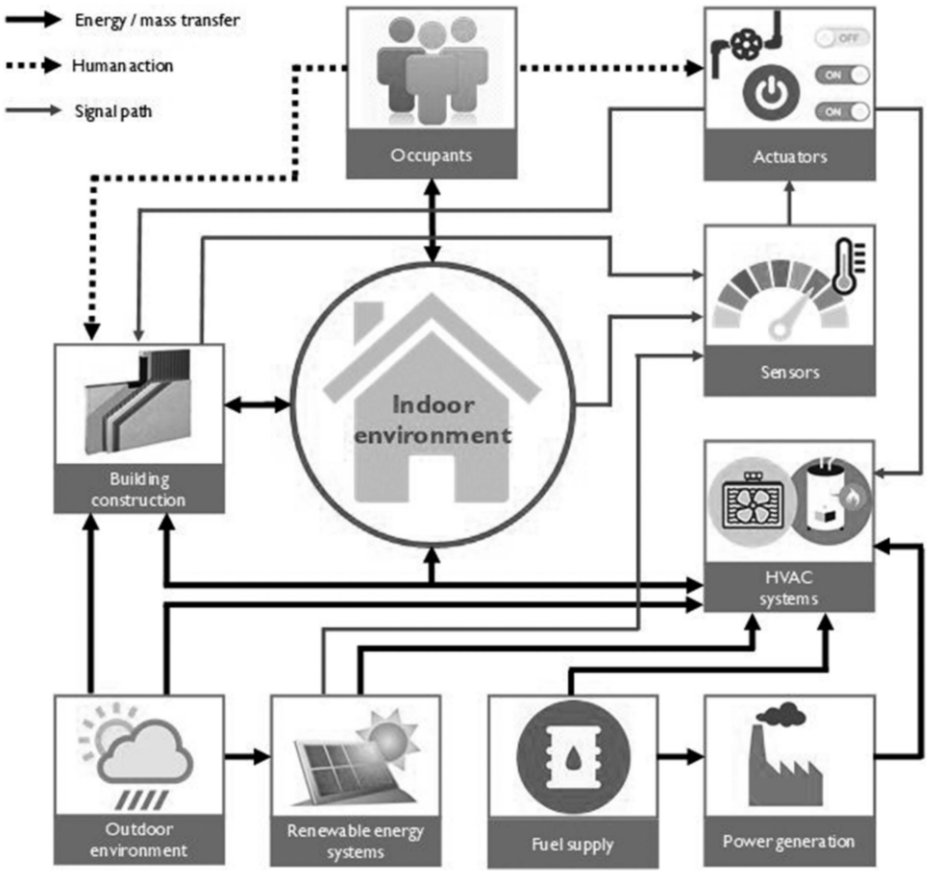


Figure 1.1: **The modelling scope of typical building performance simulation programs**, from [10].

attention in the past [5]. However, it was found that optimization is often not robust towards rapid changes at the conceptual design stage caused by uncertainty in the project requirements, or that it does not suit the need for architectural freedom by the designers [1].

Instead, methods allowing interactive *exploration* of design alternatives have recently been favoured over automated tools to find a particular optimal design [20]. Currently parametric modelling is used for this purpose. The idea is to automatically run a large number of simulations covering a multitude of design options. The simulation inputs and outputs are stored in a database such that the architect has immediate access to performance estimates without interacting with complex simulation software or waiting for a simulation run to finish. The data can also be incorporated into interactive user interfaces, e.g. parallel coordinate plots [18], that can guide the designer through the space of possible design options [24].

In a recent empirical study, the use of interactive BPS-based tools was shown to be popular among architects and also enabled them to produce better performing designs compared to conventional approaches [1].

1.2.2 Challenges

The use of interactive tools circumvents the hurdles of the BPS process, in which architects and project developers hire a BPS expert who collects all relevant project information, sets up the simulation model and conducts the simulation runs. This can be tedious and pushes BPS towards the end of the design process to ensure compliance to performance targets or to building codes. Authors have referred to this as the problem of BPS being an *elaborative* tool rather than a *proactive* element in design processes [23].

Using parametric models has been the first step to tackle these challenges - with significant drawbacks. First, the design parameter combinations must be selected prior to the design space analysis. When the studied building is large and complex the runtime of a BPS constrains the selection process to relatively few samples ($\approx$ 100). This is particularly limiting, as building design problems are commonly characterized by a large number of design parameters which span a large, multi-modal design space [21][27].

A coarse set of parameter combinations restricts the freedom of architects and also may not capture high performing design alternatives. One way around this is to use powerful computational hardware to increase simulation speed, as already available in some BPS software products [9], and the use of Design-of-Experiment methods (DoE) [6] to pick samples efficiently throughout the space of options. However, studies have shown that the required number of samples to provide a detailed view on the design space is large. For example, 5000 parametric simulations did not include any design alternative after the architect imposed filters on certain design parameters [19].

These limitations of parametric analysis on the one side, and the strength of machine learning methods to quickly and automatically extract understanding of correlations in data on the other, has brought the field of surrogate modelling to innovate traditional BPS [26][21].

1.3 Surrogate modelling for BPS

The idea of surrogate modelling is to train a machine learning model on BPS input and output data (see Figure 1.2, left). The approximate statistical method is evalua-

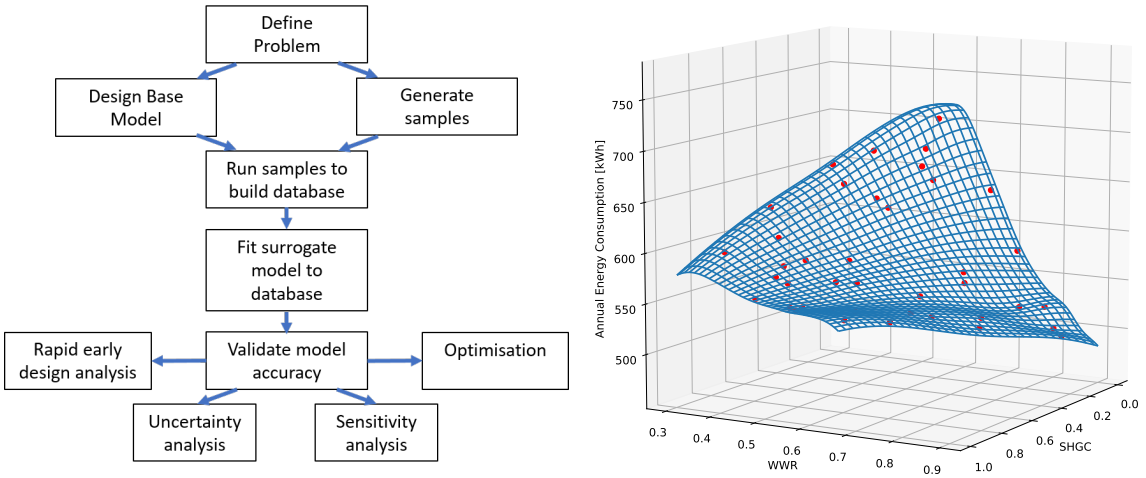


Figure 1.2: **Building surrogate modelling.** On the left, the general surrogate modelling process is showcased. Details can be found in Chapter 2. On the right, we show an example of a low dimensional design problem. The red dots depict the training data, and the blue grid shows the surrogate evaluated at the grid's nodes.

ted much faster than the BPS model counterpart, which allows to produce thousands of performance estimates within seconds, as shown in Figure 1.2 (right) by the evaluation of a surrogate model on a tight grid of points. In comparison to parametric runs, the parameters (here the window-to-wall ratio, WWR, and the window's solar heat gain coefficient, SHGC) can be chosen freely.

1.3.1 Simulating, fast and slow

The core contribution of this thesis is to integrate BPS with surrogate models which is similar to producing building performance estimates with both a fast and a slow system. We use the slow high-fidelity model to synthesise a large set of physical laws

explaining the building design performance estimates. It is considered a white-box model, where we know the underlying rational. The laws are scientific generalizations and are not bound to a certain design parameter range. The fast surrogate model, which represents the second system, is very different. It relies on statistical learning, which is bound to the domain of the training data. When using a machine learning model as surrogate, an algorithm determines the model structure making the model hard to interpret (black-box model).

The characteristics of the two systems are reminiscient of Kahneman's definition of how the brain forms thoughts, which he published in his book "Thinking, fast and slow" [14]. He found that humans use two thought processes; one is fast and one is slow. The fast system is non-logical, effortless, intuitive and emotion-driven. The slow system is more energy-intensive, based on rationales, more logical and we consciously perceive the thinking process. Kahneman points out that the two systems are concurrent and even the fast process can be used for complex tasks, e.g. a chess player is able to play speed chess after he trained reading books and playing matches over several years. Determining which system to use is crucial, and wrong decisions can cause mistakes.

This analogy inspired this work, and will be referred to throughout the thesis. For example, the challenge of determining when to use a surrogate model and when to refer to an actual simulation run was explored in the research below (see Chapter 4).

1.4 Research questions

In the following we formulate specific research objectives to advance the integration of BPS with surrogate modelling. The objectives are split into two parts, *Part I* focusses on improving the use of surrogate models to augment BPS and is the primary focus of this thesis, and *Part II* uses surrogate modelling to extract building properties from building sensor measurement data through model calibration. All objectives are based on a thorough literature review, which is presented below.

Part I

Research Question 1.1: How can surrogate models be more robust and is there a way to quantify their uncertainty in emulation?

Surrogate models inherently introduce error to building performance estimates. First comparative studies have shown that they are very accurate on average [21][26], however, this does not ensure that the surrogate model performs well for the part of the design space the architect is most interested in. The objective behind this research question is to identify these inaccuries and to quantify confidence intervals. This potentially also allows us to *hybridize* the two systems, i.e. the slow high-fidelity BPS software and fast surrogate model, to jointly produce building design performance estimates as fast as possible within a specified certainty band (see Section 1.3.1). This may include that the surrogate model may actively learn, by targeting simulation runs that it is most uncertain about.

Research Question 1.2: How can surrogate models generalize to more building design problems and more locations, which differ in climate?

In existing studies surrogate models are derived to approximate a specific building simulation model that is designed for a specific project. Hence the sampling and training of a surrogate has to be repeated if the project changes. Some authors compartmentalized surrogate modelling into multiple tasks, e.g. to specifically emulate the heat flux through walls, floors and ceilings [7]. This envisions that the compartmentalized surrogate models can be combined to approximate any geometry. Among other limitations, this approach still binds the surrogate to the specific climate it has been trained for. We aim to find representations of climate data as input to a surrogate such that it can quantify the impact of different climates on building performance. This will make surrogates much more reusable and readily applicable without the need for sampling and training prior to application.

Part II

Research Question 2.1: How can we extract fundamental building mechanical system properties from smart meter data prior to surrogate-based model calibration?

In the previous section, we introduced the challenge of finding a suitable base model for a large number of buildings. Essential parameters for a base model include building location and climate conditions, primary building usage, building geometry and mechanical system configurations. Only with satisfactory prior knowledge of these properties is it possible to derive a physically meaningful quantitative calibration of parameters like the envelope R-value, heating system efficiency, infiltration rate, or heat recovery efficiency.

Some of these underlying properties are easier to collect than others, e.g. occupancy behaviour can be extracted from load profiles and building location and geometry can be collected using satellite data. Currently, we are lacking an approach to derive

which mechanical system type is installed. An automated smart-meter-based estimate is developed in this thesis.

Research Question 2.2: How does the performance of surrogate-based building model calibration compare to other methods to extract thermal building properties?

Having accurate knowledge of the building at hand still does not guarantee that a bottom-up surrogate-based building characteristic estimate is the best option to collect quantitative building properties prior to designing the building retrofit. We benchmark surrogate-based calibration against other bottom-up approaches and top down deep learning methods [2].

1.5 Structure of the thesis

The structure of the thesis chronologically follows the outline given in the research questions. In Chapter 2, we present a thorough literature review. It is the first publication summarizing significant works on surrogate modelling for sustainable building design. Part I of the research questions follows. We start be giving a detailed example on the use of surrogate models for building design (Chapter 3). Afterwards, we tackle the research questions of Part I in Chapters 4 and 5. The research questions of Part II are addressed in Chapter 6. Additional contributions that cover the use of machine learning for related fields like building controls, or retrofit analysis, are found in the Appendix.

Chapter 2

Literature Review

The motivation of surrogate modelling is driven by the ability to provide instantaneous feedback to architects at the early design stage, but their evaluation speed makes them attractive for a variety of design analysis tasks. This includes design optimization, global sensitivity analysis, and uncertainty analysis.
Quickly mapping design parameters to building performance metrics can also be useful for determining parameters of an existing building. Either by using an optimization approach or a Bayesian paradigm, we can use the surrogate model to calibrate building parameters of existing buildings. In comparison to other calibration methods, surrogate based calibration is fast while retaining the link to detailed building performance simulation models (white-box models), whereas in other approaches rather simplified physics-based models (grey-box models) are used. Detailed BPS models allow us a larger flexibility when implementing retrofit scenarios post-calibration in comparison to simplified models.

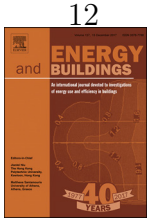
In the following we review the use of surrogate models for the design of new buildings. That review article does not feature a section on surrogate-based model calibration. The associated literature is summarized in Section 6.1.

Energy & Buildings 198 (2019) 170–186

Contents lists available at ScienceDirect

Energy & Buildings

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



Surrogate modelling for sustainable building design – A review

Paul Westermann*, Ralph Evins

Energy and Sustainable Cities Group Department of Civil Engineering University of Victoria 3800 Finnerty Road, Victoria BC, Canada

ARTICLE INFO

Article history:
Received 24 January 2019
Revised 15 April 2019
Accepted 26 May 2019
Available online 29 May 2019

Keywords:
Sustainable building design
Building performance simulation
Surrogate model
Meta-model
Early design
Uncertainty analysis
Sensitivity analysis
Building design optimisation

ABSTRACT

Statistical models can be used as surrogates of detailed simulation models. Their key advantage is that they are evaluated at low computational cost which can remove computational barriers in building performance simulation. This comprehensive review discusses significant publications in sustainable building design research where surrogate modelling was applied.

First, we familiarize the reader with the field and begin by explaining the use of surrogate modelling for building design with regard to applications in the conceptual design stage, for sensitivity and uncertainty analysis, and for building design optimisation. This is complemented with practical instructions on the steps required to derive a surrogate model. Next, publications in the field are discussed and significant methodological findings highlighted. We have aggregated 57 studies in a comprehensive table with details on objective, sampling strategy and surrogate model type. Based on the literature major research trends are extracted and useful practical aspects outlined.

As surrogate modelling may contribute to many sustainable building design problems, this review summarizes and aggregates past successes, and serves as practical guide to make surrogate modelling accessible for future researchers.

1. Introduction

The Intergovernmental Panel on Climate Change (IPCC) recognizes the potential for the current building stock to stabilize or reduce its global energy use by mid-century [1]. The high performance of current building technologies and understanding of how to integrate them, make energy efficient buildings and retrofits also economically viable.

However, the building sector transforms slowly. The International Energy Agency (IEA) observed that it lags behind in the clean-energy transition as defined in the Paris Agreement [2]. One key challenge faced by the sector is that each building and retrofit is unique and has to be customized due to varying purpose, location and cultural context. Taking into account that the existing building stock of 150 billion square meters will grow by an annual rate of 3.7 billion square meters until 2026 [3] and that buildings are currently designed in a largely individual fashion by architects and engineers, facilitating and automating the design processes will be crucial to the spread of sustainable buildings.

Recent advances in machine learning paired with growing data availability are pushing the automation of analytical problems like sustainable building design [4,5]. Three fundamental types of data exist in the building domain:

(a) Building sensor data (e.g. smart meters, internet of things (IoT) sensors, building management systems)
(b) Building stock data (e.g. annual energy demand and floor area for a large set of buildings)
(c) Building simulation data (stored results of building simulation)

The first two types are particularly useful for optimising building operation [6,7], designing building-specific retrofit options [8] *(a)*, or for conducting energy mapping and building performance benchmarking in a certain geographic area covered by the building stock data *(b)* [9].

Both types of data are composed of historical observations on already existing buildings. Statistical prediction models trained on that data clearly may not be accurate for new building technologies or unique design concepts. Hence, building simulation relying on physical laws remains crucial for the design of new buildings. Its validity is not bound to observations, but instead any new design, retrofit option or building technology can be modelled.

Abbreviations: BPS, Building Performance Simulation; GP, Gaussian Process model; ANN, artificial neural network; MARS, multivariate regression splines; SVM, support vector machine; PCE, polynomial chaos expansion; RF, random forest; RBF, radial basis function; LSTM, long-short term memory network; LHS, latin hypercube sampling; DoE, design of experiments; iid, independent and ideally distributed; SA, sensitivity analysis; UA, uncertainty analysis; BDO, building design optimisation.

* Corresponding author.
E-mail addresses: pwestermann@uvic.ca (P. Westermann), revins@uvic.ca (R. Evins).

https://doi.org/10.1016/j.enbuild.2019.05.057
0378-7788/

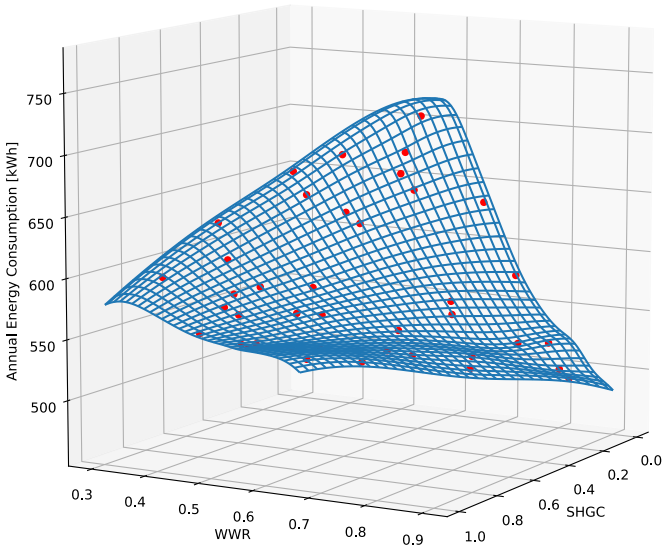


Fig. 1. Example of the application of surrogate modelling for sustainable building design evaluation. This surrogate estimates annual energy consumption based on window-to-wall ratio (WWR) and solar heat gain coefficient (SHGC). It was fitted to previously collected simulation samples (red dots) and was then evaluated at a finer resolution (every intersection of the blue mesh). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

However, current building simulation software has high computational cost and setting up a building model is time intensive [10]. Needed architects and designers do not fully integrate it into their daily work [11]. Surrogate models [12–14], or meta-models, are promising to provide building performance assessment which is physical knowledge based but much faster than simulation-based design analysis [15].

The idea of surrogate modelling is to emulate an expensive high-fidelity model, in this case a building simulation model, using a statistical model. The surrogate is trained on a small set of simulation in- and output data (c). Once it is validated to approximate the detailed simulation model well enough, it can be used to almost instantly predict outcomes of the high-fidelity simulation given an appropriate set of building design information.

In this work we are largely concerned with surrogates that predict aggregated design metrics (e.g. annual energy use) rather than detailed time series (e.g. hourly energy use). The process is illustrated in Fig. 1 for a problem with two inputs and one output. Here a Gaussian process model was trained to predict annual energy demand based on window-to-wall ratio and solar-heat-gain coefficient. In general (deep) artificial neural networks, support vector machines, or radial-basis function networks are common choices [16].

It is important to stress that the models studied in this review are trained on synthetic data. They are only accurate within the limitations of the simulation program and the input data used. The error induced by the simulation program as well as the modelling error of the surrogate must be balanced against the significant benefits that surrogate models bring. Both causes of errors must be addressed together, as the more accurate the simulation, the more accurate the surrogate must be to capture its behaviour. We assume that the reader is familiar with the possible errors in building simulation [17] and therefore take synthetic data as sufficient.

The review is structured as follows:

In the first two sections we familiarize the reader with the field. Section 2 covers the background on the use of surrogate modelling for the conceptual design stage (2.1), sensitivity and uncertainty analysis (2.2–2.2) and design optimization (2.4). Section 3 gives details on the steps to derive a surrogate model split into problem definition (3.1), simulation base model implementation (3.2), sampling (3.3) and surrogate model fitting (3.4). This is complemented with a list of existing surrogate modelling tools (3.5).

The reviewed literature is presented in Sections 4 and 5. First, we outline the scope of this review and refer to other reviews in related fields like energy demand forecasting (4.1). After giving an overview of the research topics (4.2) and the applied methods found (4.2.1–4.2.3), the papers are discussed thoroughly grouped by the four use cases as introduced in Section 2. We summarize findings drawn from the literature in a comprehensive list in Section 5 covering research trends and practical aspects of surrogate model fitting.

Finally, we conclude and give suggestions for future research in Section 6.

2. Surrogate models for building design

Based on existing literature (see Table 2), four stages of the building design process are found to significantly benefit from surrogate modelling:

1. Conceptual design stage
2. Sensitivity analysis
3. Uncertainty analysis
4. Optimisation

In the following section, each stage is explained in detail and the associated use of surrogate modelling explained. The section is summarized in Table 1.

2.1. Conceptual design stage

The early design or conceptual design stage happens at the very beginning of the building design process. At this point, the design is most flexible. Many parameters are roughly determined (e.g. building geometry and system types), which have a substantial impact on the final environmental and economic performance of the building [18].

Architects derive design concepts together with other stakeholders in a dynamic process. This can involve quick and drastic design changes [19] where the whole concept of the building is

Table 1
Summary on the use of surrogate models for building performance design analysis.

Analysis type	Use of surrogate
Conceptual design	• Fast feedback for design concepts; design space exploration • Fast analysis of impact of design decisions on design variability
Sensitivity analysis	• Fast variance-based global SA
Uncertainty analysis	• Fast building performance probability distribution derivation • (Model calibration)[a]
Optimisation	• Acceleration of optimisation process • Enabling gradient-based optimisation

[a] Beyond the scope of this review.

Table 2
Considered literature and its properties.

				Subject						Surrogate Model Type												Sampling Methodology				
Index (#)	Reference	First author	Year	Early Design	Sensitivity Analysis	Uncertainty Quantification	Optimisation	Comparison of meta-modelling techniques	Building Stock Analysis	ANN	RBF	Gaussian Processes	Linear Regression incl. (LASSO, stepwise, nth order terms)	MARS	SVM	Random Forest	Ensembles	Others	Favoured model (only if comparison done	# Inputs	# Outputs	Multiple climate zones?	Sampling Method	Sampling Method: adaptive (a) / static (s)	Simulation tool	# Samples (Total)	Checked impact of number of samples?
1	[77]	Gratia	2002	x									x							8	3		Doelhart matrix	s	TRNSYS	1980	
2	[26]	Ritter	2015	x					x				x						Polynomial	7	5		factorial	s	EnergyPlus	2187	
3	[25]	Geyer	2014	x									x							8	4		factorial	s	EnergyPlus	4096	x
4	[78]	Edwards	2017	x				x		x			x							156	80-90		Markov order	s	EnergyPlus	8*10^-6	
5	[79]	Yi	2015	x									x							8	2		L18 Taguchi	s	EnergyPlus	18	
6	[81]	Maltais	2017	x	x		x						x							15	2		manual	s	Daysim	1900	x
7	[84]	Korolija	2012	x									x					x		2	3		full factorial	s	EnergyPlus	23040	
8	[86]	Catalina	2013	x									x							3	1	x	full factorial	s	TRNSYS	8748	
9	[87]	Hygh	2012	x	x								x						Stepwise LR	28	3	x	full factorial	s	EnergyPlus	20000	
10	[88]	Lam	2010	x									x						LR	12	1	x	manual	s	DOE-2.1	1021	
11	[89]	Jaffal	2009	x									x							11	1	L12/18	Taguchi, FCC35, BB188, D68, D136, D80, D160	s	TRNSYS	60-230	x
12	[90]	Rackes	2016	x											x					38	1	x	Sobol	s	EnergyPlus	16k - 50k	
13	[91]	Romani	2015	x				x					x						Polynomial	11	2	x	D-optimal	s	TRNSYS	115-210	
14	[92]	Ascione	2017	x					x	x									ANN	29	3	x	LHS	s	EnergyPlus	500	
15	[93]	Jaffal	2017	x									x						Polynomial	10	4	x	factorial, Box-Behnken	s	TRNSYS	120	
16	[94]	Singaravel	2018	x						x									ANN	15	1	x	LHS	s	EnergyPlus	1001	
17	[23]	Ostergard	2017	x	x								x					x	Others	9	3		Sobol	s	Be10	6000	
18	[24]	Basbagill	2014	x		x														27	2		LHS, orthogonal array	s	eQuest	7623	
19	[97]	Hopfe	2012		x	x	x					x								10	2		optimality criterion	a	VA114	n/a	
20	[98]	Kim	2016		x	x		x				x							GPR	47	1		LHS	s	EnergyPlus	15	x
21	[36]	Tian	2011		x								x						SRRC	9	3		manual	s	EnergyPlus	2400	
22	[99]	Das	2014		x												x		ensemble	9	2	x	LHS	s	CONTAM	500	
23	[32]	Rivalin	2018		x	x		x				x						x	PCE	24	1		manual	s	TRNSYS	200	
24	[29]	Eisenhower	2011		x	x												x		1009	10		pseudo Monte Carlo	s	EneryPlus	5000	
25	[101]	Tsanas	2012		x			x					x			x			RF	8	2		full factorial	s	Ecotect	768	
26	[102]	de Wilde	2010		x	x							x	x					MARS	10	3		LHS	s	EnergyPlus	2000	
27	[103]	Tian	2015		x			x				x	x	x	x			x	ensemble	28	2		LHS	s	IDEAS (Modelica)	600	x
28	[105]	Chen	2017		x			x					x	x					MARS	10	3		LHS	s	EnergyPlus	5000	x
29	[21]	Hester	2017	x		x							x						Stepwise LR	20	1		full factorial	s	EnergyPlus	20000	
30	[108]	Papadopoulos	2016			x							x						MLR	6	1		LHS	s	EnergyPlus	1200	
31	[113]	Wong	2010	x			x			x									ANN	9	4		manual	s	EnergyPlus	11315	
32	[114]	Magnier	2010				x			x										20	5		LHS	s	TRNSYS	450	
33	[115]	Asadi	2014				x			x										4	3		LHS	s	TRNSYS	1045	
34	[116]	Aydin	2015				x			x									ANN	8	2		manual	s	EnergyPlus	105	
35	[107]	Eisenhower	2012			x	x								x					7-1009	2		quasi - Monte Carlo	s	EnergyPlus	5000	
36	[106]	Chen	2018		x		x								x				SVM	10	1	x	LHS	s	EnergyPlus	5610	
37	[117]	Chen	2017				x						x	x	x				SVM	9	3	x	LHS	s	EnergyPlus	5610	
38	[119]	Stavrakakis	2012				x				x									2	1		manual	s	Fluent (CFD)	63	
39	[72]	Prada	2018				x	x			x	x	x	x	x				MARS	6	2		pseudo Monte Carlo, Sobol, LHS	a	TRNSYS	50-27k	x
40	[96]	Dhariwal	2017		x		x						x							26	6		factorial	s	EnergyPlus	113-193	
41	[120]	Carreras	2016				x						x							5	2		full factorial	s	EnergyPlus	7776	
42	[121]	Gengembre	2012				x					x								8	4		expected improvement	a	R5C1	150	x
43	[123]	Tresidder	2012				x					x								10	2		optimality criterion	a	EnergyPlus	84-100	x
44	[124]	Zhang	2013				x					x								15	1		expected improvement	a	EnergyPlus	195	
45	[125]	Gilan	2016				x					x								9	2		optimality criterion	a	EnergyPlus	300-1100	x
46	[127]	Zemella	2011				x			x										5	5		optimality criterion	a	EnergyPlus	1500	x
47	[128]	Xu	2016				x						x		x					6	2		optimality criterion	a	EnergyPlus	>90	x
48	[129]	Brownlee	2015				x				x									50	3	x	optimality criterion	a	EnergyPlus	5000	
49	[130]	Wortmann	2018				x				x									13-40	3		optimality criterion	a	EnergyPlus	500	
50	[131]	Van Gelder	2013					x					x	x					MARS	14	2		LHS	s	IDEAS (Modelica)	500	
51	[132]	Cheng	2014					x						x						8	2		full factorial	s	Ecotect	768	
52	[133]	Chou	2014					x		x					x	x	x		SVR, ensemble	8	2		full factorial	s	Ecotect	768	
53	[134]	Symonds	2015					x		x	x								ANN	9	6		LHS	s	EnergyPlus	1000	
54	[15]	Van Gelder	2014					x		x	x	x		x		x			ANN	14	2		LHS	s	IDEAS (Modelica)	500	x
55	[28]	Yang	2016				x	x			x	x	x			x		x	SVM	42	2		initial: LHS, adaptive: space filling	a	EnergyPlus/Daysim	360-430	
56	[30]	Ostergard	2018		x			x		x		x	x	x	x	x			GPR	14	4		Sobol	s	BSim	32-8100	x
57	[83]	Chlela	2009					x					x						Polynomial	13	3		factorial	s	SIMBAD	32-470	x

modified. Currently, building simulation cannot keep up with the speed in the early design phase [11,20]. One reason is that setting up a simulation for one specific concept involves the manual definition of many parameters [21]. Furthermore, the simulation runtime itself is long and may interrupt the train of thought in the creativity process of the architect: ideally the program feedback time would be less than 10 seconds [22].

As a consequence of these drawbacks, researchers have derived requirements for early design tools. [23] point out that a tool for fast global design space exploration is required to quickly evaluate a large bandwidth of different initial design concepts. To reduce complexity in that process, only a few interesting parameters should be considered [20]. This may lead to facilitation of simulation, but should be balanced with simplification [19]. Lastly, Hester et al. [21] and Basbagill et al. [24] suggest early design tools should provide distributions of the performance of the building as an output. This is because at early stage many parameters are uncertain or defined as a range of possible values (*design variability*), and hence simulation results should incorporate that uncertainty.

How a surrogate model helps. Surrogate modelling simplifies the interaction between the building designer and the building simulation process in two ways. First, as surrogates are evaluated instantly (<0.1 s [15]), they are able to provide rapid point estimates [25], or distribution estimates [21] of the building performance. This enables designers to rapidly assess a design concept and explore the design space. Second, in comparison to simulation-based parametric analysis which generates discrete results, surrogate models provide continuous relationships between design variables and building performance metrics. Due to the complexity of the state-of-the art surrogate models, they are capable to capture variable interactions and extract non-linear, multi-modal behaviour [23].

Lastly, the computational layout of surrogate models is lightweight and could be embedded into existing modelling software [26].

2.2. Sensitivity analysis

Sensitivity analysis (SA) is used to rank the importance of parameters on some outcome variable [27,28]. Often it serves as a preliminary step prior to early design, uncertainty analysis (see Section 2.3) or optimisation (see Section 2.4) to reduce problem complexity. There are two different approaches: local and global methods.

In *local methods* inputs of one specific design are perturbed to approximate their partial derivatives. This provides sensitivities of inputs for the considered design. However, in a non-linear building design space sensitivities may change among different building designs [29,30] and local methods may not be suitable for general conclusions on the sensitivity of parameters.

Global methods study the influence of parameters over the whole design space. Apart from fast parameter screening methods, global analysis is computationally more demanding compared to local methods [29]. Two different methods for global analysis exist. First, the structure of the model and its parameters (or: coefficients) may be interpreted as for example in linear regression based SA. Second, in the variance-based approach a large set of simulation samples is statistically analysed. The latter is model-free and studies the impact of one parameter (*first order* sensitivity) or the combinatorial impact of multiple parameters (*total* sensitivity) on the variance of the output.

How a surrogate model helps. Local and global methods are based on simulation samples. Fast surrogate model evaluations speed up the process of sample generation [27]. They could be particularly helpful for variance-based methods which demand large number of samples. For example, the derivation of Sobol indices is sample intensive and usually limited to a small number of parameters due to computational costs [31]. In this case, the speed of a surrogate model enables an increase in the number of parameters to be studied [32].

On the other side, SA also plays a crucial role for surrogate models. Using SA, the most relevant surrogate model inputs can be determined and thus the model complexity reduced. Furthermore, when the surrogate model is very complex (as with a black-box model), SA can be used alongside the surrogate model to obtain a better understanding of the model behaviour.

2.3. Uncertainty analysis

While the purpose of SA is to quantify the effect of a change in one input on the output, uncertainty analysis (UA) studies the likeliness of a change in outputs induced by uncertain inputs [33,34]. A probabilistic view of building performance is very important. It enables quality assurance of building performance under uncertainty as for example required for energy performance contracting [32], to quantify the robustness of the design towards some exogenous variable change (e.g. climate change [35]) or to support the early design stage when many design parameters are uncertain (see Section 4.3.1.2). Sensitivity analysis may be a part of UA to screen the parameter set for the most impactful ones to reduce computational cost [31,32].

Ongoing research was reviewed in [36]. Generally, uncertainties in building design may be grouped into three categories [37]:

- Uncertainty in design parameters during the planning phase,
- uncertainty in physical parameters caused by fluctuations of material properties,
- uncertainty in scenario parameters due to assumptions of internal (e.g. usage of the building) and external (weather and climate data) conditions.

Different ways to quantify that uncertainty exist. Most commonly, uncertainty in parameters is forward propagated to receive a probability distribution of building performance like energy consumption or carbon emissions [36]. This may be done following the external or the internal approach [33].

The former assumes a building simulation model to be a black-box model. The model is used to produce a probability distribution of outcomes given a random set of possible design parameter combinations. The Monte-Carlo method may be the most popular external approach method. In the internal approach the simulation model is modified and uncertainty distributions in parameters is propagated to the model outputs [33].

To conduct the external approach the uncertainty of parameters is required. Usually, it is based on expert knowledge or results from inverse parameter uncertainty estimation if measurement data is available [38]. Bayesian calibration is a common approach for parameter uncertainty estimates and found in [38] or [39] for the building design context.

How a surrogate model helps. Surrogate models are particularly useful to accelerate the derivation of building performance distributions with the external approach which requires a significant number of simulation samples. Depending on the specific approach different numbers of simulation runs are required, varying between 60 and 80 samples for joint uncertainty propagation of all parameters in a Monte Carlo simulation [40] to larger numbers like 2^N or $2N+1$ if the impact of individual parameters and their interactions are broken down as in the factorial or differential method [33].

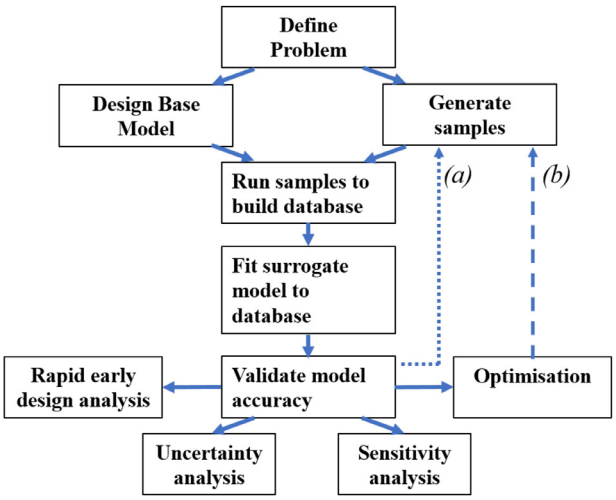


Fig. 2. Overview of the steps to derive a surrogate model. Two approaches exist. In the *sequential approach* sampling and surrogate model fitting happens subsequently. In the *iterative approach*, sampling and surrogate fitting happens iteratively where samples are picked by identifying parts of the design space with unsatisfying model accuracy *(a)* or based on an optimality criterion defined for an optimisation task *(b)*.

2.4. Design optimisation

Building design optimisation (BDO) is one of the fastest growing fields in building simulation research. It is reviewed in [41] and [42]. The goal is to find building designs which optimize a performance objective subject to constraints (e.g. comfort, system size, etc.).

In most common BDO, the fitness function to be optimized is computed using building simulation software. Different optimization algorithms exist that range from direct search, integer programming and gradient-based methods to meta-heuristics like genetic algorithms (GA). Many algorithms are introduced in the reviews above and some of them compared in [43]. The most prevalent approach is GA [41], which is easily implemented and capable of dealing with a wide variety of problems including discrete and continuous variables (e.g. heating system type versus wall thickness), multiple objectives, and discontinuities prevailing in building simulation software [44].

Following [42] an optimisation process may be split into three steps:

1) Preprocessing:	Formulation of the optimization problem; selection of optimizer
2) Optimization:	Running and monitoring of the optimizer; checking of termination criterion
3) Postprocessing:	Visualization of optimization results (e.g. Pareto front); possibly robustness evaluation

The procedure of numerical optimization is iterative, which involves many building simulation runs and may take multiple hours or days until convergence is achieved.

How a surrogate model helps. Surrogate models may speed up convergence rate of BDO. They are applied in two different ways (see Fig. 2 in [13]). In the direct surrogate-based optimisation approach the surrogate model is fitted initially and then used for optimisation.[1] The iterative approach iterates between fitting the surrogate and adding potentially optimal points to the training data.

In other engineering domains where complex simulations are imperative and too expensive without surrogate models (e.g. aerospace engineering [13,14,45]), surrogate models are well established and extensive know-how exists that is yet to be transferred to the building domain. Regarding building performance optimisation, the characteristic of surrogate models to smooth the original fitness function [46] is especially promising as building simulation results were found to have discontinuities [43]. Removing the discontinuities enables the use of optimization algorithms with potentially better performance than meta-heuristics like GA.

3. Surrogate model derivation

The steps to derive a surrogate model are shown in Fig. 2. First, the design problem and the associated design parameters have to be defined. Then the building designer implements an initial building model and picks design samples to be simulated using some sampling strategy. The parameter set defined for each sample is used to modify the base model and run building simulations with it. Results are stored in a database of inputs (design parameter values) and outputs (simulation results, e.g. annual energy consumption). Afterwards, a surrogate model is fitted to the input-output data. Last, the model is validated by computing the model accuracy. It quantifies the deviation of surrogate predictions from simulation outcomes for the same set of inputs.

Most commonly surrogate derivation happens *sequentially*. First sample locations are generated using some Design of Experiments (DoE) strategy and then the surrogate model is fitted. As the samples are defined prior to simulation and not adjusted depending on model outcomes, we refer to this approach as *static sampling*.

The *iterative* approach intertwines sample definition and surrogate model fitting. Samples are iteratively added to the database based on surrogate predictions and simulation results. Therefore, surrogate accuracy and design space complexity *(a)*, or an optimisation criterion *(b)* are evaluated to identify optimal choices for further samples.

In the following we provide details on each step in Fig. 2.

3.1. Problem definition

In the first step design parameters, the inputs to the surrogate model (also known as 'features'), and design objectives, the outputs of the surrogate model, are defined. The selection of inputs and outputs is important as changing them at later stage may require additional high-fidelity model simulations.

Outputs are chosen based on the design objective. Similar to optimisation methods, a surrogate supports studying a specific aspect of building design, e.g. energy efficiency, which is encoded in the surrogate outputs.

The number of design parameters should be limited to circumvent the curse of dimensionality: the number of simulation samples that are needed to create an accurate surrogate of the design space grows exponentially with the number of parameters [47]. Parameters may be chosen based on the design task, or global SA if the most important parameters should be considered [48,49] (see Section 2.2). Besides deciding which parameters to choose, an associated range of possible values needs to be defined.

3.2. Base model implementation

In this step, an initial building design is implemented in physics-based building simulation software like EnergyPlus [50]. Contextual parameters, i.e. those not part of the list of design parameters, are carefully set depending on the problem (e.g. building location, climate, etc.).

[1] Some existing literature refers to model-based optimisation instead of surrogate-based optimisation. This should not be confused with simulation models used for optimization. For clarity we specifically refer to surrogate models.

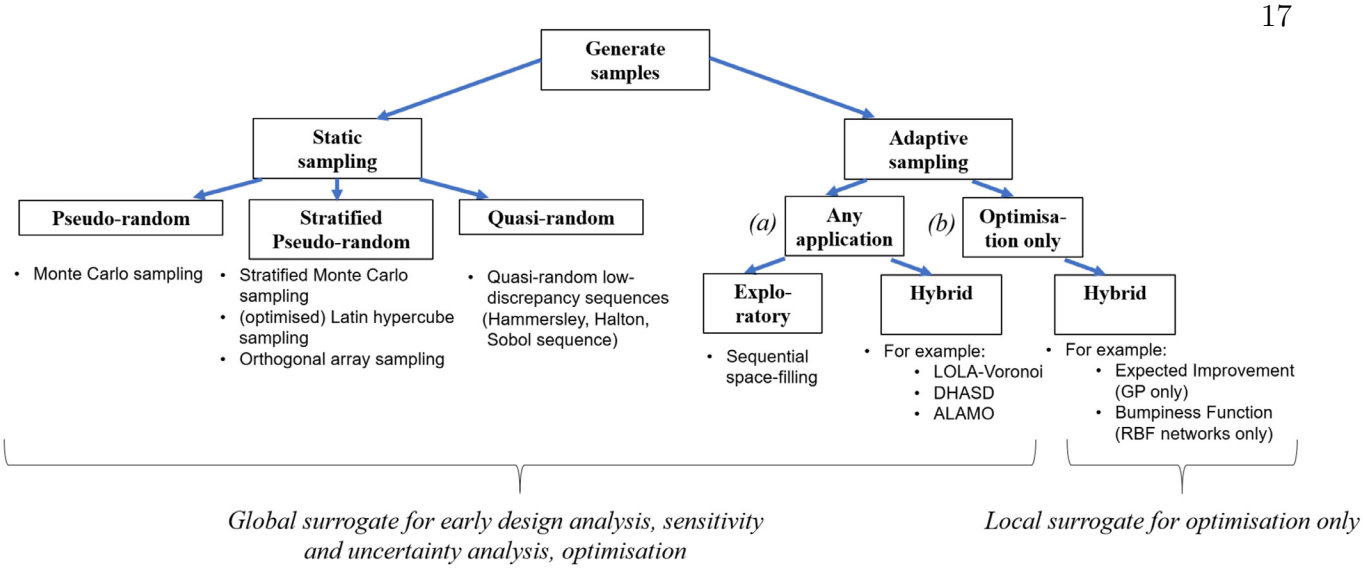


Fig. 3. Overview of different sampling methods [52].

3.3. Database generation

After the selection of parameter inputs and their range, a sampling strategy is chosen (see Fig. 3). The goal of all sampling strategies (also known as design of experiments, DoE) is to select points in the design space to maximise information gain per simulation run while minimizing sampling time. Recent reviews on DoE strategies are given by Yondo et al. [51] and Garud et al. [52].

As outlined above, two types of sampling methods exist. In static sampling all sample locations are defined in one shot prior to model fitting. This provides a global surrogate model being accurate on the whole design space. Common methods include *pseudo-random* sampling like Monte Carlo sampling, *quasi-random* sampling like Hammersly, Halton or Sobol's sequences, and *stratified pseudo-random* sampling like stratified Monte Carlo sampling, latin-hypercube sampling (LHS), or orthogonal array sampling. It is not obvious which of the provided algorithms performs best and depends on the number of variables and samples. A comparison of the methods is given in [52]. Looking at building related literature, we found that LHS is the most applied sampling scheme.

A caveat of static sampling is that it may require a lot of samples to reach an acceptable level of accuracy and therefore, adaptive sampling algorithms are sometimes favourable [51]. The goal of adaptive sampling is to balance *exploration* of under-sampled areas of the design space and *exploitation* of information gained from surrogate or simulation outcomes. Different exploration and exploitation metrics exist, called space infill criteria. They enable to identify under-sampled and complex *(a)*, or potentially optimal *(b)* areas. Before adaptive sampling is applied the surrogate is initiated on a seed of samples (found using a static sampling algorithm). While the adaptive sampling strategy *(a)* produces a *global* surrogate, *(b)* generates a surrogate model which is accurate *locally* where the design space is interesting with regard to a certain design objective. Adaptive sampling methods for global surrogate derivation *(a)* are addressed in [52] and for optimisation *(b)* in [53].

If a global surrogate is wanted, a straight-forward way of adaptive sampling is to iteratively reapply space-filling sampling (see static sampling algorithms) which is purely *explorative*. However, this may lead to inefficient sampling as it does not differentiate between complex and rather uniform areas. Therefore, taking both exploration and exploitation into account may be favourable (*hybrid*). For optimisation purposes, we only consider *hybrid* adaptive sampling methods. Pure exploitation would cause the algorithm to get stuck in local optima. An often applied sample infill criterion for optimisation is the expected improvement (EI) metric which balances model uncertainty with potential optimal performance [54].

To visualise the difference between static and adaptive sampling we derive a surrogate model (Gaussian Process) for optimisation of the Branin test function as shown in Fig. 4. We selected 20 samples using static sampling as well as adaptive sampling (path *(b)* in Fig. 3). The white dots in both plots show the locations of samples using the static approach. In case of adaptive sampling the white dots represent the initial seed to train a first model.

While static sampling leads to a uniform placement of the samples, adaptive sampling quickly identifies the areas where the test function may be optimal (here minimal). This is done by picking locations where the expected improvement criterion is the highest [54].

This small experiment showcases how sampling can follow a specific objective and possibly, increase sampling efficiency to achieve a certain accuracy in the area of interest.

3.4. Surrogate model fitting

Model construction happens in three steps.

1. Data preprocessing and model type selection
2. Model training and hyper-parameter optimisation
3. Model validation

For brevity and because of an abundance of existing literature, we only provide a small introduction to the field and the existing types of surrogate models. The interested reader is referred to [55] for an introduction on machine learning, to [14] for a book on surrogate modelling, and to [30] where different surrogate modelling techniques for building design are compared.

3.4.1. Data preprocessing and model type selection

The input and output data format must be suitable for the surrogate modelling approach of choice. For example, most approaches require the inputs to be numerical instead of categorical. In that case, categorical variables can be transformed to dummy variables [55]. Once formatted correctly, the data is split into training and test samples. A random separation of 20% of the data for testing is suitable. Finally, some model types require the inputs to

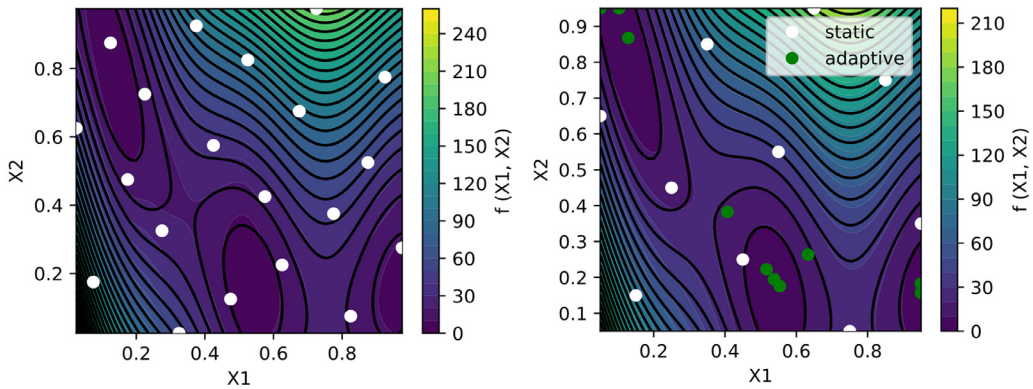


Fig. 4. Showcasing the difference between static (left) and adaptive (right) sampling. On the left 20 samples are chosen based on LHS. On the right, first an initial set of 10 samples was picked using static sampling (LHS) followed by 10 adaptively selected samples using the expected improvement criterion [54].

Characteristics	MARS	RBF	ANN	SVM	GP
Handling of different data types	●	●	●	●	●
Ability to determine variable interactions	●	●	●	●	●
Computational scalability (large number of variables and samples)	●	●	●	●	●
Accuracy	●	●	●	●	●
Interpretability	●	●	●	●	●

Fig. 5. Comparison of different non-parametric surrogate models based on [55, p. 351]. Green, blue and red dots indicate good, medium and poor performance with regard to the characteristics listed. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

be normalized to the same range which ensures equal weighting of variables during model training.

The selection of the surrogate model type is primarily driven by reaching the highest surrogate accuracy possible. Sometimes a trade-off between optimum accuracy and an interpretable model structure is favoured [48,56]. Although each model type has advantages and disadvantages with regard to certain modelling requirements as shown in Fig. 5, many authors suggest the initial use of multiple models to find the most suitable one [13,15].

Model types may be grouped into parametric models and non-parametric models [56,57]. The former uses assumptions on the functional relationship of inputs and outputs. Based on that assumption, a data model is derived whose parameters are calibrated using the collected data. In non-parametric modelling the goal is not to find the correct parameter values of a predefined data model but to find the underlying functional relationship between inputs X and outputs y [57]. In building design, performance metrics like energy consumption may behave non-linearly, featuring discontinuities and multiple modes [30,43,44]. Understanding that behaviour and manually encoding it in a parametric model may be difficult and time consuming. Non-parametric, algorithmic modelling automates this process and thus, may be more suitable for to quickly modelling the relationship of design parameters and performance metrics. In the following, examples for the two model types are given.

3.4.1.1. Parametric models. Multiple linear regression is the most popular parametric model. Its structure and variables are specified manually preliminary to model training. The structure can include variable interaction terms or variables transformed by taking its nth order as done in polynomial regression. Even if variables are combined or transformed, linear regression remains *linear in parameter* meaning no model parameter appears as an exponent or is multiplied or divided by another parameter.

Other parametric models can be developed but they all share a common disadvantage. Unless knowledge allows to derive a valid assumption for the structure of the data model, they are prone to provide questionable analytical findings and lower prediction performance in comparison to algorithmic models [57].

3.4.1.2. Non-parametric models. Different types of non-parametric methods exist. They include artificial neural networks (ANN), radial basis functions networks (RBF), support vector machines (SVM), multivariate adaptive regression splines (MARS), Gaussian Process models (GP) and others. The model types differ in their generic structure.

MARS models may be considered as an extension to linear regression models which automatically identify variable interactions and suitable variable transformations. This is done by a linear combination of multiple basis functions applied to the input vector. Here, the basis function is commonly a hinge function or a multiplication of multiple hinge functions [58]. The hinge function enables piecewise behaviour of the resulting model which is characteristic for MARS models. The multiplication of multiple hinge functions enables to model arbitrary high order relationships and variable interactions.

RBF networks also use linear combinations of basis functions [59]. They use Gaussians as basis functions and apply them to the distance of the input vector to a center vector associated to each Gaussian. Functions that only depend on the distance to a center vector are radially symmetric which explains the name of this model.

Another model type pivoting non-linear basis functions to model versatile mathematical relationships is the ANN. An ANNs consists of multiple cells, called neurons, which receive inputs from and send their outputs to other neurons. Inside a cell the inputs are weighted, summed up and used in a basis function. Typically, sigmoid basis functions are used which imitate the spiking of a neuron in a human brain. Chaining up multiple layers consisting of multiple neurons gives the ANN a high degree of flexibility and in theory, it is capable to model any mathematical function [55].

In GP, observations are considered as realisations of a multivariate Gaussian distribution. The multivariate Gaussian is used as a prior distribution and this distribution is conditioned by existing data. This leads to a posterior distribution of possible functions which generated the data [60].

Support vector machines were originally designed for classification problems. In support vector classification a hyperplane is determined with maximal margin towards the closest observation

of each class. The same method is used in support vector regression to find the hyperplane which centres all observations optimally [61].

3.4.2. Model training and hyper-parameter optimisation

After the data is prepared and the model is selected, its parameters and weights are found using a specific training algorithm. For example ANNs are trained via the well-known *backpropagation* algorithm. Apart from training the model weights, non-parametric models require *hyper-parameters* to be specified. Hyper-parameters allow to tune the variance of the predictions of the surrogate. They should be optimised to balance variance with bias to avoid *overfitting* the model on the training data. An overfitted model does not generalize well on unseen data. To select the hyper-parameters usually multiple different settings are compared in a grid-search or Bayesian optimisation is used [62].

3.4.3. Model validation

Model validation is done using dedicated test data. The accuracy of the model is quantified using different performance metrics. Typical choices are mean absolute error (MAE) or the coefficient of determination (R^2) which quantifies how much of the variance in the data is explained by the model.

3.5. Tools

Existing tools may be sorted into two groups: dedicated surrogate modelling toolboxes and those covering portions of the surrogate derivation process.

The first group of tools covers all steps from 3 to 5 from above. They offer different DoE strategies and surrogate types. Due to the excellent performance of surrogate models on optimisation problems, toolboxes are often designed specifically for optimisation purposes. Matlab users are referred to Matsumoto [63] or to SUMO [64]. A Python option is the SMT toolbox, which focusses on gradient-based optimisation,[2] although the choice of methods is rather limited.

Other tools only provide software for specific steps of surrogate model derivation. The Python toolbox PyDOE offers a set of different static sampling methods.[3] The EPPY toolbox allows to access EnergyPlus input files in Python, which enables to quickly transfer generate simulation models given a set of samples.[4] The well known machine learning toolboxes ScikitLearn [65], Tensorflow [66] and PyTorch [67] all feature different surrogate model types and model validation schemes.

Opossum, a plug-in to Grasshopper, is the only surrogate toolbox dedicated to building design [68]. It can only be used for BDO problems and not for deriving global surrogate models. Opossum is based on the Python toolbox RBFOpt [69].

4. Review of surrogate modelling for building design

A significant amount of literature exists to explore and realise the potential of surrogate models, also termed meta-model or response surface model, for building design. The literature review was started with a search through publications listed in Google Scholar and Web of Science using the terms “surrogate model”, “building design” and “building performance design”.[5] This provided a list of 30 publications. The list was extended by analysing their bibliography. Finally, 57 sources were found, shown in Table 2.

Apart from that, previous reviews in the wider context of surrogate modelling applications were collected. They are introduced in the following Section (4.1).

Although a lot of effort was invested to compile a representative set of ongoing research, the intention of this review is not to be exhaustive. In particular, applications of statistical models trained on non-simulation data (e.g. [70]) or simplified physical models are disregarded (e.g. [71]). This also involves applications of surrogate models for model calibration as in [38].

4.1. Previous reviews

This is the first review on the use of surrogate models for Building Performance Simulation (BPS). However, multiple papers include review sections addressing applications of surrogates for building design.

A review and comparison of model types are found in [16] and [15]. Prada et al. [72] looked at the suitability of different types of surrogate models for evolutionary building design optimisation. A comprehensive review of the use of data for building design may be found in [5]. In [73] and [36] surrogates are mentioned in an overview of literature in the field of uncertainty quantification and in [27] they are part of a review on sensitivity analysis. In [42] and [41] sections cover surrogate models applied to building design optimisation.

Other fields in computational building science use similar techniques as in surrogate modelling, for example [74] and [6] reviewed data driven energy demand forecasting.

In other engineering domains where computational experiments are costly surrogate modelling has been applied extensively. An overview of the application of surrogate modelling in aerospace engineering is given by Wang and Shan [13], Forrester et al. [14], Simpson et al. [45] and Queipo et al. [75].

4.2. Overview of publications

Table 2 and Fig. 6 give a summary of each publication including subject, surrogate model type and sampling strategy. Most of the papers address building design optimisation (22 publications), and leverage surrogate models at the early design stage. Also a wide distribution in the fields of sensitivity (16) and uncertainty analysis (9) was found.

Aside from the applications of surrogates for building design problems, 16 papers compare the suitability of different types of surrogate models for building design.

4.2.1. Surrogate model types

Fig. 6(b) shows that in half of the studies, parametric models are used (compare Section 3). Apart from multiple linear regression, this group encompasses polynomial, stepwise, and LASSO regression. The second most models found in literature are Gaussian Process models (GP) and third most common are artificial neural networks (ANN). Other model types include multivariate regression splines (MARS), support vector machine (SVM), random forest (RF), radial basis function (RBF) and model ensembles. An introduction to the models is found in [76].

4.2.2. Sampling strategies

Eleven studies used adaptive sampling (Fig. 6(d)). All but five use them in combination with a GP model. The majority of papers used static sampling strategies with a strong preference towards latin hypercube sampling (LHS)(15). Other sampling strategies include random, orthogonal array, full-factorial, Box-Behnken design

[2] https://github.com/SMTorg/smt.
[3] https://pythonhosted.org/pyDOE/.
[4] https://pythonhosted.org/eppy/.
[5] As Google Scholar does not support the use of parentheses multiple searches equivalent to [(“surrogate model” or “meta-model” or “metamodel”) AND (“building design” or “building performance design”)] were conducted.

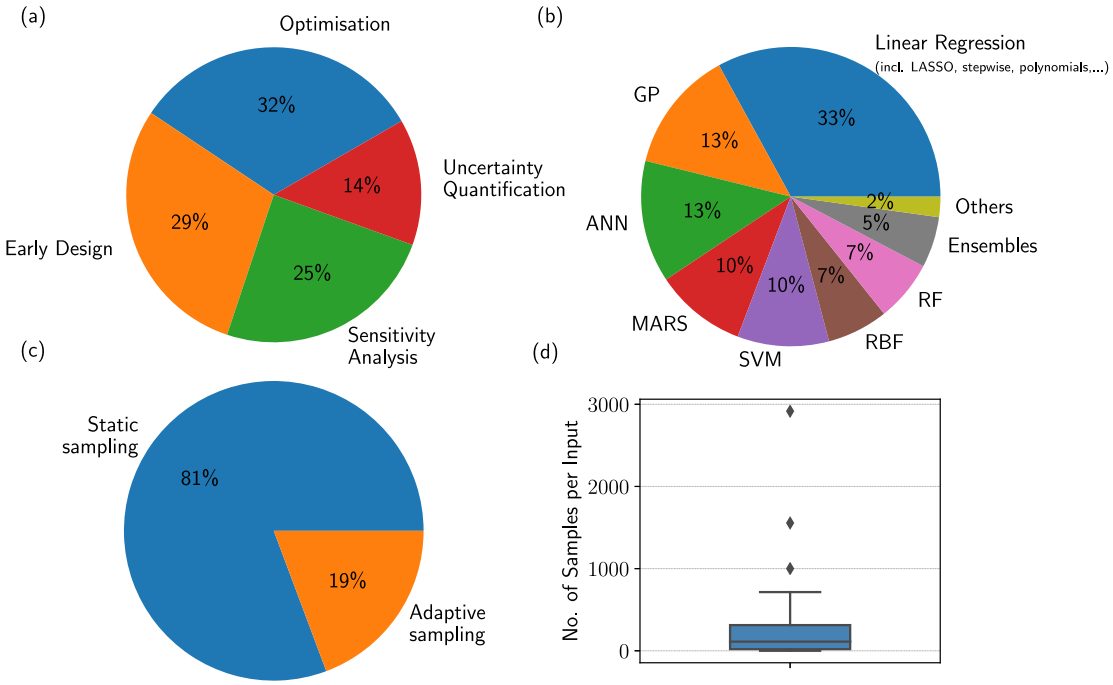


Fig. 6. Overview of publications. Figure (a) highlights the applications of surrogate modelling found in the building design research domain. Figure (b) shows the different surrogate model types used. Figures (c) and (d) focus on the sampling methods to derive an accurate surrogate. Figure (c) shows the share of papers that used static sampling instead of adaptive sampling. Figure (c) indicates how many samples per input were collected in each paper.

and L12-Taguchi tables based sampling. A limited number of studies used manual sampling or evaluated all possible combinations of design parameters (full-factorial).

The number of simulation samples per input is shown in Fig. 6(c) and Table 2. The range is large spanning from single digits to thousands of samples per input. Quantifying the sampling efficiency by number of samples per input is questionable as the design space does not increase linearly but exponentially with each input parameter added. Another option would be to quantify the share of the design space covered by samples [25]. However, as input variables may be continuous, discrete or categorical, and their ranges change drastically among the different studies, the design space size of each study would have to be calculated individually. This is beyond the scope of this review.

4.2.3. Model objectives (outputs) and parameters (inputs)

Fig. 7 shows which inputs were used for different model objectives (outputs). Annual energy demand (and energy use intensity) is the most common output. Another big fraction of papers approximated heating and cooling demand.

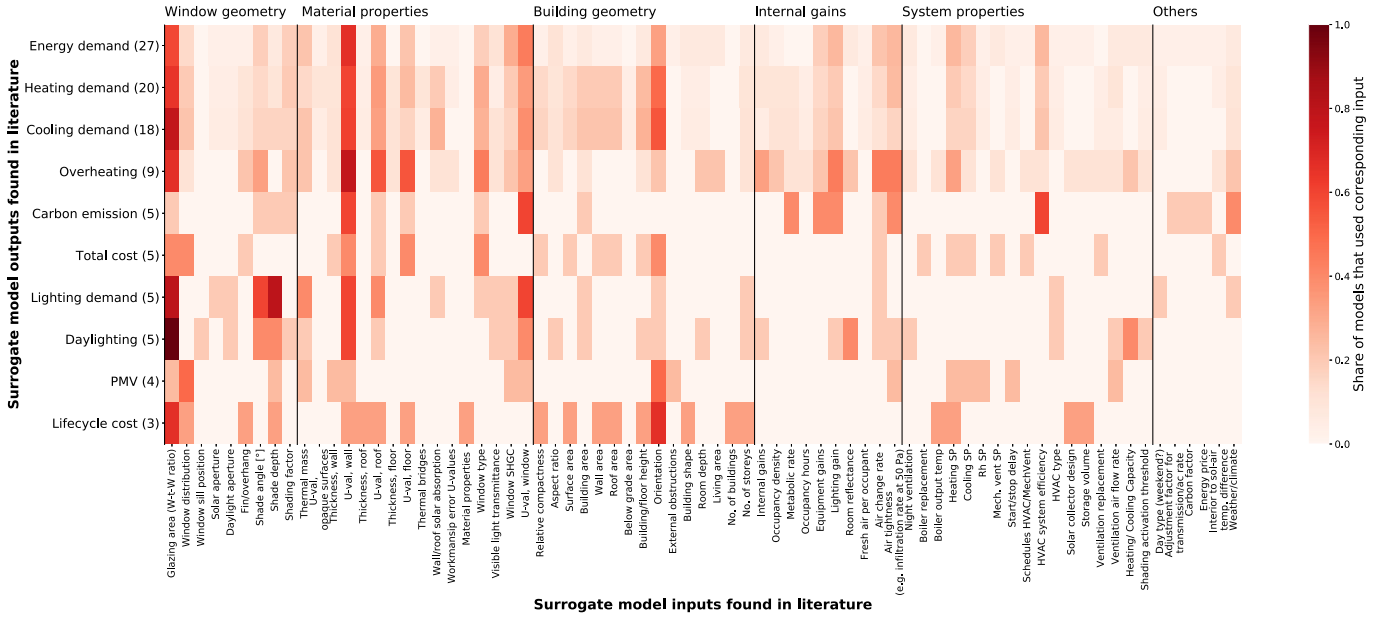


Fig. 7. Usage of in- and output variables in the literature. The figure shows the share of models which used a specific input for a specific output. Next to the outputs (y-axis) the number of associated studies is shown in brackets.

To model the energy demand mainly inputs on building geometry, window geometry and material properties are used. Many authors rely on window-to-wall ratio and wall thermal transmittance as model inputs.

4.3. Discussion of papers grouped by purpose

4.3.1. Early design

As presented in Section 2.1, surrogate models may play a key role in overcoming current limitations of building simulation tools for early design applications. The following literature used surrogates

- To provide rapid feedback during early design,
- and to run large amounts of simulations quickly to provide estimates of *design variability*.

As the amount of existing literature is large, we subdivide the literature corresponding to the geographical scope on which the surrogate model was validated. The scope ranges from one specific building location to multiple climate zones.

4.3.1.1. Rapid feedback. The first group of authors used surrogates for the design of **one building**:

In an early work from 2002, [77, #1] embedded a polynomial regression model into an early design tool which computes energy demand based on 10 parameter inputs. The paper shows that the concept of surrogates has existed for a long time. However, the provided method is not as robust in comparison to more recent publications. For example, surrogate modelling errors are not computed on separate test data.

A similar tool, the design space exploration assistance method (DSEAM), embeds a surrogate model into a CAD tool to provide instant feedback during early design [26, #2]. Performance predictions are visualised as a three-dimensional surface and in a parallel coordinate plot. The method is showcased for the design process of an urban office building. The visualisation techniques enabled to intuitively identify the most influential design parameters for reducing energy consumption.

Geyer and Schlueter [25, #3] considers the use of surrogate models covering a large geographical scope and multiple building types rather suitable for educational purposes but foresees the need for highly individualised surrogates to ensure high accuracy for a specific building. Therefore, they developed an automated surrogate derivation method customized for building design practitioners. They fitted a polynomial model with an algorithm which automatically selects exponents and interaction terms. Consequently, their modelling scheme is non-parametric, which is unusual for polynomials, and allows to combine high surrogate interpretability and accuracy.

Instead of developing holistic tools the following papers looked at more specific elements of surrogate modelling. The authors of [78, #4] tried to overcome the curse of dimensionality, i.e. the computational cost incurred by high numbers of surrogate inputs and outputs. They used ANNs and LASSO regression, which scale more efficiently than MARS, GP and RBF, to build models with 156 inputs and 80 to 90 outputs that are time resolved at 15 min. Their model was trained on an extensive EnergyPlus dataset (267 TB) generated on large cluster computer. This clearly bound their method to users with access to large hardware, unless their surrogate would generalize well, i.e. it could be reused for a variety of building design problems. Unfortunately, this was not addressed by the authors. An advantage of the 15 min-resolution is that it enables users to not only speed up building design, but also automate tasks like system sizing and demand profiling.

Yi et al. [79, #5] computed emergy[6] using EnergyPlus output data and trained a surrogate model on that postprocessed data. They achieved a prediction performance of $R^2 \approx 0.62$[7] which is low in comparison to studies where surrogates estimated energy demand.

Maltais and Gosselin [81, #6] studied daylighting with regard to comfort (glare index) and lighting demand. They fitted a polynomial model for both outputs and found a large accuracy difference although the same simulation data was used for model fitting ($R^2_{GI} \approx 0.95$, $R^2_{LD} \approx 0.78$). The finding that accuracy varies for different model outputs was observed in other publications (e.g. [30,82,83]).

Lastly, Korolija et al. [84, #7] provides an interesting extension to surrogates of whole building simulation programs. The authors state that including detailed HVAC systems into simulations increase simulation run time by 1.3 to 3.7 and that it requires advanced know-how which architects may not have. Therefore, they derived a polynomial model to map building energy demand to energy requirements of secondary HVAC systems which distribute thermal energy inside the building (e.g. ducts). They found that, although a simplistic polynomial was used, accuracy is satisfactory with a relative error of less than 10% in more than 80% of the cases for cooling and heating requirement. In future work one could study if accuracy can be improved with a non-parametric instead of a polynomial model. Nonetheless, the given performance seems good enough to save time by replacing detailed HVAC modelling with a surrogate during early design.

In the second group of studies, surrogates were derived considering **multiple climate regions**:

Catalina et al. conducted two studies to apply linear regression to receive rapid information on heating demand during early design [85][86, #8]. In their latter study the regression model included quadratic terms and was fitted using iteratively re-weighted terms. The model uses only three inputs (heat loss coefficient, south equivalent surface and difference of indoor to outdoor temperature). Although the model was trained and validated on simulation data, the model predictions were also compared to actual measurements of buildings. A significant difference of predicted to measured annual heating demand is found and could only be taken into account by introducing a building-specific correction term to their polynomial.

Hygh et al. [87, #9] applied multivariate linear regression to approximate energy use in four different climate zones. In comparison to the previous study, 27 input parameters were used leading to only slightly higher accuracy ($R^2 > 0.98$) than the previous paper ($R^2 \approx 0.97$) but allowing a wider variety of designs to be analysed using one surrogate. Similar to [87], [88, #10] and [89, #11] used one multivariate linear regression model per climate zone. Both authors point out that the accuracy varied for different climate zones. All three studies used multiple models for each climate zone.

The question arises if variables capturing the characteristics of different climate zones can be found. When used as inputs, they could enable the use of one surrogate covering all climate zones. An approach is given in [90, #12], where a SVM model is derived that estimates energy use of naturally ventilated commercial buildings in Brazil. They generated a dataset with 418 different weather files. The weather data was then reduced to a few statistical variables and used as surrogate model inputs. Romani et al. [91, #13] also used only one model (full quadratic polynomial model) to predict the heating and cooling demand in Morocco (four cli-

[6] Energy refers to the amount of solar energy embodied in the energy used up during a service or production [80].

[7] Calculated from the provided F-score.

mate zones). They found variable interaction helped to compensate for the use of one model only.

Thirdly, one study was found covering **multiple building types** within one climate region [92, #14]. They derived multiple single-output ANNs to find optimal retrofit strategies for Southern Italy. Their six separate networks estimate heating demand, cooling demand and occupant comfort of the existing non-retrofitted and retrofitted building stock.

Recently, some researchers focussed on finding **general** surrogate models without a climatic or geographical scope.

In [93, #15] the same authors as in [89] tried to find more generalizable surrogates by integrating know-how on building physics to derive more meaningful features (inputs) from common design parameters. Their features include energy gains due to transmission, air change rate and solar heat gain. Thy were calculated assuming steady-state behaviour but fed to a surrogate which approximates dynamic, i.e. non-steady-state, simulation. They achieved high accuracy scores for a single-building case study ($R^2 \approx 0.99$) but further details and benchmarking are required.

Singaravel and Geyer aimed to decompose a "monolithic surrogate model" into multiple components [82,94, #16]. In the first approach they suggest fitting multiple ANNs, each approximating heat gain through an individual building element like a wall or a window. Adding up the outputs of the individual models they computed the whole building performance. In their second approach they compartmentalised one surrogate into multiple approximating heating and cooling demand. The use of the recurrent long-short term memory network (LSTM) allowed to model dynamic effects. They studied the generalizability of both approaches on three test cases. Based on the results of the most complex building design case, the first approach ($R^2_{cooling} \approx 0.98$, $R^2_{heating} \approx 0.85$) seems to outperform the zonal LSTM model, but no final conclusion is drawn by the authors.

4.3.1.2. Design Variability. After studying methods for early building design [11, #17], Ostergard et al. developed a new design methodology to guide sustainable building design with multiple stakeholders involved [23]. They propose to first evaluate the performance of large number of designs, using a surrogate model, and sequentially filter them using specifications on the final building performance (outputs). This provides distributions of possible design choices (inputs). To visualise the impact of performance specifications, the parallel coordinates plot was favoured.

Instead of specifying outputs, Hester et al. [21] determine the change of the output distribution if one of the design parameters is decided. They used a linear regression model to run Monte Carlo simulations after each design choice. Part of the authors conclusion is that not only the speed of the surrogate is helpful, but also the reduced number of parameters required to provide a performance estimate.

A similar study was done by Basbagill et al. [24, #18]. The authors constructed probability distributions for life cycle cost and performance treating decision parameters as random variables. The distributions are derived from a database generated with eQuest [95] using orthogonal array sampling. Although no surrogate is used (but a fast physics based simulation model instead), the methodological steps are similar to those of surrogate modelling, and show that fast simulation software is an alternative to surrogate modelling.

4.3.2. Sensitivity analysis

There are three cases found in literature where surrogate modelling and SA are combined (compare Section 2.2). *(i)* use SA prior to surrogate modelling for variable selection, *(ii)* use surrogates to accelerate variance-based SA, and *(iii)* use SA complementary to surrogate modelling to increase analytical insight into the data. While in general non-parametric methods are preferred for building surrogate models, parametric methods are a regular choice for SA as one can easily access variable importance estimates by looking at the regression coefficients (e.g. linear regression).

For brevity, this section does not cover the full literature on SA in which linear regression was applied. Further literature can be found in [27].

4.3.2.1. Variable selection for surrogate models. Many studies use SA for variable selection. Here we summarize eight contributions which share a similar approach.

Dhariwal and Banerjee [96, #41] conducted fractional factorial design-based sampling and determined the most impactful parameters using Morris' method. The parameters found are used as model inputs to a second order polynomial surrogate (response surface model).

Hopfe et al. [97] computed standardized rank regression coefficients (SRRC) to quantify the impact of uncertain parameters. They chose the five most influential to complement the design parameters as inputs of the GP. Similarly, SRRCs were used in [98, #20].

Multiple SA methods (Pearsson, Spearman, Kolm and Krusk coefficient) were computed in [99, #22] to take linear, monotonic and non-monotonic, and asymmetric variable correlations into account. The most important variables were used as inputs of an ANN to emulate internal air quality simulations of a building stock. They found differing results, which may be caused by the way the four SA methods handle non-linearities. They suggest the use of simple scatter plots to discover non-linearities and to pick the SA accordingly.

Maltais and Gosselin [81, #6] used linear regression and variance-based SA prior to fitting a polynomial to estimate natural daylighting performance. As part of their study, they looked at the numbers of samples required to achieve stable sensitivity coefficients. While standard regression coefficients stabilized after 600 runs, the Sobol indices (variance-based SA) converged after 1900 runs. The data generated from those 1900 runs was subsequently used to train a surrogate model.

Ostergard et al. [30, #56] compared different surrogate modelling techniques. To facilitate model fitting of many different model types, they applied a global SA on the *hyper-parameters* using Smirnov two-sample statistics.

The same authors in [23] and [32] chose surrogate inputs by ranking parameters with the Morris screening method. The method was favoured as it is fast, requires fewer simulation samples and its qualitative ranking of variable sensitivities is close to more complex SA methods.

4.3.2.2. Surrogate model based sensitivity analysis. In [32, #23] two different kinds of surrogate models and two types of sensitivity analysis are applied. The authors fitted a surrogate (polynomial chaos expansion model) to conduct a variance-based SA with 24,000 samples. For the derivation of the surrogate, Morris screening was conducted beforehand to find input parameters as introduced in the section above.

Eisenhower et al. [29, #24] emulates the design space with high dimensional model representations [100] to compute variance-based global sensitivities for thousands of parameters. This would take multiple days without a surrogate.

Tsanas and Xifara [101, #25] used a random forest model (RF) to estimate the energy performance of a building. RFs provide parameter importance ranking through the impurity metric which the authors compared to SRRC-based ranking. They found slight differences and warned of the limitations of linear regression in dealing with collinearity.

In [102, #26] both variance-based SA metrics using a MARS model and SRRCs were computed to find the most important parameters for building-related carbon emissions taking climate change into account. Although the latter ignores variable interactions, both approaches provided similar results. 6000 simulations were required to conduct the variance-based SA in their case study on a UK office building. Without the use of a MARS model this analysis would take multiple days.

4.3.2.3. Interpretation of surrogate models. The last application of SA is to provide insight into the functional behaviour of a surrogate model, if its mathematical structure is too complex to be comprehensible intuitively.

[103, #27] use the Correlation-Adjusted corRelation (CAR) score [104] to understand variable importance for a set of campus buildings. The same data is used to derive multiple surrogates. The combination of both statistical approaches provides interpretive and predictive tools to the building designer.

Similarly, Hygh et al. [87, #9] and Chen et al. [105, #28] computed standardized regression coefficients alongside training a stepwise linear regression and a MARS surrogate model. The latter also used bootstrapping methods to validate the robustness of the sensitivity coefficients. The same authors conducted a SA and a heuristic optimisation in [106].

4.3.3. Uncertainty analysis

Output uncertainty quantification is similar to the assessment of design variability during the conceptual design stage (see Section 4.3.1), and often conducted alongside a sensitivity analysis (see Section 4.3.2). Like design variability assessment and sensitivity analysis, current uncertainty quantification methods mostly rely on sampling based methods, i.e. input parameter distributions are converted to output distributions using Monte Carlo simulations [73]. The idea is to use surrogates to accelerate Monte Carlo simulations [36], however the existing literature is rather limited.

Hester et al. [21, #29] sequentially generate probability distributions of the output after each design parameter is specified. This visualises the converging distribution of the output with each design decision taken. Here, they used a linear regression surrogate model to avoid long computation times.

Eisenhower et al. also derived probability distribution on comfort and annual energy demand based on 1009 input parameters uniformly distributed within 20% of their baseline [107]. They compared the distributions of both 5000 simulation runs as well as SVM evaluations and found high agreement in the mean and variance.

Rivalin et al. [32, #23] and Kim [98, #20] studied the use of Gaussian process emulators and polynomial chaos expansion (PCE). The former paper first applies LHS to derive the PCE model. Once they have an accurate model they re-apply LHS to derive the model output dispersion and distribution faster than with random Monte Carlo simulation.

Papadopoulos and Azar [108, #30] use a surrogate model to study the influence of varying levels of control of occupants and facility management under uncertain occupant behaviour. After training the surrogate, a linear regression model, they generate 11^3 cases each with a different level of control of occupants on lighting, equipment and thermostat setpoints. They underpin the cases with uncertainty of human behaviour and generate 25 samples for each case, such that the surrogate model is evaluated for 33275 samples. They visualise the results in an appealing three-dimensional map.

One of the reasons for the scarcity of existing literature may be the findings of Macdonald [109] and Lomas and Eppel [40], who stated that, disregarding the number of uncertain parameters, between 60 and 100 samples are required to receive an accurate probability distribution of the outputs. Based on existing literature this number of samples is probably not sufficient to derive an accurate surrogate model (see column "number of samples" in Table 2). Thus the majority of papers used standard building performance simulation for sampling, sometimes running them on high performance computing facilities [29,35,110,111]. A workbench for propagating input uncertainties to performance uncertainties using EnergyPlus may be found in [112].

4.3.4. Design optimisation

In this section we review the papers which replace simulation models by surrogate models to accelerate the search for optimal building design parameters. The literature can be sorted into two different groups: *(i)* Direct surrogate-based optimisation (*(a)* in Figs. 2 and 3) and *(ii)* iterative surrogate-based optimisation (*(b)* in Figs. 2 and 3).

4.3.4.1. Direct surrogate-based optimisation. An early application of surrogate models for BDO is found in [113, #31]. Wong et al. used an ANN based grid search to determine optimal selections of solar aperture, daylight aperture, overhangs and side fins to minimize annual energy consumption. The authors limited the grid search to only 41 surrogate model runs although their surrogate model should be cheap to evaluate much more samples.

Magnier and Haghigat [114, #32] used an ANN and the NSGA-II optimizer to minimize energy consumption and comfort. They reported that the surrogate-based optimisation achieved an accuracy of within 1% of simulation-based optimisation and only required seven minutes, but stress that generating the database underlying the surrogate took three weeks. Nonetheless, if the same number of model evaluations during optimisation would have been conducted with a simulation, the process would have taken 10 years. The relatively long simulation time might be caused by choosing 2 min time-steps for their simulation, while having a workstation with a 1.66 GHz processor.

Shortly after that Asadi et al. [115, #33] published a similar study focussing on retrofit optimisation (between one and three objectives) using a validated EnergyPlus base model for the surrogate derivation. Sample simulation took three days and their model achieved a decent accuracy (MRE $< 2.5\%$) on a validation set. Like other authors, they did not report the accuracy of the optimality candidates. They point out that the speed of surrogate-based optimisation (< 9 min) enables designers to explore different design strategies at early stage. In comparison, simulation-based exhaustive search would have taken 75 days. They suggest increasing the number of design parameters and incorporating surrogate uncertainty prediction to further expand the insight for architects. A study similar to [114] and [115] can be found in [116, #34], which focussed on L-shaped multi-story office buildings.

While the previous authors used an ANN model in combination with a genetic algorithm, Eisenhower et al. [107, #35] and Chen et al. [106, #36][117, #37] used SVM models. While the latter used NSGA-II like previous papers, Eisenhower et al. leveraged that a surrogate model enables the use of gradient-based instead of derivative-free optimizers. Comparing both on a multi-object optimization of comfort and energy demand, the results were equally stable but gradient-based optimizers converged significantly faster (a few seconds instead of some minutes). In all three studies sensitivity analysis was conducted to reduce the number of design parameters prior to optimization. Eisenhower et al. showed that similar optima were found with seven input parameters as with 1009 parameters. Hence, increasing the number of inputs barely increased the optimality score.

Constrained gradient-based optimisation (sequential-quadratic programming [118]) was used in [119, #38] together with an RBF model to maximise comfort of naturally ventilated buildings using

window geometry. Their surrogate was trained on simulation data generated by a sequence of computational fluid dynamics (CFD) and building energy simulation. To validate the optima, they also ran simulation at the optima and found a MAE of $< 10\%$ in the RBF-optimization outcomes. The validation did not include a comparison of surrogate-based and simulation-based optimization. As the building considered was simplistic, further research is required for a general conclusion on their method.

A comparison of different surrogate models and different sampling approaches for evolutionary design optimisation is given in [72, #39]. The models were compared with regard to their efficiency, efficacy and solution quality. The authors recommend MARS models over GP, RBF and SVM models due to higher accuracy. Part of their study was an analysis of potential time savings using surrogate models (including sampling). They found savings of more than 80% to be feasible, particularly for complex design spaces. In comparison to the other papers they measured optimum accuracy not only at specific points, but computed the generational distance between the surrogate-based and the simulation-based Pareto fronts. One of many findings is that increasing the number of training samples leads to a lower generational distance.

Above, only non-parametric models were presented which are complex and difficult to interpret. Some authors prefer simpler approaches like polynomials. [96, #40] used a second-order approximation and benchmarked it against simulation-based optimization with and without parameter importance analysis. Carreras et al. [120, #40] optimised a cubic house to minimize cost and environmental life-cycle performance. They used cubic spline interpolation as a surrogate and reduced gradient optimization to determine the best insulation thickness. Including the time for database generation they found a time reduction of 8 times, down to 21.3 hours in comparison to simulation-based optimisation.

4.3.4.2. Iterative surrogate-based optimisation. In comparison to direct optimisation, iterative surrogate-based optimisation relies on a space infill criterion which balances exploration and exploitation. This difference leads to a changing preference on surrogate model type. While in the previous section a lot of studies used ANN or SVM models, here often GP models are applied which can quantify model uncertainty. This can serve as an *exploration* criterion for adaptive sampling (see Section 2).

An early work on the use of GP in the BDO domain by Gengembre et al. [121, #42] minimized life-cycle cost and energy consumption using the constrained efficient global optimizer [122]. The GP model was updated with samples chosen to maximise the expected improvement criterion.

Similarly, in [123, #43] and [124, #44] the expected improvement criterion is used. In the former study, the optimisation process is benchmarked against simulation-based optimisation using NSGA-II. It was found to have a steeper convergence curve and to require fewer high-fidelity model simulations. However, in the case of multi-objective optimisation this could not be confirmed.

Gilan et al. [125, #45] used a combination of GP and NSGA-II. In comparison to other studies, they computed the space infill criterion based on a whole area of the design space instead of an individual point. They calculated the mean posterior variance of the offspring from each iteration of the optimizer (50 samples). Comparing their method to direct surrogate-based optimisation, they found good agreement of the Pareto Front while cutting runtime by two thirds.

Hopfe et al. [97, #19] performed optimisation using the SMS-EMOA algorithm [126]. As the objective function they used the mean value of 201 perturbations around a point proposed by the optimizer. The goal is to find more robust solutions. To the best of our knowledge, they could replace their GP model with any other surrogate model type. They claimed that their approach helps to reduce the number of samples needed to find an optimal solution by 5–20% compared to simulation-based optimization.

Besides the literature on GP, the following publications used methods which are independent of the surrogate type. They rely only on model predictions instead of posterior variance estimates provided by GP models. One early study [127, #46] updated ANNs at each iteration with samples selected by NSGA-II leading to a locally accurate surrogate. The model was initialised on a set of 50 samples and the optimizer provided 50 samples at each iteration. In a similar fashion, [128, #47] used an SVM and [129, #48] a RBF model to minimize building cost. The former reported their method reduced optimization time by up to 60% on a case study.

Lastly, Wortmann developed Opossum as plug-in for Grasshopper. His tool, which is based on RBFOpt [69], adaptively trains and optimizes an RBF model [130, #49]. In a comparison with eight other optimization schemes, RBFopt was found to be the fastest converging and second most stable (after direct search) to optimize the energy demand of a building with 13 design parameters. Furthermore, it was the best performing algorithm in maximizing useful daylight illuminance (UDI) while minimizing glaring effects. This paper clearly shows the great potential of iterative surrogate-based optimization for building design problems.

5. Trends and practical aspects

This review confirms that surrogate models are a strong element in current building performance simulation and optimisation research, and results have shown that they are a suitable alternative to common building simulation models in certain cases. Performance analysis during the conceptual building design stage, sensitivity and uncertainty analysis, as well as building design optimisation are more accessible, primarily due to the large reduction of computational cost.

In the following section, we list application trends and practical aspects extracted from the reviewed literature.

5.1. Trends in the application of surrogate models

- As surrogate models lower the computational burden of early design, sensitivity and uncertainty analysis, it becomes possible to get insight into building performance over the whole space of potential design options.
  A good way to visualise this is the parallel coordinates plot [23,26]. In comparison to simulation-based design exploration, this allows users to intuitively explore multi-dimensional spaces and find promising designs in a limited time.
- The value of surrogate models hinges on the decrease in time to conduct a certain analysis while maintaining high accuracy. Although many examples on the use of surrogates for early design, SA and UA exist, there is a lack of understanding how large the time savings can be. Only in papers on optimisation analysis, thorough analyses of time savings were found.
- Surrogate-based optimisation showed promising first results to speed-up building design optimisation.
  The listed publications achieved a time reduction of up to 80% [72] and the identified optima have proven to have high quality in comparison to full simulation-based black-box optimisation [72,107,130]. Examples for time savings are given in Table 3. An open question is whether direct or iterative surrogate-based optimisation better fits the requirements of building designers. Iterative surrogate-based optimisation may be fast and efficient [130], but as stated in [107], direct optimisation using a global surrogate allows to

Table 3
Examples from the literature for potential of optimization time reduction while maintaining accuracy of the optimum.

Optimization strategy		Multi objective	Optimizer	Time reduction	Comparison to simulation-based optimum
Direct surrogate-based					
	[114]	x	NSGA-II	-97% (10y to 3w)	Energy demand: $<2.5\%$, Overheating: $<25\%$, max. error on global validation set
	[96]	x	NSGA-II	-55% (27h to 12h)	$<10\%$ max. error on global validation set
	[107]		IPOPT/NOMAD	+4% (49h to 52h)	0.7% deviation from true optimum
	[117]	x	NSGA-II	n/a (7.4h to n/a)	1.59% deviation from true optimum
Iterative surrogate-based					
	[125]	x	NSGA-II	-52.2% (14h to 7h)	3.17% difference in hypervolume of Pareto Front
	[72]	x	NSGA-II	-82% (71h to 13h)	75%-80% samples of original Pareto Front found
	[128]	x	NSGA-II	-60% (23h to 9h)	optima are close to true Pareto Front but have low diversity (spread metric Δ increases from 0.41 to 1.01)

easily change the optimisation objective or optimizer settings without rerunning simulations.

- Recently, researchers have been trying to find more general surrogates applicable to many different problems. One may envision that if a surrogate is highly generalizable, it could fully replace building simulation tools for the most common types of building projects.
  The maximum scope of a single surrogate model has been broached by multiple publications. Most authors used surrogate models only for a specific building and therefore, Geyer and Schlueter [25] focussed on automating the surrogate derivation process. Others have fitted a single surrogate to estimate the performance of multiple buildings of a specific type [92] or in one climate region [86,87]. In future, one could capture weather data in a few descriptive variables and use them as inputs to a general surrogate as introduced in [90].
  Another option is given by Singaravel et al. [94]. In their grey-box approach they used domain knowledge to split one surrogate into multiple physical entities representing energy fluxes through walls, floors, etc. Further research is required to support their promising initial findings.
- Lastly, most surrogate model types lack interpretability of their mathematical structure and are not suitable to answer analytical questions. One way around is to increase the number of surrogate model outputs. For example in the aforementioned grey-box approach multiple physical meaningful metrics are estimated. Another example was given in [78] where time resolved energy use instead of annual performance metrics were reported.

5.2. Practical aspects

- In the reviewed publications, it seems feasible to explain more than 95% of the variance in simulation results ($R^2 > 0.95$) for energy, heating and cooling demand estimates with one surrogate model for one or more buildings in one or more climate zones. The accuracy was found to be lower for other kinds of output (e.g. *Max* CO_2 [30], or *Overheating* [131]).
- Both model selection and hyper-parameter optimisation are important to achieve high accuracy as shown in different comparisons of surrogate modelling techniques for building design [15,30,72,132–134]. For example, [30] advocates the use of ANN for extensive analysis, GP for non-experts to get high accuracy, and MLR for quick, automated surrogate modelling.
  In general, accuracy may be improved by using non-linear, parameter-free models instead of parametric ones. However, we observed that even with models that are *linear in parameter*, especially polynomials, an accuracy of $R^2 > 0.95$ is achievable in some cases [25,26,91].
  As important as model selection is hyperparameter optimization as standard model settings usually yield insufficient accuracy [15]. A simple grid search may already yield a large increase in accuracy [106]. It is promising that recent publications relied on validated, sophisticated hyperparameter optimisation methods using state-of-the-art toolboxes (see Section 3).
- A frequently reported problem is the limited number of inputs a surrogate models can handle without exploding computational cost (*curse of dimensionality*).
  It is popular to integrate sensitivity analysis into the surrogate derivation process to determine the most important parameters. A surrogate model using only those inputs has proven to be accurate and to provide sufficiently optimised design options [106,107].
- While good model selection does not necessarily increase accuracy, it may increase sampling efficiency. [15] and [30] both found GP models to be sample efficient while RBF and MARS models require a lot of samples to reach high accuracy.
- The best choice of sampling algorithm is uncertain. Most studies within that review used latin hypercube sampling (see Table 2).
  No study exists which compares all static sampling schemes at once. In [135] Sobol's sampling was used for Monte Carlo simulation and provided more precise and robust output distributions than latin-hypercube and random sampling.
  Furthermore, a comparison of static and adaptive sampling in the field of building surrogate models is yet to be done. In other research domains adaptive sampling strategies have successfully shown to require less simulation runs until the surrogate reaches a certain accuracy [136].

6. Conclusion

This review provides a thorough discussion of publications that use surrogate models for sustainable building design.

The publications are sorted according to application area into conceptual design, sensitivity and uncertainty analysis, and building design optimisation. In particular, the use of surrogate models as a tool to give insight and understanding into high dimensional building design spaces was found to be popular in current research. Furthermore, multiple publications have shown that em-

bedding surrogate models into optimisation procedures accelerates the process significantly.

Apart from the analysis of research trends, this review serves as a practical guide. A detailed introduction to the process of deriving a surrogate model is given. The publications reviewed are categorized in both a large table and multiple figures providing a convenient technical overview of the field. Finally, practical aspects of all publications are summarized in a separate section with regard to model accuracy, model type, input selection and sampling strategy.

We expect future research to focus on lowering the computational cost for deriving a surrogate model and to increase the interpretability of models. The former could be achieved by implementing advanced sampling strategies, or by extending the scope of a single surrogate model from one to multiple buildings such that the derivation process does not have to be repeated for every analysis. Low interpretability can be avoided by compartmentalising surrogate models into multiple physically meaningful submodels.

Surrogate modelling has already been shown to lower the burden for architects and engineers to assess sustainable building designs using advanced performance analysis. We envisage, that it will play a key role in achieving sustainability in the future building stock.

Declaration of Competing Interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Acknowledgement

This work has been supported by grant funding from CANARIE via the BESOS project (CANARIE RS-327).

References

[1] R.K. Pachauri, M.R. Allen, V.R. Barros, J. Broome, W. Cramer, R. Christ, J.A. Church, L. Clarke, Q. Dahe, P. Dasgupta, et al., Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change, IPCC, 2014.
[2] I.E. Agency, Tracking Clean Energy Progress, 2017.
[3] Navigant Research, Global building stock database, commercial and residential building floor space by country and building type: 2017–2026,, 2017, https://www.navigantresearch.com/reports/global-building-stock-database.
[4] R.S. Michalski, J.G. Carbonell, T.M. Mitchell, Machine Learning: An Artificial Intelligence Approach, Springer Science & Business Media, 2013.
[5] M. Loyola, Big data in building design: a review, J. Inf. Technol. Constr.(ITcon) 23 (13) (2018) 259–284.
[6] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, X. Zhao, A review of data-driven approaches for prediction and classification of building energy consumption, Renew. Sustain. Energy Rev. 82 (2018) 1027–1047.
[7] A. Ahmad, M. Hassan, M. Abdullah, H. Rahman, F. Hussin, H. Abdullah, R. Saidur, A review on applications of ann and SVM for building electrical energy consumption forecasting, Renew. Sustain. Energy Rev. 33 (2014) 102–109.
[8] H. Lim, Z.J. Zhai, Comprehensive evaluation of the influence of meta-models on Bayesian calibration, Energy Build. 155 (2017) 66–75.
[9] C. Hjortling, F. Björk, M. Berg, T. af Klintberg, Energy mapping of existing building stock in sweden–analysis of data from energy performance certificates, Energy Build. 153 (2017) 341–355.
[10] S. Attia, E. Gratia, A. De Herde, J.L. Hensen, Simulation-based decision support tool for early stages of zero-energy building design, Energy Build. 49 (2012) 2–15.
[11] T. Østergård, R.L. Jensen, S.E. Maagaard, Building simulations supporting decision making in early design–a review, Renew. Sustain. Energy Rev. 61 (2016) 187–201. https://www.sciencedirect.com/science/article/pii/S136403211600280X.
[12] T. Simpson, V. Toropov, V. Balabanov, F. Viana, Design and analysis of computer experiments in multidisciplinary design optimization: a review of how far we have come-or not, in: 12th AIAA/ISSMO multidisciplinary analysis and optimization conference, 2008, p. 5802.
[13] G.G. Wang, S. Shan, Review of metamodeling techniques in support of engineering design optimization, J. Mech. Des. 129 (4) (2007) 370–380.
[14] A. Forrester, A. Keane, et al., Engineering Design via Surrogate Modelling: A Practical Guide, John Wiley & Sons, 2008.
[15] L. Van Gelder, P. Das, H. Janssen, S. Roels, Comparative study of metamodelling techniques in building energy simulation: guidelines for practitioners, Simul. Modell. Pract. Theory 49 (2014) 245–257, doi:10.1016/j.simpat.2014.10.004.
[16] T. Østergård, R.L. Jensen, S.E. Maagaard, A comparison of six metamodeling techniques applied to building performance simulations, Appl. Energy 211 (2018) 89–103.
[17] P. De Wilde, The gap between predicted and measured energy performance of buildings: a framework for investigation, Autom. Constr. 41 (2014) 40–49.
[18] G. Löhnert, A. Dalkowski, W. Sutter, Integrated design process: a guideline for sustainable and solar-optimised building design., International Energy Agency (IEA) Task 23 Optimization of Solar Energy Use in Large Buildings, subtask B. Austria., 2003.
[19] S. Petersen, Simulation-based support for integrated design of new low-energy office buildings, DTU Civil Engineering, Technical University of Denmark, 2011.
[20] S. Attia, J.L. Hensen, L. Beltrán, A. De Herde, Selection criteria for building performance simulation tools: contrasting architects' and engineers' needs, J. Build. Perform. Simul. 5 (3) (2012) 155–169, doi:10.1080/19401493.2010.549573.
[21] J. Hester, J. Gregory, R. Kirchain, Sequential early-design guidance for residential single-family buildings using a probabilistic metamodel of energy consumption, Energy Build. 134 (2017) 202–211, doi:10.1016/j.enbuild.2016.10.047.
[22] R.B. Miller, Response time in man-computer conversational transactions, in: Proceedings of the December 9–11, 1968, fall joint computer conference, part I, ACM, 1968, pp. 267–277.
[23] T. Ostergard, R.L. Jensen, S.E. Maagaard, Early building design: informed decision-making by exploring multidimensional design space using sensitivity analysis, Energy Build. 142 (2017) 8–22, doi:10.1016/j.enbuild.2017.02.059.
[24] J.P. Basbagill, F.L. Flager, M. Lepech, A multi-objective feedback approach for evaluating sequential conceptual building design decisions, Autom. Constr. 45 (2014) 136–150, doi:10.1016/j.autcon.2014.04.015.
[25] P. Geyer, A. Schlueter, Automated metamodel generation for design space exploration and decision-making - a novel method supporting performance-oriented building design and retrofitting, Appl. Energy 119 (2014) 537–556, doi:10.1016/j.apenergy.2013.12.064.
[26] F. Ritter, P. Geyer, A. Borrmann, Simulation-based decision-making in early design stages, in: 32nd CIB W78 conference, Eindhoven, The Netherlands, 2015, pp. 27–29.
[27] W. Tian, A review of sensitivity analysis methods in building energy analysis, Renew. Sustain. Energy Rev. 20 (2013) 411–419.
[28] S. Yang, W. Tian, E. Cubi, Q. Meng, Y. Liu, L. Wei, Comparison of sensitivity analysis methods in building energy assessment, Procedia Eng. 146 (2016) 174–181.
[29] B. Eisenhower, Z. O'Neill, V.A. Fonoberov, I. Mezić, Uncertainty and sensitivity decomposition of building energy models, J. Build. Perform. Simul. 5 (3) (2012) 171–184.
[30] T. Ostergard, R.L. Jensen, S.E. Maagaard, A comparison of six metamodeling techniques applied to building performance simulations, Appl. Energy 211 (2018) 89–103, doi:10.1016/j.apenergy.2017.10.102.
[31] C. Spitz, L. Mora, E. Wurtz, A. Jay, Practical application of uncertainty analysis and sensitivity analysis on an experimental house, Energy Build. 55 (2012) 459–470.
[32] L. Rivalin, P. Stabat, D. Marchio, M. Caciolo, F. Hopquin, A comparison of methods for uncertainty and sensitivity analysis applied to the energy performance of new commercial buildings, Energy Build. 166 (2018) 489–504.
[33] I.A. Macdonald, Quantifying the effects of uncertainty in building simulation, University of Strathclyde Glasgow, 2002 Ph.D. thesis. http://www.esru.strath.ac.uk/Documents/PhD/macdonald_thesis.pdf.
[34] S. de Wit, Uncertainty in building simulation, in: A. Malkawi, G. Augenbroe (Eds.), Advanced building simulation, Routledge, 2004, pp. 39–73.
[35] W. Tian, P. De Wilde, Uncertainty and sensitivity analysis of building performance using probabilistic climate projections: a UK case study, Autom. Constr. 20 (8) (2011) 1096–1109.
[36] W. Tian, Y. Heo, P. De Wilde, Z. Li, D. Yan, C.S. Park, X. Feng, G. Augenbroe, A review of uncertainty analysis in building energy assessment, Renew. Sustain. Energy Rev. 93 (2018) 285–301.
[37] C.J. Hopfe, J.L. Hensen, Uncertainty analysis in building performance simulation for design support, Energy Build. 43 (10) (2011) 2798–2805.
[38] Y. Heo, R. Choudhary, G. Augenbroe, Calibration of building energy models for retrofit analysis under uncertainty, Energy Build. 47 (2012) 550–560.
[39] M. Manfren, N. Aste, R. Moshksar, Calibration and uncertainty analysis for computer models - a meta-model based approach for integrated building energy simulation, Appl. Energy 103 (2013) 627–641, doi:10.1016/j.apenergy.2012.10.031.
[40] K.J. Lomas, H. Eppel, Sensitivity analysis techniques for building thermal simulation programs, Energy Build. 19 (1) (1992) 21–44.
[41] R. Evins, A review of computational optimisation methods applied to sustainable building design, Renew. Sustain. Energy Rev. 22 (2013) 230–245.
[42] A.-T. Nguyen, S. Reiter, P. Rigo, A review on simulation-based optimization methods applied to building performance analysis, Appl. Energy 113 (2014) 1043–1058.

[43] M. Wetter, J. Wright, A comparison of deterministic and probabilistic optimization algorithms for nonsmooth simulation-based optimization, Build. Environ. 39 (8) (2004) 989–999.
[44] M. Wetter, E. Polak, A convergent optimization method using pattern search algorithms with adaptive precision simulation, Build. Serv. Eng. Res.Technol. 25 (4) (2004) 327–338.
[45] T.W. Simpson, J. Poplinski, P.N. Koch, J.K. Allen, Metamodels for computer-based engineering design: survey and recommendations, Eng. Comput. 17 (2) (2001) 129–150.
[46] J. Free, A. Parkinson, G. Bryce, R. Balling, Approximation of computationally expensive and noisy functions for constrained nonlinear optimization, J. Mech. Transm.Autom. Des. 109 (4) (1987) 528–532.
[47] Y.S. Ong, P. Nair, A. Keane, K. Wong, Surrogate-assisted evolutionary optimization frameworks for high-fidelity engineering design problems, in: Knowledge Incorporation in Evolutionary Computation, Springer, 2005, pp. 307–331.
[48] S. Shan, G.G. Wang, Metamodeling for high dimensional simulation-based design problems, J. Mech. Des. 132 (5) (2010) 051009.
[49] F.A. Viana, T.W. Simpson, V. Balabanov, V. Toropov, Special section on multidisciplinary design optimization: metamodeling in multidisciplinary design optimization: how far have we really come? AIAA J. 52 (4) (2014) 670–690.
[50] D.B. Crawley, L.K. Lawrie, F.C. Winkelmann, W.F. Buhl, Y.J. Huang, C.O. Pedersen, R.K. Strand, R.J. Liesen, D.E. Fisher, M.J. Witte, et al., Energyplus: creating a new-generation building energy simulation program, Energy Build. 33 (4) (2001) 319–331.
[51] R. Yondo, E. Andrés, E. Valero, A review on design of experiments and surrogate models in aircraft real-time and many-query aerodynamic analyses, Prog. Aerosp. Sci. 96 (2018) 23–61, doi:10.1016/j.paerosci.2017.11.003.
[52] S.S. Garud, I.A. Karimi, M. Kraft, Design of computer experiments: a review, Comput. Chem. Eng. 106 (2017) 71–95.
[53] A. Bhosekar, M. Ierapetritou, Advances in surrogate based modeling, feasibility analysis, and optimization: a review, Comput. Chem. Eng. 108 (2018) 250–267.
[54] J. Močkus, On Bayesian methods for seeking the extremum, in: Optimization Techniques IFIP Technical Conference, Springer, 1975, pp. 400–404.
[55] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., 2013.
[56] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, 112, Springer, 2013.
[57] L. Breiman, Statistical modeling: the two cultures, Stat. Sci. 16 (3) (2001) 199–231.
[58] J.H. Friedman, Multivariate adaptive regression splines, Ann. Stat. (1991) 1–67.
[59] S. Chen, C.F. Cowan, P.M. Grant, Orthogonal least squares learning algorithm for radial basis function networks, IEEE Trans. Neural Netw. 2 (2) (1991) 302–309.
[60] C.E. Rasmussen, Gaussian processes in machine learning, in: Advanced lectures on machine learning, Springer, 2004, pp. 63–71.
[61] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (3) (2004) 199–222.
[62] M. Claesen, B. De Moor, Hyperparameter search in machine learning, arXiv preprint arXiv:1502.02127.
[63] J. Müller, Matsumoto: the matlab surrogate model toolbox for computationally expensive black-box global optimization problems, arXiv preprint arXiv:1404.4261.
[64] D. Gorissen, I. Couckuyt, P. Demeester, T. Dhaene, K. Crombecq, A surrogate modeling and adaptive sampling toolbox for computer based design, J. Mach. Learn. Res. 11 (Jul) (2010) 2051–2055.
[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (Oct) (2011) 2825–2830.
[66] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: a system for large-scale machine learning., in: OSDI, 16, 2016, pp. 265–283.
[67] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, 2017. https://github.com/pytorch.
[68] T. Wortmann, Opossuma model-based optimization tool in grasshopper, in: Proceedings of the 22nd CAADRIA Conference, Hong Kong, CN, 2017, pp. 283–292.
[69] A. Costa, G. Nannicini, Rbfopt: an open-source library for black-box optimization with costly function evaluations, Optim. Online 4538 (2014).
[70] T. Olofsson, S. Andersson, Long-term energy demand predictions based on short-term measured data, Energy Build. 33 (2) (2001) 85–91.
[71] B.B. Ekici, U.T. Aksoy, Prediction of building energy consumption by using artificial neural networks, Adv. Eng. Softw. 40 (5) (2009) 356–362.
[72] A. Prada, A. Gasparella, P. Baggio, On the performance of meta-models in building design optimization, Appl. Energy 225 (2018) 814–826.
[73] S. Bucking, R. Zmeureanu, A. Athienitis, A methodology for identifying the influence of design variations on building energy performance, J. Build. Perform. Simul. 7 (6) (2014) 411–426.
[74] T. Ahmad, H. Chen, Y. Guo, J. Wang, A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: a review, Energy Build. 165 (2018) 301–320.
[75] N.V. Queipo, R.T. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, P.K. Tucker, Surrogate-based analysis and optimization, Prog. Aerosp. Sci. 41 (1) (2005) 1–28.
[76] E. Alpaydin, Introduction to Machine Learning, MIT press, 2009.
[77] E. Gratia, A. De Herde, A simple design tool for the thermal study of an office building, Energy Build. 34 (3) (2002) 279–289.
[78] R.E. Edwards, J. New, L.E. Parker, B. Cui, J. Dong, Constructing large scale surrogate models from big data and artificial intelligence, Appl. Energy 202 (2017) 685–699, doi:10.1016/j.apenergy.2017.05.155.
[79] H. Yi, R.S. Srinivasan, W.W. Braham, An integrated energy-emergy approach to building form optimization: use of energy plus, emergy analysis and taguchi-regression method, Build. Environ. 84 (2015) 89–104, doi:10.1016/j.buildenv.2014.10.013.
[80] H.T. Odum, Environmental Accounting: Energy and Environmental Decision Making, Wiley, 1996.
[81] L.-G. Maltais, L. Gosselin, Daylighting 'energy and comfort' performance in office buildings: sensitivity analysis, metamodel and pareto front, J. Build. Eng. 14 (2017) 61–72, doi:10.1016/j.jobe.2017.09.012.
[82] P. Geyer, S. Singaravel, Component-based building performance prediction using systems engineering and machine learning, Appl. Energy 228 (2017) 1439–1453.
[83] F. Chlela, A. Husaunndee, C. Inard, P. RiedeFer, A new methodology for the design of low energy buildings, Energy Build. 41 (9) (2009) 982–990, doi:10.1016/j.enbuild.2009.05.001.
[84] I. Korolija, Y. Zhang, L. Marjanovic-Halburd, V.I. Hanby, Regression models for predicting UK office building energy consumption from heating and cooling demands, Energy Build. 59 (2013) 214–227.
[85] T. Catalina, J. Virgone, E. Blanco, Development and validation of regression models to predict monthly heating demand for residential buildings, Energy Build. 40 (10) (2008) 1825–1832.
[86] T. Catalina, V. Iordache, B. Caracaleanu, Multiple regression model for fast prediction of the heating energy demand, Energy Build. 57 (2013) 302–312.
[87] J.S. Hygh, J.F. DeCarolis, D.B. Hill, S.R. Ranjithan, Multivariate regression as an energy assessment tool in early building design, Build. Environ. 57 (2012) 165–175, doi:10.1016/j.buildenv.2012.04.021.
[88] J.C. Lam, K.K. Wan, D. Liu, C. Tsang, Multiple regression models for energy use in air-conditioned office buildings in different climates, Energy Convers. Manag. 51 (12) (2010) 2692–2697.
[89] I. Jaffal, C. Inard, C. Ghiaus, Fast method to predict building heating demand based on the design of experiments, Energy Build. 41 (6) (2009) 669–677.
[90] A. Rackes, A.P. Melo, R. Lamberts, Naturally comfortable and sustainable: informed design guidance and performance labeling for passive commercial buildings in hot climates, Appl. Energy 174 (2016) 256–274, doi:10.1016/j.apenergy.2016.04.081.
[91] Z. Romani, A. Draoui, F. Allard, Metamodeling the heating and cooling energy needs and simultaneous building envelope optimization for low energy building design in Morocco, Energy Build. 102 (2015) 139–148, doi:10.1016/j.enbuild.2015.04.014.
[92] F. Ascione, N. Bianco, C. De Stasio, G.M. Mauro, G.P. Vanoli, Artificial neural networks to predict energy performance and retrofit scenarios for any member of a building category: a novel approach, Energy 118 (2017) 999–1017, doi:10.1016/j.energy.2016.10.126.
[93] I. Jaffal, C. Inard, A metamodel for building energy performance, Energy Build. 151 (2017) 501–510, doi:10.1016/j.enbuild.2017.06.072.
[94] S. Singaravel, J. Suykens, P. Geyer, Deep-learning neural-network architectures and methods: using component-based models in building-design energy prediction, Adv. Eng. Inform. 38 (2018) 81–90.
[95] J.J. Hirsch, Associates, equest - the quick energy simulation tool, 2016, http://www.doe2.com/eQUEST/.
[96] J. Dhariwal, R. Banerjee, An approach for building design optimization using design of experiments, Build. Simul. 10 (3) (2017) 323–336, doi:10.1007/s12273-016-0334-z.
[97] C.J. Hopfe, M.T. Emmerich, R. Marijt, J. Hensen, Robust multi-criteria design optimisation in building design, in: Proceedings of building simulation and optimization, Loughborough, UK, 2012, pp. 118–125.
[98] Y.-J. Kim, Comparative study of surrogate models for uncertainty quantification of building energy model: gaussian process emulator vs. polynomial chaos expansion, Energy Build. 133 (2016) 46–58, doi:10.1016/j.enbuild.2016.09.032.
[99] P. Das, C. Shrubsole, B. Jones, I. Hamilton, Z. Chalabi, M. Davies, A. Mavrogianni, J. Taylor, Using probabilistic sampling-based sensitivity analyses for indoor air quality modelling, Build. Environ. 78 (2014) 171–182, doi:10.1016/j.buildenv.2014.04.017.
[100] G. Li, S.-W. Wang, H. Rabitz, S. Wang, P. Jaffé, Global uncertainty assessments by high dimensional model representations (HDMR), Chem. Eng. Sci. 57 (21) (2002) 4445–4460.
[101] A. Tsanas, A. Xifara, Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools, Energy Build. 49 (2012) 560–567, doi:10.1016/j.enbuild.2012.03.003.
[102] P. De Wilde, W. Tian, Predicting the performance of an office under climate change: a study of metrics, sensitivity and zonal resolution, Energy Build. 42 (10) (2010) 1674–1684.
[103] W. Tian, R. Choudhary, G. Augenbroe, S.H. Lee, Importance analysis and metamodel construction with correlated variables in evaluation of thermal performance of campus buildings, Build. Environ. 92 (2015) 61–74, doi:10.1016/j.buildenv.2015.04.021.
[104] V. Zuber, K. Strimmer, High-dimensional regression and variable selection using car scores, Stat. Appl. Genet. Mol.Biol. 10 (1) (2011).

[105] X. Chen, H. Yang, K. Sun, Developing a meta-model for sensitivity analyses and prediction of building performance for passively designed high-rise residential buildings, Appl. Energy 194 (2017) 422–439, doi:10.1016/j.apenergy.2016.08.180.

[106] X. Chen, H. Yang, Integrated energy performance optimization of a passively designed high-rise residential building in different climatic zones of china, Appl. Energy 215 (2018) 145–158, doi:10.1016/j.apenergy.2018.01.099.

[107] B. Eisenhower, Z. O'Neill, S. Narayanan, V.A. Fonoberov, I. Mezic, A methodology for meta-model based optimization in building energy models, Energy Build. 47 (2012) 292–301, doi:10.1016/j.enbuild.2011.12.001.

[108] S. Papadopoulos, E. Azar, Integrating building performance simulation in agent-based modeling using regression surrogate models: a novel human-in-the-loop energy modeling approach, Energy Build. 128 (2016) 214–223, doi:10.1016/j.enbuild.2016.06.079.

[109] I.A. Macdonald, Comparison of sampling techniques on the performance of Monte Carlo based sensitivity analysis, in: Eleventh International IBPSA Conference, 2009, pp. 992–999.

[110] S. Burhenne, O. Tsvetkova, D. Jacob, G.P. Henze, A. Wagner, Uncertainty quantification for combined building performance and cost-benefit analyses, Build. Environ. 62 (2013) 143–154.

[111] K. Soratana, J. Marriott, Increasing innovation in home energy efficiency: Monte Carlo simulation of potential improvements, Energy Build. 42 (6) (2010) 828–833.

[112] B.D. Lee, Y. Sun, G. Augenbroe, C.J. Paredis, Towards better prediction of building performance: a workbench to analyze uncertainty in building simulation, in: 13th International Building Performance Simulation Association Conference, Chambéry, France, 2013.

[113] S.L. Wong, K.K.W. Wan, T.N.T. Lam, Artificial neural networks for energy analysis of office buildings with daylighting, Appl. Energy 87 (2) (2010) 551–557, doi:10.1016/j.apenergy.2009.06.028.

[114] L. Magnier, F. Haghighat, Multiobjective optimization of building design using trnsys simulations, genetic algorithm, and artificial neural network, Build. Environ. 45 (3) (2010) 739–746, doi:10.1016/j.buildenv.2009.08.016.

[115] E. Asadi, M.G. da Silva, C.H. Antunes, L. Dias, L. Glicksman, Multi-objective optimization for building retrofit: a model using genetic algorithm and artificial neural network and an application, Energy Build. 81 (2014) 444–456, doi:10.1016/j.enbuild.2014.06.009.

[116] E.E. Aydin, O. Dursun, I. Chatzikonstantinou, B. Ekici, Optimisation of energy consumption and daylighting using building performance surrogate model, in: Living and Learning: Research for a Better Built Environment: 49th International Conference of the Architectural Science Association, 2015, pp. 536–546. ⟨GotoISI⟩://WOS:000381380100052.

[117] X. Chen, H. Yang, A multi-stage optimization of passively designed high-rise residential buildings in multiple building operation scenarios, Appl. Energy 206 (2017) 541–557, doi:10.1016/j.apenergy.2017.08.204.

[118] P.T. Boggs, J.W. Tolle, Sequential quadratic programming, Acta Numerica 4 (1995) 1–51.

[119] G.M. Stavrakakis, P.L. Zervas, H. Sarimveis, N.C. Markatos, Optimization of window-openings design for thermal comfort in naturally ventilated buildings, Appl. Math. Modell. 36 (1) (2012) 193–211, doi:10.1016/j.apm.2011.05.052.

[120] J. Carreras, C. Pozo, D. Boer, G. Guillen-Gosalbez, J.A. Caballero, R. Ruiz-Femenia, L. Jimenez, Systematic approach for the life cycle multi-objective optimization of buildings combining objective reduction and surrogate modeling, Energy Build. 130 (2016) 506–518, doi:10.1016/j.enbuild.2016.07.062.

[121] E. Gengembre, B. Ladevie, O. Fudym, A. Thuillier, A Kriging constrained efficient global optimization approach applied to low-energy building design problems, Inverse Probl. Sci. Eng. 20 (7) (2012) 1101–1114, doi:10.1080/17415977.2012.727084.

[122] D.R. Jones, M. Schonlau, W.J. Welch, Efficient global optimization of expensive black-box functions, J. Global Optim. 13 (4) (1998) 455–492.

[123] E. Tresidder, Y. Zhang, A.I. Forrester, Acceleration of building design optimisation through the use of Kriging surrogate models, Proc.Build.Simul.Optim. (2012) 1–8.

[124] R. Zhang, F. Liu, A. Schoergendorfer, Y. Hwang, Y.M. Lee, J.L. Snowdon, Optimal selection of building components using sequential design via statistical surrogate models, in: Proceedings of Building Simulation, 2013, pp. 2584–2592.

[125] S.S. Gilan, N. Goyal, B. Dilkina, Acm, Active learning in multi-objective evolutionary algorithms for sustainable building design, in: Gecco'16: Proceedings of the 2016 Genetic and Evolutionary Computation Conference, 2016, pp. 589–596, doi:10.1145/2908812.2908947.

[126] N. Beume, B. Naujoks, M. Emmerich, Sms-emoa: multiobjective selection based on dominated hypervolume, Eur. J. Oper. Res. 181 (3) (2007) 1653–1669.

[127] G. Zemella, D. De March, M. Borrotti, I. Poli, Optimised design of energy efficient building facades via evolutionary neural networks, Energy Build. 43 (12) (2011) 3297–3302, doi:10.1016/j.enbuild.2011.10.006.

[128] W. Xu, A. Chong, O.T. Karaguzel, K.P. Lam, Improving evolutionary algorithm performance for integer type multi-objective building system design optimization, Energy Build. 127 (2016) 714–729, doi:10.1016/j.enbuild.2016.06.043.

[129] A.E.I. Brownlee, J.A. Wright, Constrained, mixed-integer and multi-objective optimisation of building designs by NSGA-II with fitness approximation, Appl. Soft Comput. 33 (2015) 114–126, doi:10.1016/j.asoc.2015.04.010.

[130] T. Wortmann, Genetic evolution vs. function approximation: benchmarking algorithms for architectural design optimization, J. Comput. Des. Eng. (2018), doi:10.1016/j.jcde.2018.09.001.

[131] L. Van Gelder, H. Janssen, S. Roels, Metamodelling in robust low-energy dwelling design, in: 2nd Central European Symposium on Building Physics, Mahdavi, A, OKK-EDITIONS, Vienna, 2013, pp. 93–99.

[132] M.-Y. Cheng, M.-T. Cao, Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines, Appl. Soft Comput. 22 (2014) 178–188, doi:10.1016/j.asoc.2014.05.015.

[133] J.-S. Chou, D.-K. Bui, Modeling heating and cooling loads by artificial intelligence for energy-efficient building design, Energy Build. 82 (2014) 437–446, doi:10.1016/j.enbuild.2014.07.036.

[134] P. Symonds, J. Taylor, Z. Chalabi, M. Davies, Performance of neural networks vs. radial basis functions when forming a metamodel for residential buildings, Int. J. Civil Environ. Struct. Constr. Archit. Eng. 9 (12) (2015) 1446–1450. World Academy of Science, Engineering and Technology.

[135] S. Burhenne, D. Jacob, G.P. Henze, Sampling based on Sobol' sequences for monte carlo techniques applied to building simulations, in: Proc. Int. Conf. Build. Simulat, 2011, pp. 1816–1823.

[136] S.S. Garud, I.A. Karimi, M. Kraft, Smart sampling algorithm for surrogate model development, Comput. Chem. Eng. 96 (2017) 103–114.

Part I

Surrogate modelling for design

Chapter 3

Example of a surrogate model in use.

The literature review has shown that instantaneous feedback at the early design stage is a very promising application of a surrogate model. A platform for this is being developed by the Energy in Cities group, as presented in the following paper. The tool is meant to guide building designers towards net-zero energy buildings.
The trained surrogate models are hosted on an interactive web platform. They incorporate key findings from the literature review, where the following aspects to improve the state-of-the-art are considered:

- The surrogate model uses a large set of input parameters (32) which lets users model a large set of design problems, including the ability to model various heating, cooling and ventilation systems, daylighting controls, the impact of running a server in an office building and other key factors.
- A large number of outputs are modelled to assess net energy demand, including a break down of the different end-uses and photovoltaic generation potential.
- An analysis of the impact of the training set size on the performance of the surrogate model is performed.
- An updated set of normalized error metrics that allow comparing the approximation accuracy of various outputs is given.

Net-Zero Navigator: A platform for interactive net-zero building design using surrogate modelling

Paul Westermann[1], David Rulff[1], Kevin Cant[1], Gaelle Faure[1], Ralph Evins[1]
[1]Energy in Cities group, Department of Civil Engineering, University of Victoria, Canada

Abstract

The design of high-performance buildings requires rapid iteration over many highly-integrated choices. These must be made early the design process, as certain performance targets like net-zero energy consumption may not be achievable at later stage. Existing high-resolution simulation approaches are not easily able to deliver fast, integrated design iterations.

In this paper we introduce the Net-Zero Navigator, an open-source platform for conceptual building performance design based on surrogate modelling. It leverages state-of-the-art machine learning techniques to provide surrogate models which emulate high-fidelity building performance simulation results, providing accurate design performance estimates instantly. Results are given to quantify the performance of the surrogate modelling approach, which achieved R^2 values of over 96%.

The platform builds on a suite of existing software tools (EnergyPlus, TensorFlow, KERAS API) as well as the codebase of the Building and Energy Simulation, Optimization and Surrogate-modelling (BESOS) platform. Overall, the Net-Zero Navigator platform provides a fast, interactive way to undertake concept-stage building design.

Introduction

Performance-based building design

In performance-based building design, energy metrics are used to quantify how much a design fulfils single or multiple design objectives (Kalay, 1999). This paradigm spread rapidly in recent years and as a result there are a proliferation of building energy simulation software options to address this need (Attia, 2010).

Energy performance analysis tools are used throughout the design process, from conceptual to detailed design (Östman, 2005). The Net-Zero Navigator is currently targeted to support guidance in the early design stages, where the design is most flexible and decisions have the highest impact on final performance. At this stage, absolute accuracy can be sacrificed in favour of flexibility, breadth and speed of design space exploration. this fits well with the surrogate-modelling based approach.

Conceptual design analysis tools

Existing energy simulation software tools apply building physics equations with varying level of detail. More detailed tools require careful building model implementation and potentially have long simulation run times. This mismatches the need for fast feedback during the highly dynamic process of the early design stage, where multiple, strongly differing design concepts are to be explored (Petersen, 2011).

For that reason a subfield of tools is developing which aims at low computational cost, limited number of user input requirements, and a high degree of interactivity. The underlying performance estimates are either based on *simplified physics* models or collected from *a database of pre-run parametric simulations*. Recently, the use of statistical simulation *surrogate models*, also called emulators or meta-models, gained increasing attention (Westermann and Evins, 2019).

Examples include tools from Nielsen (2005) who developed a dynamic, single-thermal-zone, lumped parameter model to estimate energy demand and indoor comfort, and from Gratia and De Herde (2002) who developed OPTI, a tool which provides annual thermal needs and thermal comfort estimates by accessing a database of pre-run simulations.

In comparison to the previous methods, Ritter et al. (2015) used a statistical emulator of a detailed, dynamic simulation tool to provide performance estimates in their DSEAM tool. Although the model is a rather simplistic second order linear regression model fitted to simulation output data it has multiple advantages over the approaches mentioned above.

First, no simplification of building physics is done, and only a small error of the statistical model is introduced. Given recent advances in machine learning even very complex physical phenomena can be emulated (Kasim et al., 2020). Second, no parametric data must be stored and the tool uses a continuous function providing the user with performance estimates for a continuous set of a large set of design inputs. The latter in particular allows design space exploration and "recognizes that different (building) forms can successfully achieve similar functions" and performances (Kalay, 1999).

When we analyse design spaces spanned by large number of parameters, advanced visualization techniques are required. Ritter et al. use parallel-

coordinate plots (PCP) which were also found popular in industry with multiple trials to develop interactive tools to building designers [1].

Contributions of this paper

The Net-Zero Navigator will be the first tool to allow building designers who are not machine learning experts to leverage advancements in surrogate modelling to drastically improve their ability to find high performance building designs.
The platform features a holistic set of building archetype surrogate models, which can be accessed using intuitive visualizations hosted on a web-platform. Beyond that, an advanced user interface allows users to access the underlying codebase and Jupyter Notebooks, such that the process of surrogate modelling and visualiziation of building design spaces can be customized if needed.
Computational optimization of buildings has been an increasing focus of research in recent years, as detailed in Evins (2013). The Net-Zero-Navigator (NZN) platform builds upon this work to deliver a practical, usable tool for design space exploration, which was often the underlying intent of many optimization studies.
The NZN platform goes beyond existing Canadian energy compliance tools (CANQuest, HOT2000 etc.) by providing rigorous simulation, multi-objective optimization, and visual exploration of design implications. It goes beyond previous simplified compliance guidance (e.g. screeningtool.ca) by providing much greater breadth and detail in results exploration. It goes beyond HTAP and BTAP (developed by Canmet) in providing surrogate modelling coupled with a user-friendly interface.
The Net-Zero Navigator is the first platform to provide visual exploration and optimization of buildings in an accessible online environment. It spearheads the application of modern computational techniques (cloud computing, machine learning, interactive visualization) to this domain.

The platform

The platform is composed of a *core* and an *advanced interfaces*. The software architecture diagram is presented in Figure 1.
The main features and the way they work together is detailed below. EnergyPlus simulations capture the whole building-level interactions between different elements and systems. Results are used to generate surrogate models (fitted statistical representations of optimization data) for effective deployment as an online platform. Optimization algorithms can then be applied to find synergies across many objectives. A prototype advanced interface allows optimization and exploration using user-uploaded models, calibration of models to measured building data prior to exploration, and user data analytics. The platform combines the best available technologies in each area to provide a user-friendly, open-source interface for optimization and visualization.

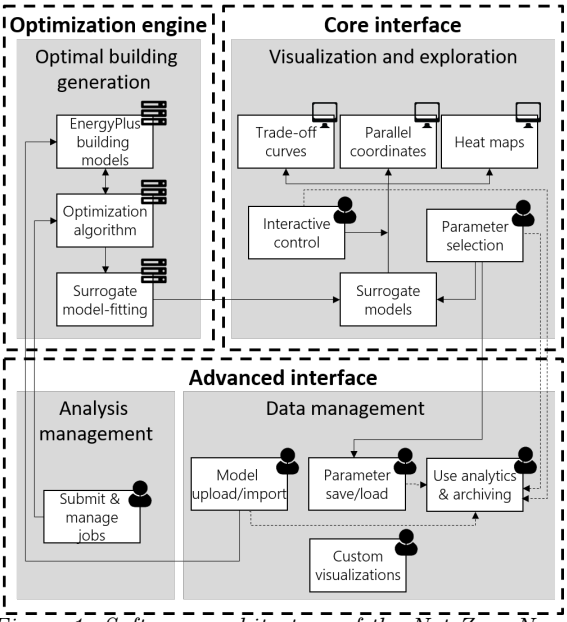


Figure 1: Software architecture of the Net-Zero Navigator platform.

The core methodologies employed are detailed building energy simulation, multi-objective optimization, surrogate modelling and web-based interactive visual data exploration, brought together through an API-based modular software implementation, as shown in the software architecture diagram in Figure 1. Icons indicate user interface modules, cluster-based computation modules, and visualization modules.
The interfaces as well as the process to fit surrogate models are described in more detail in the following sections.

User Centered Design

A core objective of the NZN project is the practical integration of surrogate modelling methods into the work processes of industry user groups and stakeholders, creating a platform to leverage the domain-knowledge of building designers and planners using more robust parametric modelling tool sets. This involves developing a refined interface for defining simulation parameters, interrogating sub-model analyses, and visualizing surrogate output. Feedback and testing by users will be central to the iterative refinement of the methods and development of an interface for practical application of these surrogate modelling modules.
Prototype interface and visualizations will be developed in JavaScript in order to interface easily with the computational modules developed in for the underlying surrogate modelling. The interface design will be refined through an iterative process

[1] Building Pathfinder (`http://www.buildingpathfinder.com/`), BTAP (`https://canmet-energy.github.io/parallel-coordinates/`), Canmet-Energy (2020)

involving user-group testing. The overall methodology and module structures will be documented and reported using user interface design and open source best practices. This will include development and implementation of training modules, workshops and information sessions that will aim to expand the user base and foster a collaborative environment for continued development of the tools and methods.
A summary of the intended user definitions, functionality and outputs are shown in Figure 2.

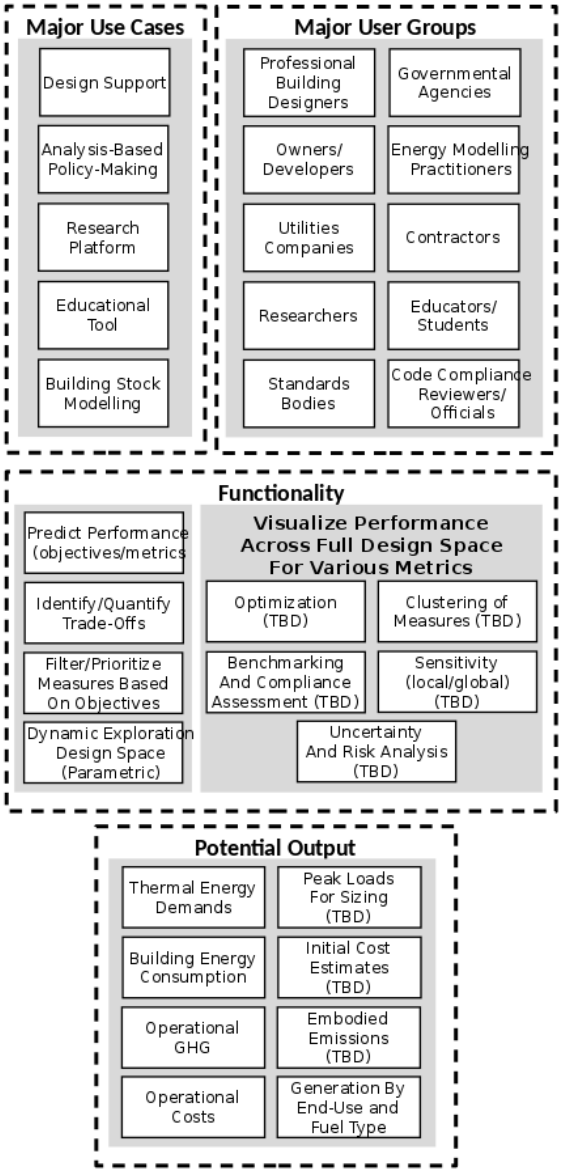


Figure 2: Summary of Intended User Definitions, Functionality and Outputs

Due to the open and adaptive development structure of NZN, the overall scope of the NZN project encompasses a broad range of potential user groups and use cases. To address the gaps in conventional building simulation practice, and to leverage the surrogate modelling framework developed for BESOS, the design focus of the NZN platform divided between two main interface modes: *core* and *advanced*.

Core Interface

The *core interface* provides interactive visualization for design exploration. The surrogate models resulting from model-fitting are provided via a cloud-based service to allow real-time optimization via an online interactive visualization interface. This core tool will allow designers to refine parameters and observe the impacts in real time, quickly gaining an understanding of the best paths to high performance for a given building type and context.
The core interface is intended to serve segments of the user group interested in quickly exploring the building design space for a predefined set of building archetypes (based on the representation in typical Canadian building stock), providing feedback on performance under selected metrics. Inputs are intuitive and adaptive, requiring minimal time from the user to generate meaningful results. Among the driving design principles for the core interface are interactivity, adaptability and intuitiveness: providing a comprehensive (but curated) range of functionality, while enabling the user to hide unnecessary information and controls, interact directly with the inputs and results, and dynamically hone in on their primary objectives.

Context variables

A variety of contextual inputs for these archetypes can be modified by the user to customize the building to suit their needs. Areas of control influencing building performance include building program (occupancy, space types, operational schedules, etc.), building geometry (floor area, height, aspect ratios, etc.), and weather (location, current or future climate conditions, etc.). Additional inputs for life cycle analysis will also be available including initial impacts (capital costs, embodied carbon, etc.), operational costs (fuel prices, emissions intensities, etc.), and financial assumptions (escalation rates, debt leveraging, interest rates, etc.). The full set of contextual inputs will be refined through exhaustive sensitivity analysis and user testing.

Measures and Parameters

The user can then choose a subset of measures for consideration – to evaluate their impact on selected metrics and outputs – adding dimensions to the problem space. The combination of the baseline building definition and measures defines the overall design space, which is constructed on the underlying surrogate models that make up the engine of NZN. Interactive modules will be incorporated that enable a variety of functionality, including assessment of specific design configurations, filtering and adjustment of

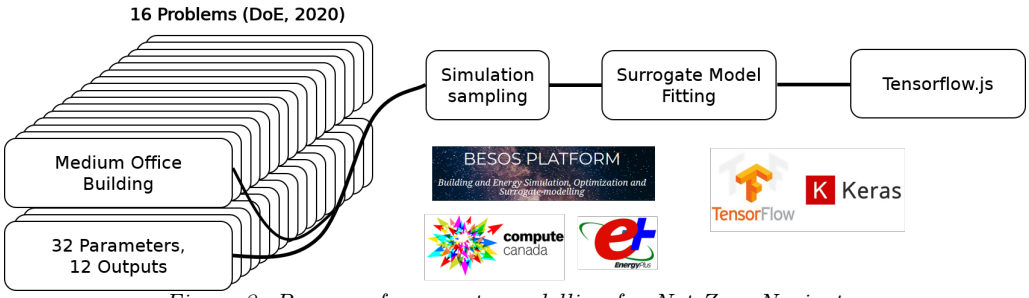


Figure 3: Process of surrogate modelling for Net-Zero Navigator

measures, or optimization based on selected metrics (such as operational energy, emissions or cost).

Outputs and Visualization

Outputs may include pre-set energy reports, data files, LCC assessment summaries, or peak sizing information for design support. Users will have the option to export EnergyPlus IDFs for specific design configurations for more explicit analysis. Results will be dynamically updated and presented to the user through interactive widgets. Visualization options include parallel coordinates plots, star diagrams, and conventional graphs. These components of interface design will be refined through an iterative process that will involve prototyping, user testing and stakeholder engagement.
The tool is designed to account for any combination of user inputs and default assumptions. The surrogate models used for the core interface are developed to provide a robust foundation for rapid, exhaustive design space exploration, and accuracy has been shown to be sufficient for this purpose; however, more detailed customization is possible in the advanced interface, which can be used to improve precision.

Advanced Interface

An *advanced interface* is proposed for energy modelling professionals as well as researchers. It provides a means to directly generate and interrogate underlying EnergyPlus and surrogate models, providing access to the full functionality of BESOS.
An advanced interface is proposed for energy modelling professionals as well as researchers. It provides the following functionalities:

- import/export EnergyPlus models,
- modify and sample EnergyPlus models,
- fit a surrogate model,
- optimize the design of a building according to various criteria,
- analyse the results using the same advanced visualization tools as for the core interface,
- customize its own visualization tool.

The advanced interface also allows custom metrics to be implemented, letting policy-makers explore the impact of measures across the design space.
All the features are accessible through Jupyter Notebooks and Python code. Moreover Net-Zero Navigator advanced interface benefits from the same functionalities as BESOS, also developed by our team, allowing an easy share of the work and an access to Compute Canada supercomputer to fasten the calculations.

Surrogate modelling procedure

The surrogate modelling procedure underlying the Net-Zero Navigator platform is shown in Figure 3. It consists of

- defining the design problem, i.e. the free design parameters and the design objectives,
- running simulation samples,
- fitting the surrogate model and exporting it as a TensorflowJS object, such that the model can be embedded into the NZN browser application.

The core objective of the process is to derive a surrogate model with a large application scope. This allows one surrogate to cover many design problems. The problem definition stage is a crucial in that regard. We approach it by collecting a holistic set of design parameters and performance metrics in an iterative process. We integrate form (window-to-wall ratio, building storeys, overhangs, orientation), materials (U-Values, Solar-Heat-Gain-coefficient, thickness), loads (people, plug, server, lighting), controls (set point schedules, area covered by daylighting sensors) and HVAC parameters (air rate, pumping rate, plant performance, DHW, heat recovery eff.).
The design parameters span a large combinatorial space. We explore it by collecting building performance simulation samples from within that space, where we aim to maximise the information gain per simulation run. Therefore, we use Latin-Hypercube sampling (LHS). The platform offers all tools for surrogate derivation, including a Python EnergyPlus API (based on EPPy), and access to computational hardware from Compute Canada. The latter enables us to run thousands of simulations within a reasonable amount of time, and to leverage GPU resources

to train deep neural network surrogate models.
To train the surrogate model, we use the machine learning toolbox Keras, which uses TensorFlow as a backend. TensorFlow is a toolbox specifically designed to train neural networks. In the NZN core interface, we are mostly concerned with predicting aggregated annual performance metrics, not time series results. This allows us to limit the complexity of the networks to relatively shallow, multi-output feed-forward neural nets ($\leq$ 3 hidden layers, $\leq$ 512 neurons per layer), which have proven to emulate aggregated simulation outcomes well (Westermann and Evins, 2019).
Once the model is trained and its accuracy validated using simulation runs not included in the training data, we export the neural network architecture and trained weights as a JavaScript file to be embedded into the NZN browser-based application, which uses TensorFlow.js to perform all surrogate model evaluations directly in the user's web-browser.
In the next section we go through all the steps above for an example case which is present in the platform, and investigate the accuracy of the surrogate model. NZN also allows the user to go through these steps for bespoke problems using the *advanced interface*.

Example case

In this section, we explain how we derive a surrogate model for one of the 16 DoE archetype buildings and explain how the user can interact with it. We briefly describe the base building model (a medium-sized office (Canmet-Energy, 2020)), provide accuracy estimates of the surrogate model for each individual performance objective, and present the parallel coordinate plot as example of a visualization tools.

Building details, parameters and objectives

The baseline building definition presented in this research leveraged the work of National Resources Canada (NRCan), who have developed a platform called BTAP to generate NECB versions of the Commercial Prototype Building Models originally created by Canmet-Energy (2020). For the purposes of this illustrative case study, the Medium Office archetype was selected.
Baseline assumptions for building program, space loads, basic controls, geometry and enclosure performance were directly derived from the NECB 2015 requirements. An alternative approach was taken to represent the mechanical systems, with the intention of capturing a wide variety of configurations and parameters through direct manipulation of air-side system and plant equipment performance in the EnergyPlus Energy Management System. This allows high-level exploration of a vast HVAC system design space through variation of a subset of core parameters, set up independently of any specific proposed design, while maintaining a consistent basis for generating building demands.

Parameter List Problem 1, Thermal Loads

1.-2.	Wall, Window U-Value
3.	Window SHGC
4.-5.	Infiltration Rate, Ventilation Rate
5.	Horizontal Shading Depth
6.	Thermal Mass
7.	Daylighting Sensors
8.	People
9.-11.	Plug, Lighting, Server-Roam Loads
12.-15.	Humidity, Temp. Setpoint (min, max)
16.-19.	WWR (North, East, South, West)
20.	Orientation
21.	Number of Storeys
22.-23.	Heat Recovery (Sensible, Latent)
23.-25.	Plant Performance (Heating-Fuel, Heating-Elec., Cooling)
26.-27.	Fan-Power (Air Rate, Conditioning Share)
28.-29.	Heating Plant Fuel Mix (Share of Total, Biofuel Share)
30.-31.	Pumping (Rate, Hydronic Share)
32.	Domestic Hot Water (Share)

Table 1: List of design parameters for Problem 1, Medium office)

Output List Problem 1

1.	Heating Supply, Gas	**7.**	Interior Lights
2.	Heating Supply, Electricity	**8.**	Pumps Power
3.	Heating Supply, Other	**9.**	Fan Power
4.	Cooling Supply, Electricity	**10.**	PV Generation
5.	Water Heating, Gas	**11.**	Heating Demand
6.	Interior Equipment	**12.**	Cooling Demand

Table 2: List of surrogate model outputs for Problem 1, Medium office)

Sampling

We generate 10,000 simulation samples using the widely applied latin-hypercube sampling, which stratifies the design parameter space into equally large hypercubes and randomly collects samples from within each hypercube. We run the 10,000 samples on an HPC cluster (250 jobs with 1 CPU and 2Gb RAM running for 2 hours each).
We fit a 2-layered feed-forward neural network on the retrieved simulation results using the Adam optimizer minimizing the mean-squared-error (MSE), where we specified the learning rate schedule parameters with $\beta_1 = 0.9$, $\beta_2 = 0.999$. Other training hyperparameters were determined in a grid search (l2-

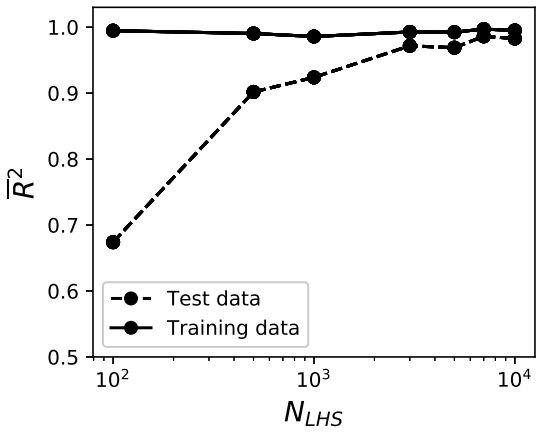


Figure 4: Number of samples vs. surrogate model accuracy.

regularization coefficient α, number of neurons per layer $n_{neurons}$). To control for the variation introduced by random weight normalization, we repeat each run of the grid search three times. For more details on feed-forward neural network training see (Bishop, 2006).

In this section, we only present the surrogate modelling performance for one of the 16 considered buildings. To efficiently run simulations for all other archetype buildings, we investigated the number of samples required to reach a satisfying surrogate model performance. In Figure 4 the accuracy as a function of the number of samples is shown. On the y-axis we plot the R^2 score, averaged over all outputs, and on the x-axis we plot the number of simulation samples, where 80% is used for training and 20% for testing and hence, corresponds to the total number of samples required to train and validate a surrogate model.The plot highlights that the increase in accuracy levels off after $N_{LHS} = 3000$ samples. However, it keeps increasing slightly for higher numbers of samples. We suggest that more than 100 samples per parameter are required, which compares well to existing literature (Westermann and Evins, 2019).

Surrogate model fitting

The final surrogate model is fitted using 8,000 simulation samples (plus 2,000 more for testing). The optimal 2-layered network architecture has 256 neurons per layer, where the weights are l2-regularized with $\alpha = 10^{-1}$. It achieves a mean $\overline{R}^2 = 0.986$ on the test data averaged over all outputs when predicting annual aggregated performance metrics.

The surrogate model accuracy for each output is shown in Figure 5, where we compare simulation outcomes with surrogate model predictions. The red lines indicate the 0% and ±10% error borders.

The overall accuracy varies between $R^2 = 0.962$ when predicting the heating demand covered by natural gas, and achieves the highest accuracy when predicting the energy demand for water heating $R^2 = 0.999$. Alongside the explained variance, R^2, we also compute the mean absolute error for each of the models providing a slightly better physical insight.

Whereas the thermal demand can be accurately estimated, the supply causes a certain degree of inaccuracy of the model. All thermal supply outputs do not surpass an accuracy of $R^2 > 0.98$. As a core objective of net-zero buildings is the reduction of energy demand, we specifically highlight the accuracy of the surrogate for low-demand buildings. Therefore, we quantify the absolute percentage error (APE) and mean absolute percentage error (MAPE) for the 50% of the samples with the lowest demand for each output. This is visualized in the top left corner of each subplot. For completeness this metric is also computed for PV generation. In future, further refinement of the surrogate model is required to optimize the prediction of thermal sources.

Instant feedback for visualization

The major advantage of the surrogate model for the early design stage, is the instantaneous generation of simulation estimates. Here the model evaluates 100,000 inputs in 3.2s.

This allows for fast and interactive visualization. A popular example is the use of a parallel coordinates plot (e.g. see Figure 6). It features sliders to specify a favoured building design where the column on the right of the plot, here the heating demand, lets the user see the impact of design choices on some performance metric.

Other visualization techniques will be explored in Net-Zero Navigator platform, but are not discussed in this paper.

Outlook and Conclusions

In this paper we introduce the NetZero Navigator platform for early-stage design of low-energy buildings. We provide details on the software architecture, the application realm, and the underlying machine learning models.

In a case study on one of the buildings hosted on the core interface of the platform, we show how the underlying machine learning models are derived, report their accuracy to emulate simulation models and exemplify the design space visualization of that building using a parallel coordinates plot. For this case we show that surrogate models fitted using relatively few samples (8,000 for a 32 parameter problem) can achieve respectable accuracy ($R^2 \geq 96\%$), which is suitable for early-stage decision-making.

The platform will continue to be refined in alignment with user groups' needs. This includes better visualization of the building design spaces, higher accuracy of surrogate models, and refined input sets for the surrogate models. Many different archetype buildings and climates will be available in the core interface, and many parameter configuration and visualization

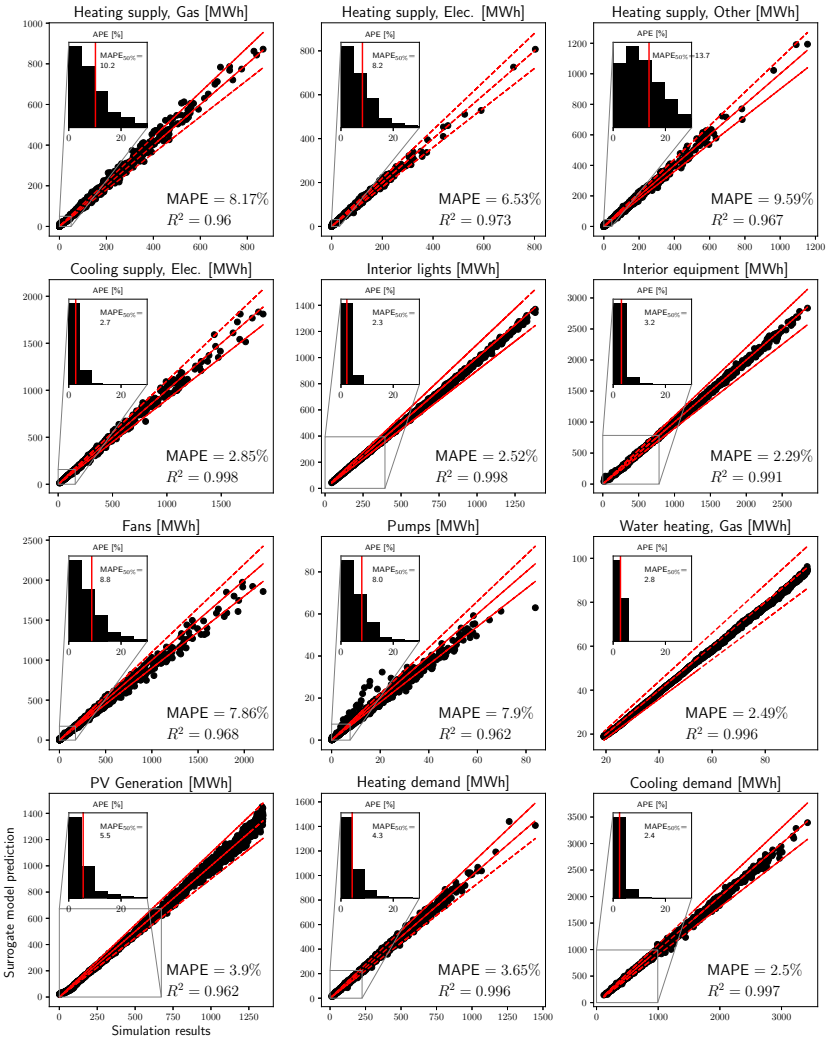


Figure 5: Surrogate model accuracy on test data for one of 16 buildings (Medium office) part of the NetZero Navigator platform.

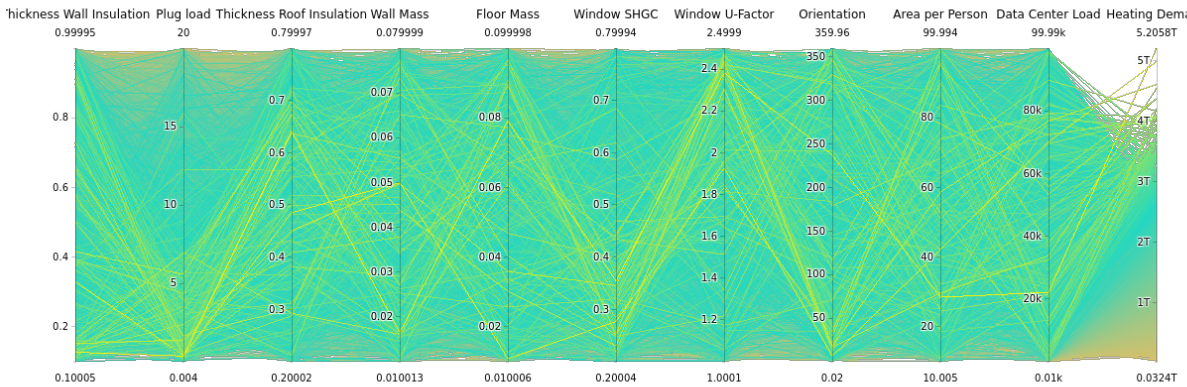


Figure 6: Parallel coordinate plot with 10 of the design parameters and heating demand as output.

options will be available in the advanced interface. Overall, this paper gives an introduction to the exciting developments that are available by combining machine-learning based surrogate models with online visualisation tools.

References

Architecture et climat (2010). *Building performance simulation tools: selection criteria and user survey*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

(2020). *BTAP*. `https://github.com/canmet-energy/btap`.

Evins, R. (2013, June). A review of computational optimisation methods applied to sustainable building design. *Renewable and Sustainable Energy Reviews 22*, 230–245.

Gratia, E. and A. De Herde (2002). A simple design tool for the thermal study of an office building. *Energy and buildings 34*(3), 279–289.

Kalay, Y. E. (1999). Performance-based design. *Automation in construction 8*(4), 395–409.

Kasim, M., D. Watson-Parris, L. Deaconu, S. Oliver, P. Hatfield, D. Froula, G. Gregori, M. Jarvis, S. Khatiwala, J. Korenaga, et al. (2020). Up to two billion times acceleration of scientific simulations with deep neural architecture search. *arXiv preprint arXiv:2001.08055*.

Nielsen, T. R. (2005). Simple tool to evaluate energy demand and indoor environment in the early stages of building design. *Solar Energy 78*(1), 73–83.

Östman, L. E. (2005). *A pragmatist theory of design: The impact of the pragmatist philosophy of John Dewey on architecture and design*. Ph. D. thesis, KTH Royal Institute of Technology.

Petersen, S. (2011). *Simulation-based support for integrated design of new low-energy office buildings*. Technical University of Denmark, Department of Civil Engineering.

Ritter, F., P. Geyer, and A. Borrmann (2015). Simulation-based decision-making in early design stages. In *32nd CIB W78 Conference, Eindhoven, The Netherlands*, pp. 27–29.

Westermann, P. and R. Evins (2019). Surrogate modelling for sustainable building design-a review. *Energy and Buildings*.

Epilogue

This study serves as an example of a typical software product based on surrogate models. Similar ideas have been showcased before in other studies using surrogate models [13][24] or using simplified physics models [8]. We derived a much larger surrogate model that can be applied to more building design problems. Apart from that, the study serves as an example of the surrogate derivation process, and provides us with the following insights.

The performance of the multi-output, feed-forward neural network was high, reaching an error of $< 10\%$ and $R^2 > 0.96$, which aligns with other studies in the field and validates the approach for following research [21]. By plotting the error distribution, this study unveiled that the predictive performance is not robust and can produce large outliers of $> 20\%$. When being provided to users, i.e. building designers and architects, uncertainty in the surrogate model estimates must be taken into account and will be addressed in the following study (Chapter 4).

The careful selection of a large number of inputs and outputs let us derive a surrogate model that can be used for many design questions but significant limitations in the types of design problems that are represented by the inputs remain. At this point the surrogate model is constrained to model performance for only one geometry (only the window size and orientation of the building can be varied). Furthermore, the given performance estimates are derived assuming one climate (Victoria, British Columbia, Canada). If a different geometry or location needs to be modelled, a new set of thousands of simulation runs is required (see surrogate model derivation process in Chapter 2). Augmenting the surrogate model architecture to accommodate multiple climates will be tackled below (Chapter 5).

Chapter 4

Uncertainty-aware surrogate models

In the following paper we address research question 1.1 and develop a surrogate model that produces uncertainty estimates alongside the actual building design performance estimate. The goal is to derive a surrogate model that can independently identify design parameter combinations that produce large prediction errors.

In the following paper, we use Bayesian neural networks and stochastic-variational Gaussian Process models as surrogate models for the same BPS model that we considered in the previous chapter. A core element of this work is to quantify the quality of the uncertainty estimates, represented here by posterior variance estimates given by the two models. A high quality uncertainty estimate guarantees that accurate confidence intervals can be provided to the end user. In addition, we propose uncertainty estimates as a means to *hybridize* simulation and surrogate modelling (see Section 1.3.1), as explained in the paper.

Bayesian learning for uncertainty-aware surrogate models

Paul Westermann[a,*], Ralph Evins[a]

[a]*Energy and Cities Group*
Department of Civil Engineering
University of Victoria, Canada

Abstract

Fast machine learning based surrogate models are trained to emulate slow, high-fidelity engineering simulation models to accelerate engineering design tasks. This introduces uncertainty as the surrogate is only an approximation of the original model.

Bayesian methods can quantify that uncertainty, and deep learning models exist that follow the Bayesian paradigm. These models, namely Bayesian neural networks and Gaussian process models, enable us to give predictions together with an estimate of the model's uncertainty. As a result we can derive uncertainty-aware surrogate models which can automatically suspect unseen design samples to cause large emulation errors. For these samples the high-fidelity model can be queried instead. This outlines how the Bayesian paradigm allows us to hybridize fast, but approximate, and slow, but accu-

Abbreviations: BDL: Bayesian deep learning; BNN: Bayesian neuraln network; SVGP: stochastic-variational Gaussian Process; DoE: design-of-experiment; ReLU: rectified linear unit;

*Corresponding Author

Email addresses: pwestermann@uvic.ca (Paul Westermann), revins@uvic.ca (Ralph Evins)

rate models.

In this paper, we train two types of Bayesian models, dropout neural networks and stochastic variational Gaussian Process models, to emulate a complex high dimensional building energy performance simulation problem. The surrogate model processes 35 building design parameters (inputs) to estimate 12 different performance metrics (outputs). We benchmark both approaches, prove their accuracy to be competitive, and show that errors can be reduced by up to 30% when the 10% of samples with the highest uncertainty are transferred to the high-fidelity model.

Keywords: Surrogate modelling, metamodel, building performance simulation, uncertainty, Bayesian deep learning, Gaussian Process, Bayesian neural network

---

Highlights

- Training of uncertainty-aware engineering surrogate models.
- Comparing deep Bayesian neural networks and Gaussian process models.
- Uncertainty estimates can identify and mitigate errors in surrogate models.

1. Introduction

A wealth of concepts exist to explore the design of new and existing buildings to improve the building sector's large climate footprint [1]. Scaling

them is challenging, as usually each building is designed individually corresponding to the cultural context, climatic conditions, surrounding buildings and design preferences. This impedes the distribution of centrally derived design paradigms to the level of individual building projects.

Architects and engineers play a vital role to bridge the gap between high-level ideas and the individual building projects. Often they use building performance simulation (BPS) to assess the energy and environmental performance of various design options and balance them against design preferences. The computational expense and associated waiting time, however, prohibits an exhaustive design space exploration and optimization. This has led researchers to train machine learning models on simulation input and output data to emulate building simulation models [2].

The computational speed of so-called surrogate models has been the basis for a range of new innovations in the field of building simulation, for example complex, interactive early design tools (e.g. ELSA [3], Building Pathfinder [4], [5]), faster optimization algorithms [6], and detailed design sensitivity and uncertainty analysis [7][8]. A recent survey of building designers confirms that a cohort which received realtime feedback from a surrogate model arrived at higher performing building designs [9].

The growing application of surrogate models draws attention towards the robustness of their performance. Studies have shown satisfactory average accuracy on test data [10] which can be slightly influenced by the type and the complexity of inputs [11] and the selection of outputs [5].

Nonetheless, average errors computed on test data can be deceiving (see Figure 1). Test data usually consists of design samples distributed uniformly

in the design space and may not reflect the portion of the space the building designer is interested in. Large errors on specific building designs may occur (heteroscedasticity of the errors), affecting important design choices and potentially lowering the energy performance of the final building.

Bayesian methods offer a framework to quantify the uncertainty stemming from the inadequacy of an approximate model (epistemic uncertainty) and recent developments in Bayesian deep learning (BDL) managed to integrate them into large machine learning models [12][13]. As a result BDL models can express for which inputs their estimates are uncertain. In our case, a Bayesian surrogate model produces a building performance estimate as a probability distribution, where the entropy or variance of that distribution allow us to quantify the uncertainty. The architect or building designer is therefore provided with a level of confidence in the performance results and thus, can define uncertainty thresholds above which the high-fidelity model, here the BPS tool, is queried to guarantee high confidence results (see Figure 2).

In this study, we explore two different Bayesian models, Bayesian neural networks [14] and stochastic variational Gaussian process models [15], to quantify epistemic uncertainty in surrogate models (see Section 2). We benchmark the overall accuracy against non-Bayesian surrogate models, validate the quality of the uncertainty estimate, and quantify how a *hybridization* of fast but approximate, and slow but accurate models reduces the error of a surrogate model while computational costs increase only slightly (see Section 5 ff.).

5

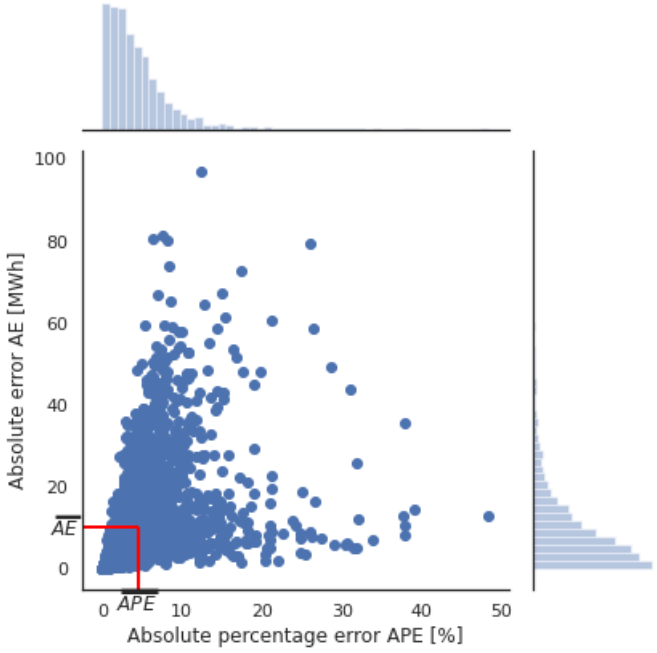


Figure 1: Distribution of errors of a surrogate model. The plot shows the error of a surrogate model which emulates the simulation of the heating demand of an office building (see case study in Section 4). While the average absolute error $\overline{AE}$ and absolute percentage error $\overline{APE}$ are low, large errors can occur. This study aims at identifying the large errors using estimates of the surrogate model's uncertainty.

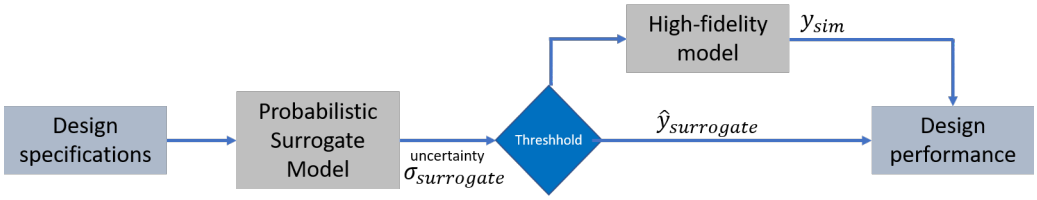


Figure 2: Uncertainty estimates to link high-fidelity model and a surrogate model. The surrogate model provides both a performance estimate $\widehat{y}_{surrogate}$ and an uncertainty estimate $\widehat{\sigma}_{surrogate}$. If the uncertainty is large, a high-fidelity model (e.g. a building energy simulation) is querried to produce accurate estimates y_{sim} of an engineering design (e.g. a building). Compare to [16]

2. Background

2.1. Motivation for surrogate modelling

The fundamental motivation to emulate a physics-based high-fidelity model is computational efficiency; simulation outputs can be estimated many orders of magnitude faster, effectively in real-time. This allows a holistic, intuitive design space exploration and analysis, which would be infeasible with a slow simulation model. Various applications are found in the building domain as well as other domains [17][18]:

- General design space exploration: The relationship between design parameters and performance is interactively explored to improve the user's understanding of the design problem [19][9]. This can happen on the single building level or on the urban level [20]. Often a parallel-coordinates plot is used to visualize the multi-dimensional problem space [5].

- Design optimization: The surrogate model is trained and queried to

accelerate iterative optimization algorithms [21][22][23]. Adaptively training the surrogate model on new simulation samples collected at each optimization iteration can further increase optimization performance [6].

- Sensitivity analysis: The surrogate model is used to run the extensive sampling (thousands of simulation runs) required for global sensitivity analysis methods [7].

- Design uncertainty analysis: Several types of uncertainties exist during the building design process - caused by undetermined design parameters, uncertain contextual parameters (e.g. surrounding buildings, carbon factors, etc.), and vague design constraints [24]. This uncertainty is often quantified using Monte Carlo sampling methods, where samples from uncertain parameter distributions are drawn and simulated to quantify how that parameter uncertainty propagates to building performance uncertainty. With a surrogate model, these uncertainties can rapidly be calculated and updated throughout the design process [8].

- Simulation model calibration: An accurate calibration of a simulation model is required to assess retrofit design choices for an existing building. The calibration, i.e. the process of determining uncertain building parameters, often relies either on iterative optimiziation algorithms [25], or on Bayesian calibration of these uncertain parameters [26]. In both cases simulations are iteratively run to closely match simulation outputs with measured sensor data by adjusting the unknown parameters. One can use surrogate models to reduce the computational

limitations of these approaches. Note that simulation model calibration can be done both for a specific building [27] or for multiple buildings [28]. The latter commonly requires an archetype model whose parameters are repeatedly calibrated using measurements of the considered buildings [29].

2.2. Surrogate model derivation

In surrogate modelling, we fit a machine learning model to a simulation dataset $D = \{x_n, y_n\}_{n=1}^{N} = (X, Y)$, where the inputs X correspond to the simulation parameters and Y to real-valued outputs of the simulation run [18].[1] In the case of building energy surrogate models, the simulation parameters are the building design parameters (e.g. insulation value of the walls) and the outputs are the simulated building performance metrics like total energy consumption or greenhouse gas emissions [2]. Studies also exist with sequential outputs, like hourly energy demand [20].

For deriving the surrogate model the modeller first needs to carefully specify the design problem, i.e. to chose the free design parameters and the performance objectives as well as all other important contextual parameters (surrounding buildings, etc.). Then simulations are run to create the simulation dataset D. The idea is to gain maximum information about the design space (the collection of all possible parameter combinations) per simulation run. Tailored sampling schemes exist, called design-of-experiment methods [30], e.g. Latin-Hypercube-sampling that uniformly distributes samples in

[1] Also categorical outputs can be considered but practical examples are lacking in building simulation literature.

the multidimensional input space. The number of samples must be specified (e.g. 10-1000 samples per parameter dimension [2]) and is adjusted if model accuracy on test samples is too low.

Metrics like the coefficient of determination (R^2), the mean absolute percentage error ($MAPE$), or the root-mean-squared-error ($RMSE$) can be used. Based on [10] and [5], accuracies of $R^2 > 0.99$ are feasible when estimating annually aggregated performance metrics, e.g. heating demand, but they can be significantly lower when more complex performance metrics are estimated.

As mentioned above, surrogate model accuracy is commonly reported as one metric, implying homoscedastic errors. This may not always hold, i.e. the errors may depend on the choice of inputs (heteroscedasticity). By using Bayesian deep learning [12], we aim to train surrogates that are aware of where in the design space, i.e. for which kind of building designs $x \in X$, the model is uncertain and may produce large errors.

2.3. Uncertainty in surrogate models

The true simulation function $y = f(x)$ is not explicitly available. We use the surrogate model to find an estimate $\hat{f}$ to approximate that function. The central root of uncertainty in surrogate modelling is how plausible the determined $\hat{f}$ is (model uncertainty or *epistemic* uncertainty) [12]. This uncertainty is particularly caused by the training set $D = (X, Y)$ which contains only a finite set of points within the space of possible simulation parameter combinations X (the design space) and associated building performance Y. Theoretically, epistemic uncertainty can be reduced to zero given more and more data [12].

We consider the problem of surrogate modelling as free of *aleatoric uncertainty*, which represents the noise inherent in observations.[2] Therefore, we only deal with epistemic uncertainty. We propose that quantifying this can be a powerful aid in surrogate modelling as it acknowledges that we have to train our model with a limited number of simulation samples representing a fraction of the design space, which makes the emulation inaccurate. Bayesian modelling now allows us to reason under that uncertainty, while still benefiting from the advantages of surrogate modelling, i.e. the computational efficiency for large scale design space exploration.

3. Bayesian modelling for surrogates

Bayesian probability theory offers us grounded tools to quantify model uncertainty [32].

To understand the core idea of Bayesian modelling, we consider a parametric model $y = f(x, \Theta)$, where x is the input, f is a space of possible models (see Figure 3) and Θ is the set of model parameters (for example the weights in a neural network). Instead of finding a single Θ, in Bayesian modelling we search for a collection of Θ, that likely has produced the output Y given X. In our case we search for a collection of surrogate models with different weights.

The Bayesian theorem, as shown in Eq. 1, is applied to find a collection which

[2] In the case of sensor data, this can correspond to sensor noise. Here, we consider simulation runs to be deterministic, i.e. the impact of numerical noise to be small. In the case of numerical building simulation, here EnergyPlus [31], this corresponds to the numerical noise of solving the thermodynamic-based differential equations.

likely has produce Y given X. Based on our prior knowledge on the distribution of the model weights $p(\Theta)$ and combined with the likelihood function $p(Y|X,\Theta) = \prod_{n=1}^{N} p(y_n|x_n,\Theta)$, which quantifies the probability that a specific model parameter set generated the observations (X,Y), the posterior of the model parameters can be computed.

$$p(\Theta|Y,X) = \frac{p(Y|X,\Theta)p(\Theta)}{p(Y|X)} \tag{1}$$

where $p(Y|X)$ is called the marginal likelihood. It represents the probability of the observed data given the model f with all possible model parameters. It is a scalar that normalizes the posterior. Given the posterior, we can now infer about future data in form of a predictive distribution:

$$p(y_*|x_*,X,Y) = \int p(y_*|x_*,\Theta)p(\Theta|X,Y)d\Theta \tag{2}$$

The mean and variance or entropy can be derived, where the latter two provide information on the uncertainty in the estimated values. In the building surrogate modelling setting, we predict an expected building performance, e.g. annual heating demand, and an associated uncertainty given building design parameters, e.g. the thickness of the wall (see Figure 3).

3.1. Variational inference

The true posterior of the weights $p(\Theta|Y,X)$ however, is commonly intractable. This is particularly the case in the big data regime when more complex models are required [15]. In the small data regime (below a few thousand samples) posterior inference with a standard Gaussian Process Bayesian model is feasible and was successfully applied for building surrogate models [33][27]. However, with increasing complexity , for example more inputs and outputs (e.g. [11]), standard GPs have major shortcomings:

12

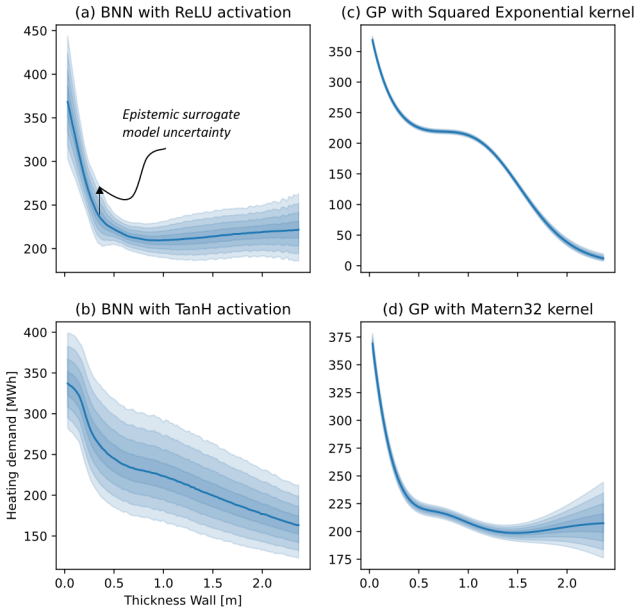


Figure 3: Bayesian neural network heating demand estimate and associated epistemic uncertainty. In particular, out-of-sample the uncertainty of the surrogate model is large. Out-of-sample is that part of the design space, where no (or few) simulations to train the surrogate model on a were collected.

- The model complexity is limited as it only consists of one layer, i.e. the outputs of the GP are not used as inputs to another GP. This prohibits modeling hierarchical structures and abstract information [13].
- Computational cost increase with the cubically ($\mathcal{O}(n^3)$) with the number of samples n. This prohibits increasing the size of the surrogate model training set to improve the model accuracy (for example to train a complex, tailored kernel with many hyperparameters [32]).

Instead, recent advances in variational inference (VI) allow us to approximate the true posterior of Θ in big data problems [34]. We pick an approximate variational distribution over the (latent) model parameters $q_\nu(\Theta)$ with its own variational parameters ν. Now we search for ν that minimizes the divergence to the true posterior which is quantified by the so-called *Kullback-Leibler (KL) divergence*. Thereby the marginalization, i.e. the integration required to calculate the true posterior, is turned into an optimization problem which is often easier to solve. The approximative distribution of q can be used to form predictions about unseen samples.

Scalable variational inference methods have been developed both to do approximative inference with Bayesian neural networks (BNN) [12] and with sparse variational Gaussian process (SVGP) models [15]. The two approaches are introduced in the following section.

3.2. Deep Bayesian Neural Networks

The concept of a Bayesian neural network (BNN) is an extension of standard network architectures (e.g. feed-forward neural network, convolutional

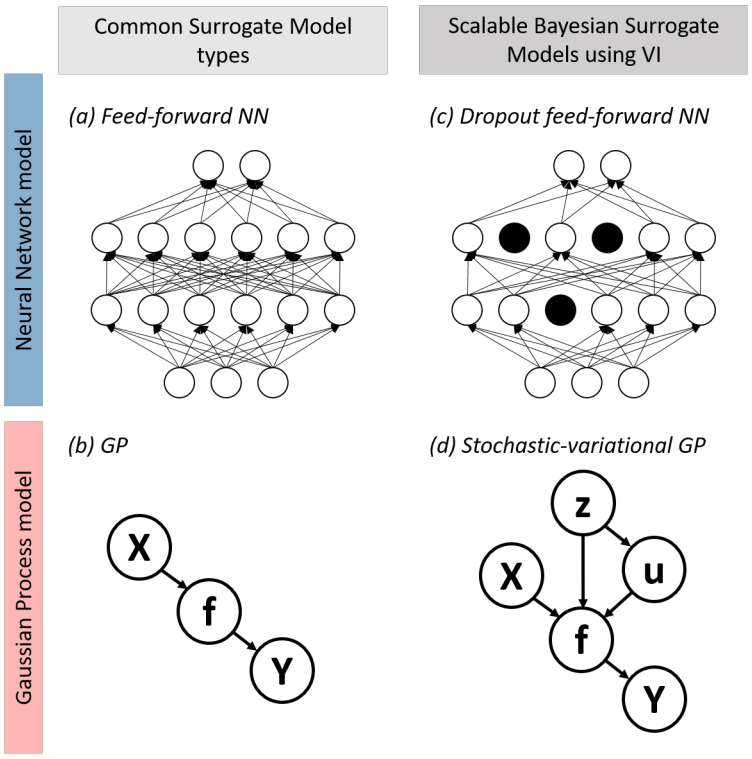


Figure 4: Considered variational-inference approaches to turn existing surrogate modelling architectures into scalable Bayesian models [15][14].

neural network, or recurrent neural network) to follow the Bayesian modelling paradigm [35]. In a BNN we sample the neural network weights from a prior distribution rather than having a single fixed value as in normal neural networks, for example from a Gaussian $\Theta \sim N(0, I)$ [36]. Instead of optimising the network weights directly we average over all possible weights, called marginalisation. Given the stochastic output of the BNN $f^{\Theta}(x)$, we receive a model likelihood $p(y|f^{\Theta}(x))$. Based on the dataset D, Bayesian inference is used to compute the posterior over the weights $p(\Theta|X, Y)$. This posterior captures the set of all plausible model parameters. This distribution allows predictions on unseen data.

As mentioned above the exact posterior is intractable, and different approximations exist [14][37]. In these approximate inference techniques, the posterior $p(\Theta|X, Y)$ is fitted with a simple distribution $q(\Theta)$. Here we consider the Dropout variational inference approach as it has shown great performance when benchmarked against other methods [14][16].

3.2.1. Dropout variational inference

Dropout variational inference is a variational inference approach, i.e. it allows to find a $q_{\nu}(\Theta)$ that minimises the Kullback-Leibler divergence to the true model posterior, that neither requires to change the architecture of common network architectures nor to change the optimisation algorithm for training the network [36]. The inference of the posterior is done by training a model which uses stochastic dropout on every neuron layer [38] (see Figure 4). This stochastic dropout is also used to remove neurons when performing predictions. By repeating the predictions (stochastic forward passes), we create a distribution of outputs, which was shown to minimize

the KL divergence [36].

This KL divergence objective is formally given in the following, where we approximate $p(\Theta|X,Y)$ with $q(\Theta)$ [36][12]:

$$Ł(\Theta, p) = \frac{1}{N}\sum_{i=1}^{N} \log p(y_i|f^{\widehat{\Theta_i}}(x_i)) + \frac{1-p}{2N}||\theta||2 \tag{3}$$

with N data points, dropout probability p, weight samples $\widehat{\Theta_i} \sim q\theta(\Theta)$, and θ the set of the sample distribution's parameters to be optimised (weight matrices in the dropout case). Note that for each data point in the training set dropout is applied, which provides us with N samples of Θ_i.

When performing dropout variational inference the T stochastic forward passes provide us with the epistemic uncertainty given by the variance $Var(y)$:

$$Var(y) \approx \frac{1}{T}\sum_{t=1}^{T} f^{\widehat{\Theta_t}}(x)^T f^{\widehat{\Theta_t}}(x_t) - E(y)^T E(y) \tag{4}$$

with predictions in this epistemic model done by approximating the predictive mean: $E(y) \approx \frac{1}{T}\sum_{t=1}^{T} f^{\widehat{\Theta_t}}(x)$. Note that in this formulation we assumed no noise inherent in the data and therefore, $Var(y)$ is zero when we have no parameter uncertainty.

3.2.2. Model architecture and implementation

We implemented a dropout neural network using the Keras Tensorflow API [39][40] based on the work from Gal and Gahramani [14]. Our network is a feed-forward neural network with 2 hidden layers of 512 neurons which are activated with a leaky rectified linear (ReLU) function. Training was done within 1200 epochs using a batch size of 128 samples. A dropout rate of 5% was set. All mentioned parameters ($n_{layers} \in [1,2,3]$,

$n_{neurons} = [256, 512, 1024]$, dropout rate $\in [5\%, 10\%, 20\%]$) were analysed in a 5-fold cross-validation. The model with the highest accuracy on the test set was picked. Furthermore, we analysed the impact of the dropout rate on the uncertainty quality (see Section 4.3), but no significant change in the performance was observed, which agrees with the observation from [14], that the uncertainties of models with different dropout rate converge with the training progress.

3.3. Gaussian Processes in the Big Data regime

Gaussian Processes models are attractive for non-parametric Bayesian modelling [32]. They use a Gaussian Process prior for a stochastic, latent function f to describe the relationship between X and Y (see Figure 4). The function values $f(x)$ are assumed to be sampled from that Gaussian with zero mean and covariance matrix K, i.e. $f \sim \mathcal{N}(0, K)$. The choice of covariance function impacts various aspects of the GP model and also determines which model parameters Θ to be tuned. These model parameters are optimized when training the GP model.

However, given the above mentioned limitations of standard Gaussian Process models (see Section 3.1), sparse GP approximations have been developed to handle large datasets by lowering the computational complexity to $\mathcal{O}(nm^2)$ [41][42].[3] They rely on the use of inducing variables (or pseudo-inputs), i.e. a reduced set of latent variables with size $m << n$ to represent the actual data

[3]This blog post provides a summary on the history on sparse Gaussian Process models: `https://www.prowler.io/blog/sparse-gps-approximate-the-posterior-not-the-model`.

set D with n samples. The m inducing points are GP realisations $u = f(z)$ at the inducing locations Z which are in the same space as the observed inputs X (but not necessarily part of X). When training the SVGP, the locations of the inducing points Z and the covariance parameters Θ are optimally chosen to minimize the KL divergence. Important is that the locations Z are parameters to shape the variational approximate distribution $q(f)$, rather than being part of the model parameters Θ, i.e. the covariance function with parameters Θ are calculated *for* the inducing locations Z.

In comparison to sparse GPs [41], stochastic variational GPs [15] allow mini-batch training which further reduces computational complexity to $\mathcal{O}(n_{batch}m^2)$. Since [15] and others, deep Gaussian Process models have been developed, too, but are not considered in this study as our case study data set is still of limited size and complexity [13][43]. However, our SVGP model may be regarded as a one-layered deep GP [44].

3.3.1. Model architecture and implementation

Here we train a one-layered stochastic variational Gaussian Process model on batches with 100 samples with a Matern32 kernel covariance function using the GPy implementation based on [15][45]. Again we ran a 5-fold cross validation to pick the covariance function as also a simpler squared-exponential kernel was analysed. Furthermore, although the observed dataset is deterministic, we considered a fixed noise level in the model ($\approx 0.001\%$ of the mean absolute value of the outputs) as it produced much more accurate models. This implies that a deep Gaussian process may be a better choice than o one-layered SVGP.

4. Case Study: Surrogates for the design of NetZero Energy building

4.1. Objective

We use a case study on a popular topic in the building domain, the design of buildings with net-zero energy demand, to train and assess the two Bayesian model types introduce above. It shall serve as an example showcasing the use of Bayesian modelling for building surrogate modelling, but should not be considered as a benchmarking study to find the best approach. For that purpose the reader is referred to other studies, e.g. [16] or [43].

4.2. Case study building

We emulate simulation outcomes of one archetype building contained in the NetZero navigator project [5]. The NetZero navigator projects hosts building simulation surrogate models on a web-platform, which enable to predict building energy consumption of archetype buildings given a large set of building design parameters in real time. So far the platform relied on common deterministic neural network surrogates, whose building performance estimation accuracy was validated on separate building designs not contained in the training data. All the simulation runs for training and testing were collected with the well-known building performance assessment program EnergyPlus [46]. To date, no design-specific uncertainty estimate is produced to tell the user when the surrogate model estimate is not trustworthy.

For this case study, we look at a medium office archetype building, where 35 design parameters are free to choose and the building energy performance

is quantified by 12 separate performance metrics. The office architecture is based on work from the US DOE Canmet-Energy which derived commercial prototype building models. The development of the parameter set, the choice of performance metrics, and software to generate the (parametric) simulation data set, however, was developed individually for that project, where the parameter ranges are directly based on requirements in the Canadian building sector [47]. The mechanical systems are parametrized to capture a wide variety of configurations allowing direct manipulation of the air-side system (incl. heat recovery ventilation, various pump efficiencies) and plant equipment performance of various systems (heat pump, electric resistance heater, biogas furnace, natural gas furnace, air conditioning system). This allows us to explore a large HVAC system design space on a high-level (incl. multi-system setups). All details on the building may be found in [5].

4.2.1. Data set and transformations

We sample the large design space using 10'000 simulation runs, where each individual parameter combination was picked using the space-filling Latin-Hypercube-sampling (LHS) [30]. Similarly, we run additional 3000 simulations and use it as a separate test set. Each individual building simulation run took approximately 2 minutes and 10 seconds using 1 CPU and 4 GB RAM, but varied depending on the parameter choices.

Prior to training, we standardized the uniformly distributed inputs with different ranges to be normally distributed with zero mean. Furthermore, we transformed the 12 output variables to also be close to a normal distribution. Therefore, adaptive Box-Cox transformations was applied [48]. It adaptively finds transformation parameters to transform various kinds of distributions

(here of 12 different outputs) to normal distributions. This, in particular, increased the accuracy of the multi-output neural network compared to other transformations.

4.3. Evaluation criteria

We evaluate the models with regard to multiple objectives: (i) the model accuracy, (ii) uncertainty accuracy , (iii) the effectiveness of uncertainty-estimate-based issue-raising.

4.3.1. R^2 score, MAPE and RMSPE score to quantify overall surrogate accuracy

Our error metrics cover common metrics in the field, i.e. the R^2 [10] and the Mean Absolute Percentage Error (MAPE) [49]. Furthermore, we added the APE_{90} error, i.e. the 90st-percentile of the absolute errors sorted by ascending magnitude, to quantify the robustness of the surrogate model [50].

$$R^2(Y, \hat{Y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{Y})^2} \tag{5}$$

$$\text{MAPE}(Y, \hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{6}$$

where $\hat{Y}$ corresponds to the matrix of predicted values, Y is the matrix of simulated building performance values. When the error term, $Y - \hat{Y}$ approaches zero, R^2 approaches one, and MAPE goes to zero.

4.3.2. Accuracy of Uncertainty estimate

In a well-calibrated Bayesian model the uncertainty estimates capture the true data distribution, for example a 95% posterior confidence interval also contains the true simulation outcome in 95% of the times [51]. Quantifying the level of calibration is a well-known concept in classification [52] but has also been used for regression problems recently [53][51].
Formally, we say that the uncertainty estimates of the surrogate model are well calibrated if

$$\frac{\sum_{n=1}^{N}\{y_t \leq F_t^{-1}(p)\}}{N} \rightarrow p \text{ for all p} \in [0,1] \tag{7}$$

where F_t is the cumulated density function targeting y_t and $F_t^{-1} = inf\{y : p \leq F_t(y_t)\}$ is the quantile function. Here we consider each prediction as a standard, symmetric Gaussian distribution $\mathcal{N}(\mu(X), \sigma(X))$ and the confidence intervals can be computed using the inverse cumulated density function. [4] In practice, we count the fraction of observations falling in the discrete confidence levels derived from the quantile function (see Figure 6, left).

We show the level of calibration in the calibration plot, where perfectly calibrated uncertainty estimates are aligned with the diagonal. To quantitatively compare different calibration curves, one can also compute the absolute difference between the confidence curve and the diagonal, called the calibration error or the area under the curve (AUC) [53]. The problem of the calibration error is that, it can be zero even for homoscedastic uncertainty

[4] This is not necessarily true an recalibration would be required [51].

estimates (constant for any input). Therefore, we also quantify the *sharpness* of the uncertainty estimates by calculating the overall variance in the uncertainty [51] (see Section 5).

4.3.3. Discard-ranking to quantify the effectiveness of uncertainty estimates for surrogate model application

While having accurate uncertainty estimates is the one thing, in building surrogate modelling we are mostly concerned in warning model users, when the model is uncertain and recommend to rather run a simulation instead (see Figure 2). Therefore, we derive a ranking of the samples in the test set based on the magnitude in their uncertainty. This provides two conclusions. First, if it strongly overlaps with the actual surrogate model error the uncertainty estimates are an effective heteroscedastic warning mechanism. Second, we can use the ranking to calculate how much the average error can be reduced when referring a certain percentage of most uncertain samples (here 10% or 20%) to the high-fidelity simulation program than processing it with a surrogate model.

Both aspects are addressed when plotting the mean error computed on discrete percentiles of the test data, where the test data is sorted by the magnitude of the uncertainty. We can compare that curve to the mean error computed using test data sorted by the magnitude of the computed error (oracle ranking). A large distance between the two curves can tell us that the surrogates uncertainty estimates are not helpful to predict when it is inaccurate. Furthermore, by looking at the slope of the curve, we can see by how much the mean error can be reduced if we discard all samples with

uncertainties above a certain threshold.

5. Results

In this section, we show the results of using surrogate models to include epistemic uncertainty estimates. We considered to different Bayesian machine learning models to provide uncertainty estimates, i.e. a deep Bayesian dropout neural network and a stochastic variational Gaussian Process model (SVGP) approach. We scrutinize the performance of both approaches by comparing their predictive accuracy, by comparing the quality of the uncertainty estimates, and by quantifying how effectively the uncertainty estimates allow us to identify possible surrogate prediction errors.

- introduce that we use uncertainty estimates to communicate between high-fidelity and surrogate model.
- introduce structure of this section

5.1. Accuracy and Uncertainty quality

5.1.1. Accuracy

We benchmark the accuracy of the two model types, Dropout Neural Networks and SVGP models. The performance was quantified using three performance metrics as introduced above (see Section 4.3). Each model was trained five times to generate robust results (see Sections 3.2.2 and 3.3.1). The results are shown in Figure 5 and Table 1 in the Appendix.
Both considered models reach a high accuracy of $R^2 > 0.97$, where in particular the neural network explains the largest amount of the variance in

25

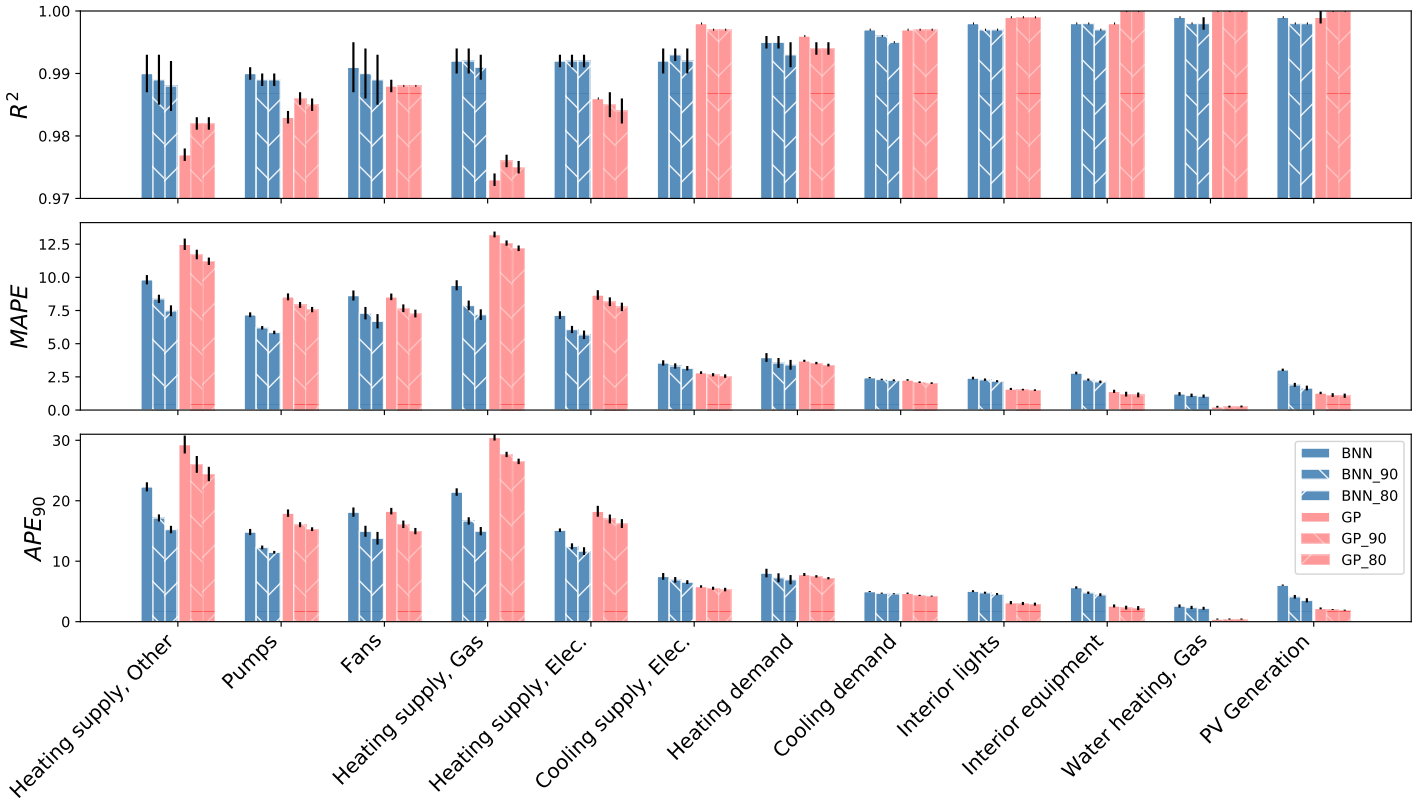


Figure 5: Summary of results on the use of deep, uncertainty-aware surrogate models. The plot shows the accuracy, quantified using three different error metrics, of both Bayesian learning approaches for all twelve outputs considered in the case study.

the testing data ($R^2 > 0.99$). Mean percentage errors of $MAPE < 13.2\%$ for the GP model and $MAPE < 9.82\%$ were found. Where the errors are the largest when the surrogate is used for estimating the supply of different heating systems (using different fuel types) to cover the heating demand, or the energy demand to run the air-side system. Much lower mean percentage errors are found for the other building performance outputs produced by the surrogate like the PV Generation or energy demand for interior lights and equipment.

To check the robustness of surrogate model estimates, we are specifically interested in the highest errors it produces. Therefore, we computed the 90-percentile of the absolute percentage errors observed on the test data, APE_{90}. High errors are found reaching up to 22.3% (30.5%) for the BNN model (GP model).

5.1.2. Uncertainty calibration

When uncertainty estimates are perfectly calibrated, the derived confidence interval, e.g. 90% confidence interval, contains the true outcome in the right number of cases, e.g. 90% of the times. This is illustrated in Figure 6, where we counted for how many times the true simulation outcome was contained in the estimated confidence interval. With a perfectly calibrated Bayesian model, estimated confidence and fraction of the samples within that interval should perfectly align (dashed line). A line below the dashed line would indicate an overly confident model (i.e. confidence bands are too narrow), above the dashed line means that the model is too careful having too large confidence bands.

In this case we find that the BNN model is almost perfectly calibrated, while

the GP model is overly confident. The low level of calibration of the GP model can also be seen on the right, where we display the distribution of all uncertainty estimates on the test data to display the *sharpness* of the uncertainty estimates (see Section 4.3). The average magnitude of uncertainty in the GP model is very small, and its distribution is narrow indicating that the uncertainty estimates tend to be homoscedastic. In case of the BNN they are larger and depict a significant level of variance. This tells us that the uncertainty estimates vary for various inputs, such that we can conclude that the BNN is well calibrated.

5.1.3. Using uncertainty estimates to increase robustness

In this section we study how effective the epistemic uncertainty estimates can be used to predict inaccuracies of the surrogate model.

The concept is simple. We sort the uncertainty estimates on the test data by scale to identify samples where surrogate model estimates are inaccurate. Samples with high uncertainty will be simulated using the high fidelity simulation program instead of the surrogate model (see Figure 2). As a consequence the user of the surrogate model, here a building designer, is facing lower inaccuracies. This must be traded-off against increased runtime, as the expensive high-fidelity simulation program is queried. This trade-off is best handled by defining the uncertainty threshold above which the simulation program is querried. Here, we define the threshold as the 90- or 80-percentile of all uncertainties observed. When using the 90-percentile, we approximately only transfer 10% of all samples to the simulation program, while this is only approximate as it depends on the building design choice of

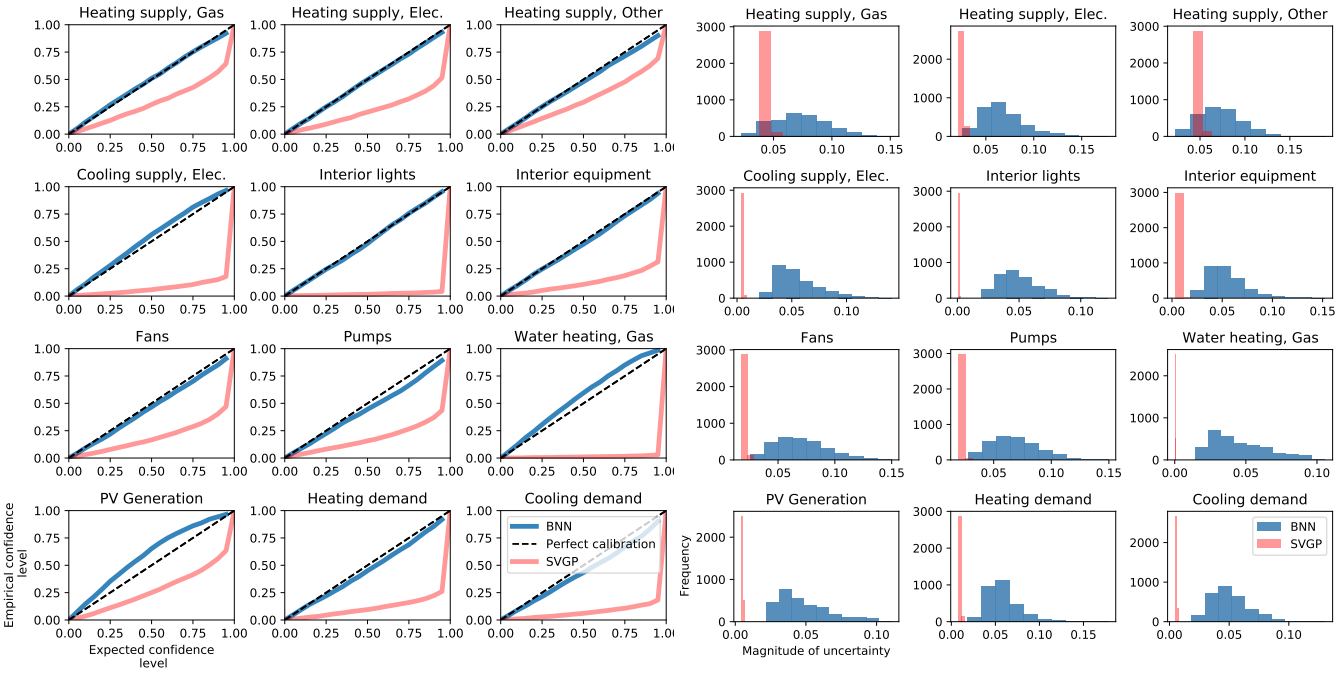


Figure 6: Visualization of the quality of uncertainty estimates of the BNN and the SVGP. The quality is quantified by how well-calibrated and sharp the uncertainty estimates are. In both regards, the BNN outperforms the SVGP in this study.

the architect.

In Figure 7, the decrease in the error is illustrated for the heat supply of different fuel sources. As seen in section 5.1.1, the surrogate model produces the largest errors when estimating the different sources of heat supply, and thus, we focus on increasing the surrogate robustness particularly for them. Discarding the 10% samples with the highest uncertainty, we can decrease the APE_{90} error in estimate the annual heating supply with a gas furnace from 24.9% to 18.9%.[5] This is equivalent to a reduction of $\approx 25\%$.

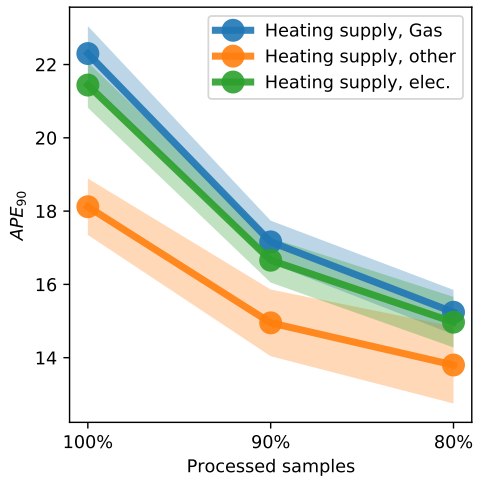
The estimation error on the other surrogate model outputs were reduced by 4% to 18% ($MAPE_{90}$), and the APE_{90} by 5% to 25% (see Figure 7). In particular, the significant reduction of the APE_{90} error, provides the user with higher robustness. Compare BNN and SVGP.

6. Discussion

Surrogate models have shown to help architects and building designers to rapidly assess the energy performance of their designs [9]. However, by being only approximative, concerns about the robustness of the surrogate model accuracy arise. A Bayesian approach for surrogate modelling, allows to not only provide a performance estimate but also inform about the confidence of the approximating surrogate model and potentially, to identify parts of the design space where the surrogate model may provide inaccurate results.

This first analysis of the use of Bayesian surrogate models revealed essential properties on the robustness of surrogate models, and how Bayesian mo-

[5] The 18.9% error was computed on the 90% remaining samples in the test set.



	$\Delta MAPE_{90}$	$\Delta MAPE_{80}$	ΔAPE_{90}	ΔAPE_{80}
Heating supply, Gas [MWh]	-16.17	-23.51	-22.29	-30.18
Heating supply, Elec. [MWh]	-15.10	-20.70	-17.58	-22.87
Heating supply, Other [MWh]	-14.66	-23.83	-23.05	-31.66
Cooling supply, Elec. [MWh]	-6.48	-11.27	-7.61	-12.68
Interior lights [MWh]	-4.98	-9.54	-5.15	-9.70
Interior equipment [MWh]	-17.92	-23.66	-15.40	-21.24
Fans [MWh]	-15.41	-22.48	-17.49	-23.84
Pumps [MWh]	-13.65	-18.52	-17.19	-22.59
Water heating, Gas [MWh]	-9.02	-13.93	-9.27	-14.67
PV Generation [MWh]	-37.29	-45.21	-31.79	-41.56
Heating demand [MWh]	-10.35	-13.89	-9.70	-13.68
Cooling demand [MWh]	-5.33	-7.79	-5.42	-7.43

Figure 7: Measured error reduction by sending uncertain samples to high-fidelity model. The data shows the error if either 100%, 90% or 80% of the building design samples are processed by the surrogate model. If 10% or 20% are processed by the high-fidelity simulation model, errors produced by the surrogate can be avoided and the overall error decreases (here quantified by the 90-percentile absolute percentage error).

delling can be an aid for effective reasoning on the energy performance of buildings under epistemic uncertainty of surrogates. The goal was to augment surrogates such that we can maintain the benefits of surrogate models while minimizing the risk associated to the uncertainty of surrogate models.

6.1. Lacking robustness of surrogate models

Surrogate model accuracy is often reported with error metrics like the R^2 or $MAPE$ score. They are important but can be deceiving. A high coefficient of explained variance (R^2) or low mean absolute errors $MAPE$, may disguise that the surrogate may actually produce quite large errors in certain fractions of the design space. For example, we found that the 90-percentile absolute percentage error can be as high as 22.3% although an $R^2 = 0.99$ suggests very high performance (see Table 1). This motivates, that indeed measures to identify surrogate inaccuracies could greatly lessen the risk associated to surrogate modelling.

6.2. Bayesian learning to express surrogate confidence

Results on the level of calibration of the dropout neural network validated that it can be used to effectively express confidence on its predictions, e.g. one could express that in Figure 3 the heating demand when the wall is $1m$ thick is between $220MWh/year$ and $230MWh/year$ with a 90% confidence. On the other hand, while being almost as accurate as the neural network model, we found that the stochastic variational Guassian Process model produces miscalibrated uncertainty estimates. Please note, that this finding cannot be generalized. First, methods exist to calibrate estimates [51], and deep Gaussian process models were found to produce larger variance in the

uncertainty estimates [43]. Nonetheless, is shows that the quality of Bayesian uncertainty estimates must be validated.

6.3. Accuracy of the Bayesian model compared to deterministic surrogate model

We can compare the results of this study to a common neural network trained on the same dataset, whose accuracy metrics were reported in [5]. The R^2 are higher throughout the study but, $MAPE$ errors increase slightly for most of the outputs (e.g. Heating supply, Gas; Heating supply, Elec.; Heating supply, Other.; Fans), after filtering the uncertain samples however, the Bayesian model is more accurate.

6.4. Bayesian learning to identify erroneous surrogate estimates

We leveraged the uncertainty estimates to express warnings when the surrogate model is highly uncertain. By defining a threshold, here the 90-percentile or 80-percentile of the uncertainty estimates on the test data, we could reduce the APE_{90} error by up to 40%.

This is a significant first step towards the hybridization of fast, low-fidelity and slow, high-fidelity models. Still, practical issues have to be solved. For example, the question arises how to implement the high-fidelity model runs. They could be carried out in the background while the surrogate model user would be working with the vague estimates as a start. In our case the results would be updated after 2 minutes and 10 seconds, which corresponds to the approximate runtime of one simulation.

Another issue is that the computational cost of evaluating a Bayesian model increases compared to a deterministic surrogate model. This is particularly

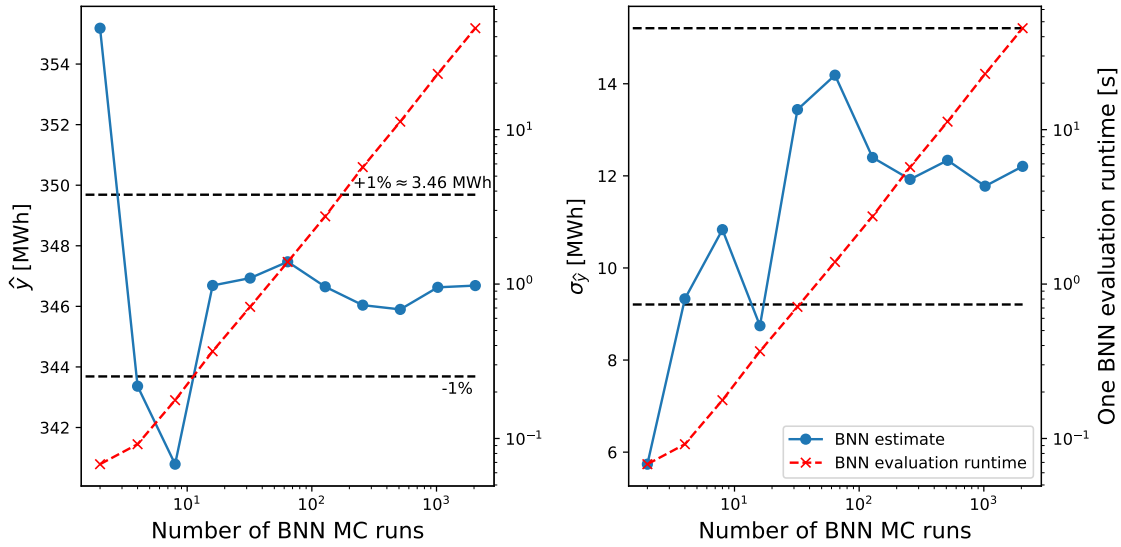


Figure 8: Convergence of BNN estimates with increasing number of Monte Carlo dropout samples. The plot shows BNN heating demand estimates and uncertainty estimates with increasing number of MC samples (see case study in Section 4). Both approximately converge after conducting 30 random dropout runs, which takes around 0.8 seconds (without parallelization).

the case for BNNs, whose uncertainty estimates are generated with Monte Carlo dropout. The BNN estimates converge with increasing numbers of BNN evaluations, which is shown in Figure 8. The plot implies that uncertainty estimates for a single sample take at approximately 0.8 seconds to guarantee convergence. This may be too slow for interactive engineering design tasks but can be easily fixed parallelizing the MC dropout sampling. These and other questions have to be addressed when integrating Bayesian surrogates into software products for building designers.

7. Conclusion and Outlook

In this study we contributed with the use of Bayesian (deep) learning models to mitigate risks associated to the use of surrogate models. By quantifying the model (epistemic) uncertainty, the Bayesian paradigm acknowledges that surrogate models will always remain approximative no matter how large the training set is, and offer a tool to effectively reason under that incurred uncertainty.

In a case study, we could show that dropout neural networks provided well-calibrated uncertainty estimates, which could be used to identify building designs for which the surrogate produced large errors. The latter enables to refer those designs back to the high-fidelity building simulation tool to assure accurate estimates for the architect or building designer. When that filtering process is applied, lower errors compared to a deterministic surrogate model could be reached.

Although all findings are bound to a case study on a building simulation surrogate, results motivate to apply Bayesian learning to other field where surrogate models are common.

In future, we foresee that Bayesian models will allow us to further *hybridize* data-driven surrogate models and high-fidelity simulation models [17]. For that purpose enriching the Bayesian surrogate models with physical know-how could be a key element. Furthermore, Bayesian learning forms a foundation for adaptively sampling simulation runs, for which the surrogate model is particularly uncertain. This progress, called active learning, will be explored in an upcoming study.

Code and Data availability

The entire source code of this work, the EnergyPlus description file (*.idf*) of the building template, and instructions on how to download the data used in this study is available in a GitLab repository.[6]

Acknowledgements

This research was supported by grant funding from CANARIE via the BESOS project (CANARIE RS-327).

References

[1] C. D. John Dulac, Thibaut Abergel, Tracking buildings, Tech. rep., Internation Energy Agency (2019).
URL https://www.iea.org/reports/tracking-buildings

[2] P. Westermann, R. Evins, Surrogate modelling for sustainable building design – a review, Energy and Buildings 198 (2019) 170–186. doi:10.1016/j.enbuild.2019.05.057.

[3] T. Jusselme, Data-driven method for low-carbon building design at early stages, Ph.D. thesis, EPF Lausanne (2020).

[4] Open Technologies, The building pathfinder, Online.
URL http://www.buildingpathfinder.com/

[6] https://gitlab.com/energyincities/building_surrogate_modelling

[5] Paul Westermann, David Rulff, Kevin Cant, Gaelle Faure, Ralph Evins, Net-zero navigator: A platform for interactive net-zero building design using surrogate modelling.
URL http://www.enerarxiv.org/page/thesis.html?id=1975

[6] C. Waibel, T. Wortmann, R. Evins, J. Carmeliet, Building energy optimization: An extensive benchmark of global search algorithms, Energy and Buildings 187 (2019) 218–240.

[7] L. Rivalin, P. Stabat, D. Marchio, M. Caciolo, F. Hopquin, A comparison of methods for uncertainty and sensitivity analysis applied to the energy performance of new commercial buildings, Energy and Buildings 166 (2018) 489–504.

[8] J. Hester, J. Gregory, R. Kirchain, Sequential early-design guidance for residential single-family buildings using a probabilistic metamodel of energy consumption, Energy and Buildings 134 (2017) 202–211. doi:10.1016/j.enbuild.2016.10.047.
URL <GotoISI>://WOS:000390624800018

[9] N. C. Brown, Design performance and designer preference in an interactive, data-driven conceptual building design scenario, Design Studies (2020) .

[10] T. Ostergard, R. L. Jensen, S. E. Maagaard, A comparison of six metamodeling techniques applied to building performance simulations, Applied Energy 211 (2018) 89–103. doi:10.1016/j.apenergy.2017.10.

102.
URL <GotoISI>://WOS:000425075600008

[11] P. Westermann, R. Evins, Using a deep temporal convolutional network as a building energy surrogate model that spans multiple climate zones, Applied Energy 264 (2020) 114715.

[12] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, in: Advances in neural information processing systems, 2017, pp. 5574–5584.

[13] A. Damianou, N. Lawrence, Deep gaussian processes, in: Artificial Intelligence and Statistics, 2013, pp. 207–215.

[14] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, 2016, pp. 1050–1059.

[15] J. Hensman, N. Fusi, N. D. Lawrence, Gaussian processes for big data, in: Uncertainty in Artificial Intelligence, Citeseer, 2013, p. 282.

[16] A. Filos, S. Farquhar, A. N. Gomez, T. G. Rudner, Z. Kenton, L. Smith, M. Alizadeh, A. de Kroon, Y. Gal, A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks, arXiv preprint arXiv:1912.10481.

[17] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, et al., Deep learning and process understanding for data-driven earth system science, Nature 566 (7743) (2019) 195–204.

[18] G. G. Wang, S. Shan, Review of metamodeling techniques in support of engineering design optimization, Journal of Mechanical design 129 (4) (2007) 370–380.

[19] F. Ritter, P. Geyer, A. Borrmann, Simulation-based decision-making in early design stages, in: 32nd CIB W78 conference, Eindhoven, The Netherlands, 2015, pp. 27–29.

[20] J. Vazquez-Canteli, A. D. Demir, J. Brown, Z. Nagy, Deep neural networks as surrogate models for urban energy simulations, in: Journal of Physics: Conference Series, Vol. 1343, IOP Publishing, 2019, p. 012002.

[21] A. Prada, A. Gasparella, P. Baggio, On the performance of meta-models in building design optimization, Applied Energy 225 (2018) 814–826.

[22] B. Eisenhower, Z. O'Neill, S. Narayanan, V. A. Fonoberov, I. Mezic, A methodology for meta-model based optimization in building energy models, Energy and Buildings 47 (2012) 292–301. doi:10.1016/j.enbuild.2011.12.001.
URL <GotoISI>://WOS:000301989800034

[23] F. Bre, N. Roman, V. D. Fachinotti, An efficient metamodel-based method to carry out multi-objective building performance optimizations, Energy and Buildings 206 (2020) 109576.

[24] C. J. Hopfe, J. L. Hensen, Uncertainty analysis in building performance simulation for design support, Energy and Buildings 43 (10) (2011) 2798–2805.

[25] D. Coakley, P. Raftery, M. Keane, A review of methods to match building energy simulation models to measured data, Renewable and sustainable energy reviews 37 (2014) 123–141.

[26] M. Manfren, N. Aste, R. Moshksar, Calibration and uncertainty analysis for computer models - a meta-model based approach for integrated building energy simulation, Applied Energy 103 (2013) 627–641. doi:10.1016/j.apenergy.2012.10.031.
URL <GotoISI>://WOS:000314669500059

[27] Y. Heo, R. Choudhary, G. Augenbroe, Calibration of building energy models for retrofit analysis under uncertainty, Energy and Buildings 47 (2012) 550–560.

[28] J. Sokol, C. C. Davila, C. F. Reinhart, Validation of a bayesian-based method for defining residential archetypes in urban building energy models, Energy and Buildings 134 (2017) 11–24.

[29] M. H. Kristensen, R. E. Hedegaard, S. Petersen, Hierarchical calibration of archetypes for urban building energy modeling, Energy and Buildings 175 (2018) 219–234.

[30] S. S. Garud, I. A. Karimi, M. Kraft, Design of computer experiments: A review, Computers & Chemical Engineering 106 (2017) 71–95.

[31] D. B. Crawley, L. K. Lawrie, F. C. Winkelmann, W. F. Buhl, Y. J. Huang, C. O. Pedersen, R. K. Strand, R. J. Liesen, D. E. Fisher, M. J. Witte, et al., Energyplus: creating a new-generation building energy simulation program, Energy and buildings 33 (4) (2001) 319–331.

[32] C. E. Rasmussen, Gaussian processes in machine learning, in: Advanced lectures on machine learning, Springer, 2004, pp. 63–71.

[33] T. Østergård, R. L. Jensen, S. E. Maagaard, Building simulations supporting decision making in early design–a review, Renewable and Sustainable Energy Reviews 61 (2016) 187–201.
URL https://www.sciencedirect.com/science/article/pii/S136403211600280X

[34] D. M. Blei, A. Kucukelbir, J. D. McAuliffe, Variational inference: A review for statisticians, Journal of the American statistical Association 112 (518) (2017) 859–877.

[35] R. M. Neal, Bayesian learning for neural networks, Vol. 118, Springer Science & Business Media, 1995.

[36] Y. Gal, Uncertainty in deep learning, University of Cambridge 1 (3).

[37] T. Pearce, M. Zaki, A. Brintrup, N. Anastassacos, A. Neely, Uncertainty in neural networks: Bayesian ensembling, arXiv preprint arXiv:1810.05546.

[38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, The journal of machine learning research 15 (1) (2014) 1929–1958.

[39] F. Chollet, et al., Keras (2015).

[40] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin,

S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: a system for large-scale machine learning., in: OSDI, Vol. 16, 2016, pp. 265–283.

[41] M. Titsias, Variational learning of inducing variables in sparse gaussian processes, in: Artificial Intelligence and Statistics, 2009, pp. 567–574.

[42] M. Bauer, M. van der Wilk, C. E. Rasmussen, Understanding probabilistic sparse gaussian process approximations, in: Advances in neural information processing systems, 2016, pp. 1533–1541.

[43] H. Salimbeni, M. Deisenroth, Doubly stochastic variational inference for deep gaussian processes, in: Advances in Neural Information Processing Systems, 2017, pp. 4588–4599.

[44] D. H. Svendsen, P. Morales-Álvarez, A. B. Ruescas, R. Molina, G. Camps-Valls, Deep gaussian processes for biogeophysical parameter retrieval and model inversion, ISPRS Journal of Photogrammetry and Remote Sensing 166 (2020) 68–81.

[45] GPy, GPy: A gaussian process framework in python, `http://github.com/SheffieldML/GPy` (since 2012).

[46] D. B. Crawley, C. O. Pedersen, L. K. Lawrie, F. C. Winkelmann, Energyplus: energy simulation program, ASHRAE journal 42 (4) (2000) 49.

[47] National Research Council Canada (NRCan), National Energy Code of Canada for Buildings 2017 (2017).
URL `https://nrc.canada.ca/en/certifications-evaluations-standards/codes-canada/codes-canada-publications/national-energy-code-canada-buildings-2017`

[48] G. E. Box, D. R. Cox, An analysis of transformations, Journal of the Royal Statistical Society: Series B (Methodological) 26 (2) (1964) 211–243.

[49] R. E. Edwards, J. New, L. E. Parker, B. Cui, J. Dong, Constructing large scale surrogate models from big data and artificial intelligence, Applied Energy 202 (2017) 685–699. doi:10.1016/j.apenergy.2017.05.155.
URL <GotoISI>://WOS:000407188500055

[50] A. Rackes, A. P. Melo, R. Lamberts, Naturally comfortable and sustainable: Informed design guidance and performance labeling for passive commercial buildings in hot climates, Applied Energy 174 (2016) 256–274. doi:10.1016/j.apenergy.2016.04.081.
URL <GotoISI>://WOS:000377728700022

[51] V. Kuleshov, N. Fenner, S. Ermon, Accurate uncertainties for deep learning using calibrated regression, in: International Conference on Machine Learning, 2018, pp. 2796–2804.

[52] J. Platt, et al., Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Advances in large margin classifiers 10 (3) (1999) 61–74.

[53] G. Scalia, C. A. Grambow, B. Pernici, Y.-P. Li, W. H. Green, Evaluating scalable uncertainty estimation methods for deep learning based molecular property prediction, Journal of Chemical Information and Modeling.

Appendix

	R^2		$MAPE$		$APE90$	
	BNN_{ReLU}	SVGP_{M32}	BNN_{ReLU}	SVGP_{M32}	BNN_{ReLU}	SVGP_{M32}
Pumps [MWh]	0.990 ± 0.001	0.983 ± 0.001	7.180 ± 0.180	8.530 ± 0.260	14.830 ± 0.510	17.950 ± 0.610
Heating supply, Other [MWh]	0.990 ± 0.003	0.977 ± 0.001	9.820 ± 0.350	12.490 ± 0.430	22.300 ± 0.750	29.300 ± 1.480
Fans [MWh]	0.991 ± 0.004	0.988 ± 0.001	8.630 ± 0.380	8.530 ± 0.250	18.120 ± 0.770	18.280 ± 0.540
Heating supply, Elec. [MWh]	0.992 ± 0.001	0.986 ± 0.000	7.150 ± 0.290	8.670 ± 0.360	15.130 ± 0.290	18.260 ± 0.900
Heating supply, Gas [MWh]	0.992 ± 0.002	0.973 ± 0.001	9.400 ± 0.380	13.230 ± 0.220	21.440 ± 0.620	30.480 ± 0.520
Cooling supply, Elec. [MWh]	0.992 ± 0.002	0.998 ± 0.000	3.550 ± 0.200	2.820 ± 0.100	7.490 ± 0.560	5.820 ± 0.200
Heating demand [MWh]	0.995 ± 0.001	0.996 ± 0.000	3.960 ± 0.330	3.710 ± 0.080	8.040 ± 0.710	7.800 ± 0.250
Cooling demand [MWh]	0.997 ± 0.000	0.997 ± 0.000	2.440 ± 0.050	2.270 ± 0.060	4.980 ± 0.090	4.700 ± 0.110
Interior lights [MWh]	0.998 ± 0.000	0.999 ± 0.000	2.410 ± 0.100	1.590 ± 0.080	5.050 ± 0.160	3.150 ± 0.270
Interior equipment [MWh]	0.998 ± 0.000	0.998 ± 0.000	2.790 ± 0.100	1.410 ± 0.120	5.650 ± 0.200	2.600 ± 0.250
Water heating, Gas [MWh]	0.999 ± 0.000	1.000 ± 0.000	1.220 ± 0.130	0.250 ± 0.070	2.590 ± 0.260	0.430 ± 0.090
PV Generation [MWh]	0.999 ± 0.000	0.999 ± 0.001	3.030 ± 0.090	1.290 ± 0.090	6.040 ± 0.100	2.200 ± 0.150

Table 1: Numeric results on the accuracy of the Bayesian models.

Epilogue

The study shows that Bayesian neural networks (BNN) can be used off-the-shelf to produce accurate uncertainty estimates. When using surrogates to provide instantaneous feedback to architects as proposed in Chapter 3, we can significantly improve the robustness of the tool by:

- Giving accurate confidence intervals to the user.
- Querying the high-fidelity simulation model once the interval becomes too large.

Further work is required to develop a sound software implementation that handles the hybridization of the surrogate model and the BPS model. Considering that a simulation run in the study takes 2.2 minutes, a smooth integration is difficult. Sequential updating of surrogate estimates with simulation outcomes might be an option.

Another aspect that the study did not cover is active learning. As the surrogate becomes aware of flaws, we could efficiently collect training samples in high-uncertainty-regions of the design space. This approach of active learning, or adaptive sampling (see Section 2), was conducted in the following study by using the LOLA-Voronoi algorithm, which can be used without uncertainty estimates.

4.1 Active learning

Adaptive Sampling For Building Simulation Surrogate Model Derivation Using The LOLA-Voronoi Algorithm

Paul Westermann, Ralph Evins[1]
[1]Energy Systems and Sustainable Cities group,
Department of Civil Engineering, University of Victoria, Canada

Abstract

Statistical surrogate models, or meta-models, are used to emulate building simulation models. Their key advantage is the reduction of computational cost. This in particular matters if building design analysis demands to explore a large number of different building designs options as in optimization or uncertainty analysis problems.
To derive a surrogate model, a data set consisting of simulation in- and output data is generated. This set is then used to train the surrogate. This process of collecting simulation data may be time intensive and a building designer has to wait until surrogate model is available.
In this study we construct a global surrogate model using adaptive sampling to speed up the data collection. In comparison to static sampling, it balances both exploration of the design space while exploiting the iteratively growing information of simulation outcomes. The advantage of adaptive sampling is not only that it can cut simulation time, but also that it rapidly provides a preliminary low-accurate surrogate to the building designer which is sequentially improved while he/she is working with the low accuracy model already.

Introduction

With a 40% share in global carbon emission and 36% share in global final energy consumption, the building sector is a key element for policy makers to address climate change and foster energy efficiency (IEA, 2017). Many policies aim to improve the design of new and existing buildings, or their systems.
Architects and engineers are key to put those policies, often encoded in compulsory annual energy consumption targets, into practice. Therefore, they may either use their own experience on sustainable building design, third party design recommendations or run physical building performance simulation (BPS) tools. In theory, BPS should be the best option. It not only enables to asses the finalized building design but rather to explore a large variety of design options. On the other hand, setting up a BPS model and exploring the design space by multiple simulations can be labour intensive and even more computationally costly. This may cause building designers to rather avoid instead of integrating BPS into their design processes (Petersen, 2011).
Surrogate models are a promising option to remove the barrier of computational cost in BPS (Ostergard et al., 2018). They are used to approximate original BPS models with a statistical machine learning model that is trained on BPS in- and output data (simulation samples). A surrogate is computationally significantly cheaper to evaluate (e.g. 10^6 designs in 1 sec. Ostergard et al., 2018) and enables to return design performance estimates almost instantaneously. Nonetheless, the cost of collecting simulation samples remain and some authors consider the surrogate model approach to only "shift simulation time" to prior to the design process. Indeed, reducing the sampling time (hours) in the surrogate derivation process is crucial and outweighs surrogate model training (minutes) and evaluation (seconds).
Using an optimum sampling plan, also called design of experiment, the information gain per simulation run can be maximised. Two different paradigms for selecting simulation samples exist. In *static sampling* all samples are chosen in one shot. In this case the individual design inputs (sample) are picked to fill the space of possible design options homogeneously. In *adaptive sampling* as shown in Figure 1 samples are picked sequentially to adapt the sampling plan depending on simulation outcomes. This enables to balance space *exploration* with *exploitation* of simulation outcomes. For example, exploitative sampling may be used to identify complex, non-linear regions in the simulation outcomes. It has been observed, that adaptive sampling may outperform static sampling schemes by lowering the number of samples required to achieve a certain level of accuracy of a surrogate model to approximate a high-fidelity simulation model (Garud et al., 2017).
In this study, we implemented the LOLA-Voronoi adaptive sampling algorithm. We identified three key advantages:

- maximise information gain per sample

https://doi.org/10.26868/25222708.2019.211232

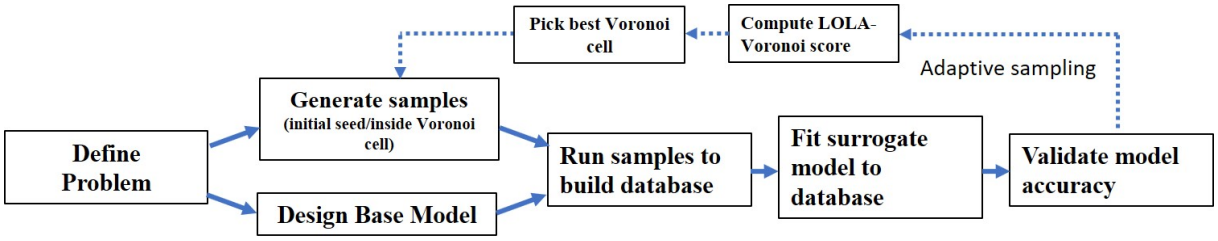


Figure 1: Adaptive sampling with the LOLA-Voronoi algorithm.

- allow *online* surrogate derivation
- no initial choice of no. of samples required

The advantage of the LOLA-Voronoi algorithm over other adaptive sampling schemes, is that can be used in combination with any of the popular surrogate modelling techniques in the building simulation domain and is not bound to Gaussian Process models like other adaptive sampling schemes. We applied the algorithm to fit a surrogate model to a simulation model of a small 5-zone office building. We quantify sampling efficiency to reach a certain level of accuracy and benchmark the results against Latin Hypercube sampling (LHS), the most popular static sampling scheme (Burhenne et al., 2011).

Surrogate Models

The use of Surrogate Models for building design

Surrogate models, or metamodels, have been successfully applied to different types of building performance analysis. Either they are leveraged to study large number of samples as for example in sensitivity analysis (Rivalin et al., 2018), uncertainty analysis (Hester et al., 2017), and optimization (Wortmann, 2018), or used to provide rapid performance assessment to building designers during the early design stage where many different designs are considered (Geyer and Schlueter, 2014). In particular for optimization purposes surrogates have lead to time savings of up to 80% compared to BPS based optimization (Prada et al., 2018).
Surrogates may be either trained to be globally (whole design space) or locally (parts of the design space) accurate. While the former serves as a full replacement of a simulation model, local surrogates are often derived in optimization schemes where only specific parts of the design space are out of interest. The latter cannot be reused in subsequent analyses. In this study we focus on sampling for global surrogates given their general range of applications.

Sampling for building simulation surrogate models

In the following we introduce one static and one adaptive sampling algorithm. Static latin-hypercube sampling is one of the most popular sampling method in surrogate modelling research (Ostergard et al., 2018). We use it as a benchmark for the adaptive LOLA-Voronoi sampler. LOLA-Voronoi has proven to outperform LHS on test functions before (Crombecq et al., 2011) and is widely applicable as it is not tied to a specific surrogate model type or certain number of variables.

Latin-hypercube sampling

Latin-hypercube sampling (LHS) is a stratified sampling scheme. In stratified sampling, the design space is divided into multiple subintervals from which samples are drawn. This reduces the risk of clustering or gaps in the sample set as found in random sampling schemes Garud et al. (2017).
LHS divides each dimension of the design space into K equal bins resulting in K^N hypercubes, where N is the number of samples. K sample points x are listed in a sampling matrix $L = [x^{(1)}, x^{(2)}, ..., x^{(K)}]^T$ where the columns represent the different design parameters and the rows the sample points. In LHS the samples are chosen in that way that in each column there are no two samples that fall in the same bin. Hence the number of bins equals the number of samples drawn. In Figure 2 the initial seed of samples (red dots) where determined using an LHS design with $K = 15$ samples. Hence, both the window-to-wall ratio and the solar-heat-gain coefficient are binned into 15 equal bins. Each bin is represented by one sample.

LOLA-Voronoi sampling

The LOLA-Voronoi adaptive sampling strategy was developed in (Crombecq et al., 2011). Like other adaptive or sequential design strategies it is designed to balance the *exploration* and *exploitation* objective for exploring a design space.
Exploration aims at filling under-sampled parts of the design space. This is very similar to the idea of most common *static sampling* schemes like LHS. *Exploitation* focusses on finding interesting or complex parts of the design space. In surrogate model derivation, exploitative sampling capitalizes simulation outcomes to identify complex (e.g. high gradient) regions in the model outputs. If adaptive sampling is integrated into an optimization scheme, exploitation rather aims to pick samples which are interesting with regard to the optimization objective.

Different metrics are used for balancing the exploitative and explorative value of sample candidates and a list is provided by Garud et al. (2017). Sample candidates are either picked around existing samples with a high sampling score, determined randomly throughout the design space, or they are actively picked using an optimization approach.
In the LOLA-Voronoi algorithm, existing samples are assigned with a hybrid sampling score H.

$$H = V + E \quad (1)$$

where E, the local-linear approximation (LOLA) score, quantifies if the region around a specific sample is non-linear and V if it is under-sampled in comparison to the other samples. Once H for each of the existing samples is calculated, the sample with the highest score serves as a reference to pick additional simulation samples around it. This is done by taking its Voronoi cell, i.e. the region consisting of all points closer to that sample than to any other, and randomly pick a sample inside that Voronoi-cell.
The LOLA score of each sample is computed by fitting a local-linear hyperplane through the simulation outcomes of its neighbours. The length of the normal of that hyperplane serves as non-linearity estimate. Determining the neighbouring samples among all existing ones is crucial to receive accurate non-linearity estimates. A detailed explanation can be found in (Crombecq et al., 2011).
The Voronoi cell size of a certain sample, V, is large if the neighbouring samples are far away where the distance among points is quantified using the Euler distance. Consequently, the larger V, the lower the density of points. Computing the actual cell size is not straight-forward and usually done by Delauny triangulation. In the LOLA-Voronoi algorithm is is estimated to lower computation cost. Therefore, a random set of points is generated within the overall design space. Subsequently, the number of points closest to each individual of the existing points are counted. Samples with the lowest number of assigned points have the lowest density and hence, are under-sampled.
Once the sample with the highest hybrid score H is found. One of the points assigned to that sample during Voronoi cell estimation, is picked as new simulation sample. Here, we limit the number of new samples to one per LOLA-Voronoi iteration but this can be modified.
The LOLA-Voronoi sample selection process is illustrated in Fig. 2 which shows the three samples with the highest exploration (a), exploitation (b) and hybrid score (c). The red dots show initial simulation samples. The blue lines are generated from a Gaussian Process surrogate model trained on this initial set of simulation data. The surrogate was then evaluated at the intersection of the blue lines. To simplify the visualization, this surrogate model was fitted to annual energy demand simulations given two inputs only (window-to-wall ratio and solar heat gain coefficient).
Here, the three samples with the highest corresponding score are encircled with a green line. If only the exploration score is considered, the samples are chosen as reference samples with the closest neighbouring samples being far away. If only the exploitation is considered, those samples with a high gradient in the surrounding region are picked. By summing up both scores a balance between the two objectives can be found (c).

Table 1: Considered parameters, their Morris coefficient and their value range.

Parameter	μ^*	Range(min, max)
Solar heat gain coeff.	$8.4 * 10^8$	(0.1,0.9)[ ]
Equipment gains	$3.5 * 10^8$	(10,15)[$\frac{W}{m^2}$]
Window-to-wall ratio	$4.6 * 10^8$	(0.1,0.9)[ ]
Lighting gains	$3.3 * 10^8$	(10,15)[$\frac{W}{m^2}$]
U-value window	$1.8 * 10^8$	(0.1, 5) [$\frac{W}{m^2K}$]
Infiltration	$1.0 * 10^8$	(10^{-4}, $2 * 10^{-3}$) [$\frac{m^3}{m^2}$]
Conductivity wall	$2.9 * 10^7$	(0.02,0.2) [$\frac{W}{mK}$]
Thickness wall	$2.1 * 10^7$	(0.1,0.5)[m]

Experiment

We applied LOLA-Voronoi sampling to derive a surrogate of a whole building simulation model of a small office building with five thermal zones (Small Office, new construction 90.1-2004; see Deru et al., 2011). The output of the surrogate is total annual energy demand of the building and the inputs are chosen by looking at previous literature where similar surrogates were constructed (Ostergard et al., 2018) and by conducting Morris screening to filter the candidate inputs (features) by their sensitivity (Tian, 2013). The remaining eight most sensitive parameters are shown in Table 1. Morris screening was conducted with five value levels using the SALib library (Herman and Usher, 2017).
Note that, parameter distributions affect the samples collected using static sampling (here LHS). Here all distributions were chosen to be normal with the min and max value serving as 95-percentile.The distributions implicitly consider which design choices are the most likely to happen or are the most common. While static sampling uses parameters distributions, adaptive sampling with the LOLA-Voronoi algorithm is independent to parameter probabilities.

To take the influence of problem size into account, we conducted two experiments with four and eight design parameters. In case of four design parameters only the four most sensitive ones were considered. Here we sequentially trained a neural network model. However, LOLA-Voronoi sampling can be used together with any type of surrogate model. The architecture

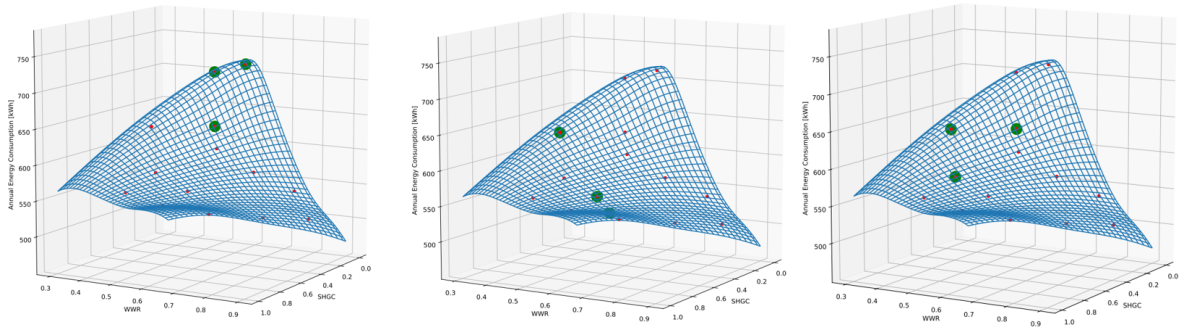


Figure 2: Exploitation, exploration and hybrid score based selection of regions for further samples.

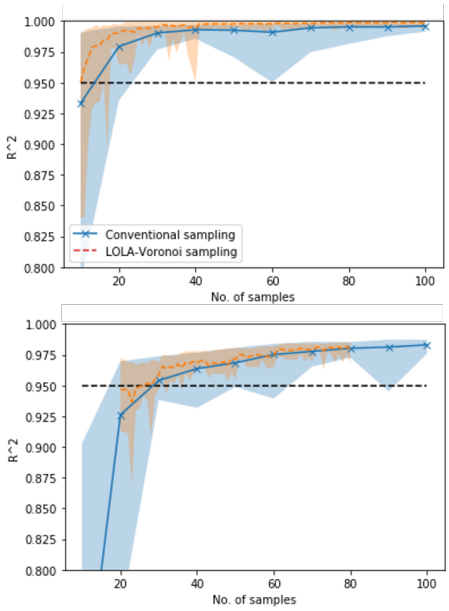


Figure 3: Results. Neural network surrogate model accuracy per number of simulated samples in problem with 4 (top) or 8 design parameters (bottom) were trained.

of the neural network model and its hyperparameters (number of hidden layers, number of samples) were optimized in a grid search and cross-validated.
We quantify the performance of LOLA-Voronoi for building simulation surrogates by plotting the surrogate model accuracy achieved for a given number of samples. The model accuracy is quantified by the coefficient of determination R^2 and is computed on a separate test set of 100 randomly selected samples in either of the two experiments. We benchmarked the results against Latin-Hypercube sampling (LHS), which is a popular static sampling approach for surrogate model derivation (Forrester et al., 2008).

Results and Discussion

The results of the two surrogate model fitting experiments are shown in Figure 3. In the top plot the result of fitting a surrogate with four design parameters (inputs) and in the bottom plot of a surrogate with eight design parameters are shown. In either case we collected enough samples to reach an accuracy of larger than $R^2 \approx 0.95$. Note that all reported accuracy scores were achieved with cross-validated (5-fold) and optimized neural network models. Due to randomness in sampling and model fitting we repeated the process twenty times. The orange dashed line shows the mean results achieved with adaptive sampling and the blue line shows the results for LHS. The band shows the maximum and minimum values found. The lowest x-axis entry corresponds to the initial seed of samples for adaptive sampling.
First, we can see that LOLA-Voronoi is more sampling efficient than LHS sampling in case of four design parameters. Not only the mean accuracy is higher but also the band of observed accuracies is smaller. In case of eight design parameters the performance of both schemes is rather similar.

The results can be discussed with regard to accuracy and implications on the applicability of adaptive sampling during early building design are given in the following.

Accuracy: On first sight the results of the presented experiments indicate that static and adaptive sampling using the LOLA-Voronoi algorithm provide similarly accurate surrogate models with slight benefits of using adaptive sampling if the number of parameters is small. With increasing number of parameters this benefit appears to vanish.
Based on the given results, if one aims for a model with high accuracy using as little simulation samples as possible, adaptive sampling may be a better choice than static sampling. Definitely, further experiments with more number of samples are required to confirm these findings.

Applicability: The given results in Figure 3 outline the advantage of LOLA-Voronoi to adapt sample selections depending on the already existing set of samples to increase surrogate model accuracy. This enables to provide a preliminary surrogate after

an initial set of simulations was conducted and then, constantly updating that surrogate while it may be in use already. Looking at Figure 3 (bottom), a building designer can use a surrogate trained on 20 samples which may have an accuracy of roughly $R^2 \approx 0.90$ (lower end of the band). While he uses the model for first building performance analysis, further samples can be acquired and after some time his surrogate reaches an accuracy of more than $R^2 \approx 0.95$. In static sampling all samples are selected in one-shot. To add further samples, one could only rerun the static sampling scheme to increase the sample density within the design space. No information on the design space complexity would be integrated in this case.

Conclusions and Future Work

This study contributes with an experiment on adaptive sampling for global surrogate model derivation in the building performance domain. A first comparison of static and adaptive sampling is given. The results show that none of the two sampling schemes clearly outperforms the other. However, the tendency is found that adaptive sampling is more sampling efficient if a high surrogate model accuracy is required. Apart from that, we saw the potential that adaptive sampling helps for fast preliminary surrogate derivation whose accuracy is improved while it is already applied by building designers for building performance analysis.
This study is a first step to study the potential use of adaptive sampling for building simulation surrogate models. The scope of the experiments should be extended in future. For example, further adaptive and static sampling algorithms could be considered and the number of design parameters increase to more than eight variables. Furthermore, LOLA-Voronoi currently only accomodates continuous variables and has to be modified such that it can also be used with discrete variables.

Acknowledgements

This work has been supported by grant funding from CANARIE via the BESOS project.

References

Burhenne, S., D. Jacob, and G. P. Henze (2011). Sampling based on sobolsequences for monte carlo techniques applied to building simulations. In *Proc. Int. Conf. Build. Simulat*, pp. 1816–1823.

Crombecq, K., D. Gorissen, D. Deschrijver, and T. Dhaene (2011). A novel hybrid sequential design strategy for global surrogate modeling of computer experiments. *SIAM Journal on Scientific Computing 33*(4), 1948–1974.

Deru, M., K. Field, D. Studer, K. Benne, B. Griffith, P. Torcellini, B. Liu, M. Halverson, D. Winiarski, M. Rosenberg, et al. (2011). Us department of energy commercial reference building models of the national building stock.

Forrester, A., A. Keane, et al. (2008). *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons.

Garud, S. S., I. A. Karimi, and M. Kraft (2017). Design of computer experiments: A review. *Computers & Chemical Engineering 106*, 71–95.

Geyer, P. and A. Schlueter (2014). Automated metamodel generation for design space exploration and decision-making - a novel method supporting performance-oriented building design and retrofitting. *Applied Energy 119*, 537–556.

Herman, J. and W. Usher (2017). Salib: An open-source python library for sensitivity analysis. *Journal of Open Source Software 2*(9), 97.

Hester, J., J. Gregory, and R. Kirchain (2017). Sequential early-design guidance for residential single-family buildings using a probabilistic metamodel of energy consumption. *Energy and Buildings 134*, 202–211.

IEA (2017). Energy technology perspectives. Technical report, International Energy Agency.

Ostergard, T., R. L. Jensen, and S. E. Maagaard (2018). A comparison of six metamodeling techniques applied to building performance simulations. *Applied Energy 211*, 89–103.

Petersen, S. (2011). *Simulation-based support for integrated design of new low-energy office buildings*. DTU Civil Engineering, Technical University of Denmark.

Prada, A., A. Gasparella, and P. Baggio (2018). On the performance of meta-models in building design optimization. *Applied Energy 225*, 814–826.

Rivalin, L., P. Stabat, D. Marchio, M. Caciolo, and F. Hopquin (2018). A comparison of methods for uncertainty and sensitivity analysis applied to the energy performance of new commercial buildings. *Energy and Buildings 166*, 489–504.

Tian, W. (2013). A review of sensitivity analysis methods in building energy analysis. *Renewable and Sustainable Energy Reviews 20*, 411–419.

Wortmann, T. (2018). Genetic evolution vs. function approximation: Benchmarking algorithms for architectural design optimization. *Journal of Computational Design and Engineering*.

Epilogue

[1] This conference paper is the first in the domain of building performance surrogate models to investigate whether the accuracy of surrogate models can be improved with more efficient, active training sampling selection, often called active learning or adaptive sampling. We explored the use of the LOLA-Voronoi algorithm to balance exploration of design space and exploitation of uncertainty estimates.

In the results we could not show that active learning performs much better than commonly used Design-of-Experiment methods, where all samples are selected prior to model fitting. Many other methods to conduct active learning exist. We did not use the Bayesian neural network uncertainty estimates derived in Section 4 because the active learning study was conducted at an earlier point of this PhD studies. The high quality of the BNN uncertainty estimates motivates their use for active learning and will be studied in future.

[1]Note that in Equation (1), E includes a factor to balance exploration and exploitation.

Chapter 5

Generalization of Surrogate models

Machine learning surrogate models generalize within the domain determined by the input parameters (features) and output parameters (targets) as well as contextual constants that define the basis model. The challenge to overcome is to find a model structure that can process a large variety of input parameters, such that the surrogate model can be applied to many different design problems without the need to re-sample simulation runs and retrain the model.

In the most general case, a surrogate model would process the exact same inputs as its simulation counterpart. In building performance simulation these inputs are described by a weather file that gives the outdoor conditions for the simulation period and a building description file. The latter usually contains many parameters that are of less interest to end users and can be ignored. On the other hand, it contains relevant open-ended information which would produce surrogate inputs of variable length, for example the building geometry is defined by variable amounts of rooms, where each room is described separately. [7] developed component-based surrogate models to handle different construction elements individually. They show that their surrogate model components can be combined to approximate the performance of

buildings with variable geometry.

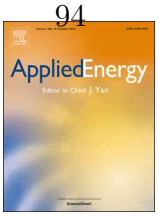
In the following paper, we contribute towards the goal that a surrogate model can process the same inputs as the BPS software. We develop a surrogate model that can estimate building design performance given any weather file. In comparison to the building description file, the weather file is usually of the same size (8760 hourly values of around 20 weather variables). The challenge is to process that large amount of information. We use deep convolutional networks to extract relevant features from the weather input file to predict the annual heating, hourly heating, and annual cooling demand for a building that has thirteen design parameters.

Applied Energy 278 (2020) 115563

Contents lists available at ScienceDirect

Applied Energy

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



Using a deep temporal convolutional network as a building energy surrogate model that spans multiple climate zones

Paul Westermann[a,*], Matthias Welzel[a,b], Ralph Evins[a]

[a] Energy in Cities Group, Department of Civil Engineering, University of Victoria, Canada
[b] Institute of Control Systems, University of Hamburg, Germany

HIGHLIGHTS

- Using deep temporal convolutional networks for building simulation surrogate models.
- Deep network processes annual hourly weather time series data (≈150,000 inputs).
- Accurate emulation of simulation outcomes for all locations in Canada.
- 3% error in estimating annual heating demand for unseen locations and building designs.
- Reasonable accuracy (R^2 = 0.92) in estimating hourly demands given weather data.

ARTICLE INFO

Keywords:
Surrogate model
Metamodel
Building performance simulation
Temporal convolutional neural network
Machine learning
Climate modelling

ABSTRACT

Surrogate models can emulate physics-based building energy simulation with a machine learning model trained on simulation input and output data. The trained model is extremely fast to run, allowing us to estimate simulation outcomes for thousands of different building designs in seconds. Recent studies have shown the diverse benefits for sustainable building design. Surrogates were applied to provide rapid feedback at the early design stage, to accelerate sensitivity analysis, uncertainty analysis and design optimization, or to improve building model calibration.

However, the current process of surrogate modelling offers much room for improvement. In particular, a surrogate model is bound to the specific building design problem it has been trained for. This includes a specific site, requiring time-intensive retraining if the building performance at another location is to be analysed.

In this paper, we develop a single surrogate model that spans arbitrarily many locations. For that purpose, we are among the first to use a deep temporal convolutional neural network to process annual multivariate weather data with hourly resolution (≈150,000 inputs). The network learns features relevant to estimate heating or cooling demand. We combine these location-specific weather features with building design parameters to serve as input to a single surrogate model (feed-forward neural network). In a case study with 569 weather files from locations in Canada, we show that the surrogate model deviates by less than 3% when predicting annual heating demand for new building designs at locations outside of the training data set.

1. Introduction

The building sector is responsible for 28% of global energy-related carbon emissions, which are at an all-time high of 9.6 $GtCO_2$ [1].

The International Energy Agency explains the rise in emissions by the rapidly increasing demand for building energy services which cannot be compensated by the growing availability of carbon-free power.

Growth in emissions is exacerbated by the ever-growing building stock (2.5% in 2017) while energy use intensity reductions are low (0.6% decrease in 2017) [1].

Transforming the building sector is challenging. Each project varies in climate, surroundings, purpose and occupant preferences. Design recommendations for one project may not be suitable for another, leading policymakers to implement purely performance-based building energy codes [1]. Such codes only regulate the whole-building energy performance, no matter which architecture, materials and building systems are chosen. An example is the BC Energy Step Code in British

* Corresponding author.
E-mail addresses: pwestermann@uvic.ca (P. Westermann), matthias.welzel@posteo.de (M. Welzel), revins@uvic.ca (R. Evins).

https://doi.org/10.1016/j.apenergy.2020.115563
Received 30 April 2020; Received in revised form 1 July 2020; Accepted 18 July 2020

Nomenclature

Θ	temperature [°C]
$\overrightarrow{h}_{X_{WTH}}$	vector of extracted weather feature
b_c	kernel bias [–]
c	kernel index [–]
C_l	total number of kernels in layer l [–]
E_l	output of layer l [–]
K_l	kernel size in layer l [–]
l	neural network layer index
$MAPE$	mean absolute percentage error [%]
n	number of samples [–]
$nMBE$	normalized mean bias error [%]
R^2	coefficient of determination [–]
$RMSPE$	root mean squared percentage error [%]
S	stride (step size) of convolutional kernel [–]
T	number of time steps [–]
W_l	tensor of all weights in layer l [–]
X	vector of inputs to surrogate model
X_P	vector of building design parameters
X_{WTH}	matrix of weather input data, values from.epw files
.epw	EnergyPlus weather data file
.idf	EnergyPlus input data file
CDD	cooling-degree days
CNN	convolutional neural network
DoE	design-of-experiment
FFNN	feed-forward neural network
HDD	heating-degree days
HVAC	heating, ventilation and air conditioning system
LSTM	long-short term memory neural network
ReLU	rectified linear unit neuron activation function
ResNet	residual neural network
RNN	recurrent neural network
TCN	temporal convolutional neural network
TMY	typical meteorological year

Columbia, Canada [2]. It defines upper limits for air-leakage, thermal energy demand and mechanical energy demand, but the design choices to meet these limits are unrestricted. As part of the policy, energy advisors are certified to ratify if a proposed design meets the code using building simulation software.

Existing whole building simulation software, like EnergyPlus [3] or IES-VE [4], is reasonably accurate in simulating the performance of almost any building design and system [5]. However, according to practitioners and researchers the tools lack scalability for the customized analysis of many building designs [6]. Setting up the simulation model for a specific project can be labour intensive, and the simulation runtime can be high if a broad set of design variations is analysed. The latter is indispensable for an extensive design space analysis including interactive conceptual design, design parameter sensitivity analysis, uncertainty analysis or design optimization.

The lack of scalability of current building simulation tools, and the growing demand for performance analysis supported by building energy codes, urge innovation of the existing simulation paradigm. Using simplified physics-based, lumped parameter approaches instead [7] is regarded an alternative but may lack accuracy and flexibility to depict the specifics of a design. In this paper instead, we consider to augment existing *slow, white-box* physics-based building simulation tools with *fast, black-box* machine learning-based surrogate models [8,9]. The idea of surrogate models is to emulate building simulation software by training a machine learning model on simulation input and output data. Given the very low computational cost to evaluate the surrogate model, it can provide building performance estimates much faster than physics-based simulation. Although the surrogate is a black-box data-driven model, we retain the link to the underlying physics as the training data is generated using a theory-grounded simulation model.

There is a growing body of work on surrogate modelling for sustainable building design [9]. It has been applied to provide fast feedback in design processes [10], to do performance uncertainty analysis [12,11,13], sensitivity analysis [14,15], and building design optimisation [16–18]. Another promising field is sampling-based building model calibration as a step before retrofit options analysis for existing buildings [19]. The technical aspects of deriving surrogate models are reviewed in [20,21].

A common drawback of the listed applications is that the surrogate models are bound to the specific building design problems covered by the training data. For example, [13] derived a surrogate model for commercial buildings in Nantes with 49 uncertain parameters. If the building site or the design parameters change, for example on a subsequent project, the surrogate model has to be retrained which involves running many simulations. Following [22], the current generation of surrogate models *"shift computational effort for simulation from within a design process to a prior time"* and therewith allow interactive, and sample-intensive detailed design. Nonetheless, the actual reduction in the number of simulation runs may be low. By increasing the number of design problems covered by a single surrogate model, this drawback can be tackled.

In this paper, we derive one single building simulation surrogate model which can generalize over many locations. This requires finding features that effectively describe the weather of each location. We use a deep temporal convolutional network to learn such features directly from raw annual hourly weather files which are fed as input to the model alongside other building design parameters. We test the approach using a large database of typical meteorological year (TMY) weather files for 569 locations in Canada [23]. Canada as one of the largest countries in the world is a suitable test case covering multiple different climates (4 ASHRAE-90.1 climate zones).

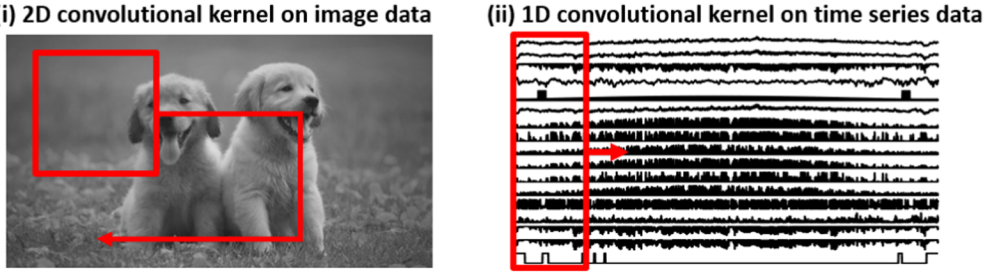


Fig. 1. Comparison of 2D convolutions for image recognition and 1D convolutions for weather time series analysis. A temporal convolutional neural network applies a kernel, i.e. a multi dimensional array of trainable weights, to multivariate time series data where the kernel is shifted in one dimension along time.

1.1. Convolutional neural networks for building simulation surrogates

The choice for deep convolutional networks to interpret weather files is inspired by findings in the domains of image recognition (see Fig. 1). Ref. [24] points out that deep networks composed of multiple layers of twodimensional convolutional kernels are powerful to extract minute, discriminative features from raw image data. These kernels are matrices of weights applied to pixels in close proximity (red box in Fig. 1). They allow to collect information carried in the local coherence of pixels (local connectivity [25]), rather than carried by the values of each pixel individually. Multiple kernels form a network layer, and multiple layers (including other layer types than convolutional layers) form a network. Each individual kernel's weights differ and extract different information from the image (features).

Recently, networks of convolutional kernels were also found helpful for sequential data modelling tasks [26] as for example classifying sensor time series data [27], and *temporal convolutional neural networks* have developed to an active field of research. In this paper, we apply a convolutional neural network to extract features from large, hourly, multivariate weather time series data commonly used to simulate building energy performance. The extracted features serve as input to surrogate models and allow them, to generalize over various climates.

1.1.1. Invariant convolutional kernels and building physics

A key feature of convolutional kernels is their invariance, i.e. the same kernel weights are used at every position of the input [25]. In the case of 1D convolutions applied to time series data (see Fig. 1), this provides a time-invariant kernel shifted along time.

Time-invariant kernels appear promising when modelling weather dependent building behaviour. The building, i.e. a time-invariant physical structure, is excited by the dynamic weather. The kernel size (number of time steps) is crucial to match the dynamics of the building structure. A thermally massive building requires a larger time window than a thermally lightweight building structure.[1] The use of kernels, serving as rolling time windows, sets convolutional neural networks apart from more frequently used surrogate models in the building simulation domain like artificial neural networks or Gaussian Process models. The idea of kernels allows CNNs to learn significant features from very large inputs (here $\approx$ 150,000 values), which is prohibitively large for GP models and is troublesome to train deep dense neural networks on. Also recent studies have shown that convolutional networks outperformed other neural network architectures for time series analysis, like long short term memory networks [26].

1.2. Structure of the paper

We present our approach for location-independent surrogate as follows. In Section 2, we revise surrogate modelling for sustainable building design. This involves *(i)* an introduction into the training process; *(ii)* an overview of the application realm of surrogate models in building performance design; and *(iii)* we review methods to extract features from time series data. In Section 3, we provide details on the applied convolutional neural network implementation and on the case study we use to validate its performance. In Section 4, the results of the case study are given. In Section 5 and 6, we discuss the results and derive conclusions for the field of building simulation.

[1] In the case of our network, we picked a relatively small window size of 8 (8 h). However, by adding convolutional layers to the network a cascade of multiple windows is generated, providing a larger path view.

2. Background

2.1. Building surrogate model derivation

The surrogate derivation process splits into (a) Problem definition, (b) Sampling and (c) Model fitting [8,28].

(a) The task of designing a building with high energy performance is defined by the free design parameters (e.g. window-to-wall ratio, number of floors, building width/length), their range of possible values, fixed design parameters (e.g. local climate, neighbouring buildings, etc.), and an objective, represented by a performance metric. There are usually between 5 and 50 continuous or discrete free parameters [9] which span a highly multidimensional span of possible designs. The machine learning surrogate model is trained to only provide performance estimate for that design space. Hence, a careful parameter selection is crucial.
(b) Samples across the design space (training data) are chosen using design-of-experiment (DoE) methods. Their goal is to maximise information gain per simulation sample to limit the number of runs as much as possible [29].
(c) After sampling, a tabular data set is compiled. The columns include the design parameters and the performance metrics, and each row represents one simulation run. The machine learning surrogate model is trained on that dataset having the variable design parameters as inputs and the performance metrics as outputs. Different machine learning models like feed-forward neural networks, support vector regression models, and Gaussian Process models, are in use and were reviewed in [20,21].

A generic problem definition with a large sampled design space lets us derive surrogate models applicable to a multitude of design projects. In this paper, we advance the research domain by incorporating location, represented by annual hourly weather data, as a dimension of the covered design space. As the weather data is multivariate with hourly resolution the number of additional surrogate inputs is large ($\approx$ 150, 000). This leads to challenges for both sampling such a large space and training a surrogate model with that many inputs. As a result, it allows us to reuse a surrogate for different locations with no retraining.

2.2. Application realm of surrogate models for building design

The low computational cost allows us to use surrogate models to conduct fast energy performance-based building design exploration, for example it can instantaneously generate interactive parallel coordinate plots.[2] The utility of surrogates hinges on their accuracy to emulate the physics-based building performance simulation. The accuracy suffers greatly if the design problem at hand differs from the data set it was trained on. In that case expensive retraining and refitting are necessary.

Based on this, we suggest realms of design problems for which either simulation software, design problem-specific surrogate models, or reusable, generalized surrogate models are suitable. In Fig. 2 we use the stage of the building design process (*y-axis*) and the frequency of building designs (architecture, materials, building systems; *x-axis*) as proxies for the specificity of a design problem and thus as drivers to decide whether to apply simulation software or surrogate models to assess the performance of a building design.

Building performance simulation (including computational fluid dynamics and HVAC simulation [30]) is indispensable for highly complex, innovation-driven building design projects. Always, they can be complemented by *problem-specific* surrogate models to save time in

[2] Tools to derive a parallel coordinate plot using a building surrogate model are available on https://besos.uvic.ca/.

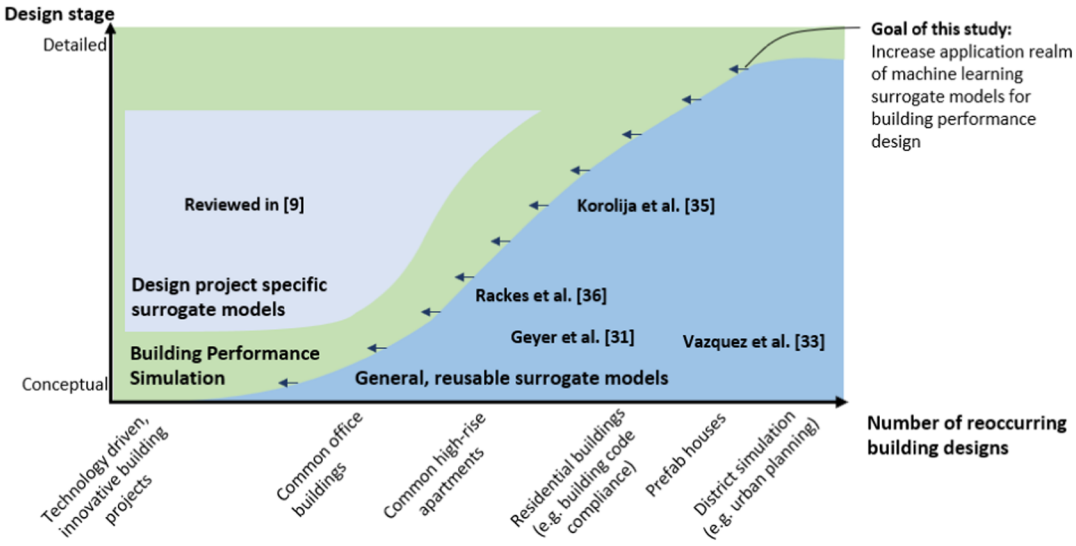


Fig. 2. Application realm for surrogate models in building performance design. At the early design stage and for a high number of reoccurring building designs, surrogate models are becoming a promising option to augment traditional building performance simulation.

design space exploration [9].

On the other hand, building projects often have a lower degree of customization with reoccurring architectural and building system choices. Examples could include fairly standard residential buildings [11], office buildings [31,32], or prefab houses. In that case, surrogate models may be reused for multiple building design processes. Taking that one step further, this allows modelling the energy performance of entire districts or cities, where similar building designs reoccur. In that case a significant reduction of simulation time can be achieved [33].

Apart from the design process stage and frequency of design, other factors like particulars of building usage are crucial to consider when choosing general surrogate models. In this study, we will show that the specific location of a design project does not prevent reusing a surrogate model. This drastically reduces the demand for retraining and even allows to analyse country-wide shifts in building design (see Fig. 12).

Note that apart from building design exploration, building model calibration could become a promising application for general surrogates. Surrogate models enable us to generate energy usage data for a large number of different design parameter combinations. This data is compared to measured data to find the best matching set of parameters of an existing building [34]. A general surrogate model can enable *large scale automated calibration* of individual buildings [34] and districts [19].

2.3. Literature on the generalization of building surrogate models.

By generalizing surrogate models to more building design problems, we can reuse the models on various building projects without running any slow, physics-based simulations. Essential is the number and type of inputs the surrogate model is capable to process. In the most general case, a surrogate model can interpret the same inputs as the building simulation program it emulates. These inputs usually include details of building design objects (geometry, materials, window sizes, etc.) and a weather file (see Fig. 4).

Taking the raw simulation input data is problematic as the format of the building design input file is open-ended, i.e. it changes from building to building. For example, a building with fewer rooms also has a smaller input file. Surrogate models always require the same number of inputs. The authors of [31] tackle that by deriving component-based surrogate models to predict thermal fluxes in building components like walls, ceilings or roofs. They show that their model reached acceptable accuracy on building geometries different from the training cases. Another example of compartmentalized surrogates is given in [35], where a surrogate specifically for the conversion of energy demand to energy supply for various building systems (heat pump, resistance heater, etc.) is given. The more common way to generalize surrogates over a variety of building designs is to calculate features representing the building geometry. Popular choices are the surface averaged insulation value, aspect ratio, the window-to-wall ratio and the number of floors [36,33,9].

While the format of the building design input file may change, the format of the weather input data, usually annual multivariate hourly data, is the same for all locations. This matches the requirement of machine learning models to use one input format, and potentially allows us to use the weather data as input. However, currently engineered features are commonly calculated to discriminate different weather input data, for example heating-degree days (or cooling-degree days) [37], which approximates the number of days in a year for which a certain amount of heating (cooling) is required. Sometimes statistical features are calculated like the mean and standard deviation of outside air temperature, humidity, solar radiation, and wind speed. In ref. [36], 418 different locations in Brazil were considered to estimate comfort in naturally ventilated buildings. They achieved high accuracy on a test set ($R^2 > 0.97$) which, however, included the same locations as the training data. This approach is simple to apply and allowed them to drastically reduce the dimensionality of the weather inputs from 8760 hourly values for up to 25 variables to a few parameter values.

Other authors have trained multiple surrogate models to span different locations. Ref. [38] used four linear regression models with weather data from four locations across the USA. Ref. [39–41] each had similar approaches in which several linear or quadratic polynomial regression models were fitted for locations in different climate zones. In the latter study, 30 models were trained consisting of one surrogate model for each of the five window types for each of the five locations. As soon as the number of locations and categorical parameters increases, this approach becomes infeasible.

2.4. Extracting features from weather data

Feature extraction is the process of finding data representations which explain the variation among the observations in a dataset, and is a core motivation for deep learning [42]. In the case of weather data, any hourly value of the multivariate time series data (8760∗25 values) can contribute to a feature. In this study, we focus on using automated feature learning and compare the determined features against human-based engineered features, which is common in the field of building

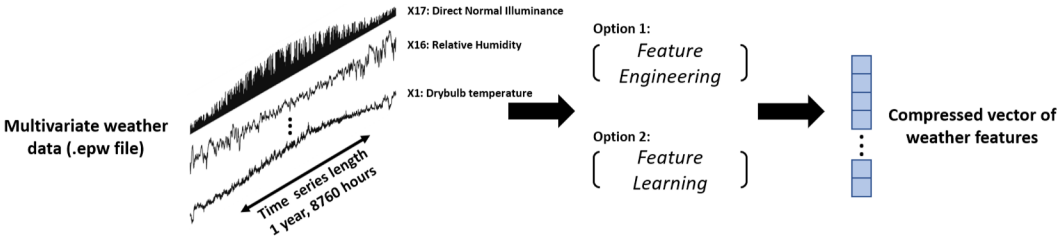


Fig. 3. Feature extraction from weather time series data using feature engineering or feature Learning.

surrogate modelling.

Whereas feature learning can be unsupervised, discriminating the input of the observations without taking any target variable into account, we follow a supervised approach where we search for features that best explain the variation in the target variable, here the heating demand of buildings for various locations (see Fig. 3).

2.4.1. Feature engineering

Engineering features using domain expertise is a common way to capture differences in time series, in particular as input to statistical models estimating building energy demand based on weather data [37]. Weather features are derived in either the time or frequency domain. Compared to other machine learning domains, like image recognition, in surrogate modelling the causalities in the dataset are usually well understood, and manual selection of features can be done based on the physics equations encoded in the simulation software. In our case for example, it is well understood that outside-air-temperature explains the thermal behaviour of a building to a large portion.

In practice, functions are implemented to extract the engineered features. This can be heating and cooling degree days or the mean, standard deviation, maximum and minimum value of each of the variables in the weather time series [36,41].

2.4.2. Feature learning

In feature learning, the machine learning model learns discriminative features from large amounts of raw data [24]. The motivation is to automate the process of feature engineering and apart from that, it potentially increases the feature quality as no data is discarded prior to fitting the model.

As mentioned above, applying feature learning directly to raw building simulation input data is problematic. Building design data is open-ended in format and weather data is very large and complex. Regarding the weather data, a reason may be that some machine learning algorithms, like feed-forward neural networks, are not time-aware and not readily applicable to time series problems.

Outside of building surrogate modelling the growing amounts of time series data, like human activity data, financial recordings, industrial observations, or smart meter data, have catalysed the development of time series analysis methods. In this work, we focus on deep learning approaches. Specifically, we use *end-to-end* time series models, where the model is directly applied to unprocessed raw data [27]. Temporal convolutional neural networks (TCN) recently have attracted attention due to their high performance on sequence modelling problems [26].[3] Similar to deep convolutional neural networks (CNN) on image data, they use kernels sliding over time series data samples. While in image processing kernels are shifted in two dimensions, TCNs use 1D kernels only being shifted along the time dimension (see Fig. 1). More details can be found in Section 3 and in [25].

[3] Note that in comparison to [26], we drop the time causality constraint of the target y_t only being dependent on past inputs $x_0 \ldots x_t$.

3. Methodology

In Fig. 4, the methodological concept of this study is shown. We train a single building energy surrogate model that is applicable over multiple different locations, represented by different annual hourly weather data. The training data consists of EnergyPlus simulation outcomes, i.e. the building simulation software we aim to emulate, which takes a building description file (.idf) and a weather file (.epw) as inputs.

Our location-independent surrogate model can estimate building performance based on both variations in building design parameters and variations in climate (compare Fig. 12). While a change of a design feature only typically affects few entries in the building description file, a change of location demands an entirely different weather file as simulation input, i.e. all entries of the hourly weather data vary. We learn relevant features $\vec{h}(X_{WTH})$ of the high-dimensional weather file X_{WTH} which serve as input to a surrogate model to estimate building performance for various climates (e.g. climates in Canada).

No feature extraction of the design parameters is conducted, as the number of considered parameters in a simulation-based building design exploration is typically low (up to 50 [9]) and can be handled well by a neural network. The varied values of the parameters $\overrightarrow{X_P}$ are therefore used directly. They are concatenated with the extracted weather features $\vec{h}(X_{WTH})$, and together form the inputs X to the surrogate model. While for feature learning we use a deep convolutional neural network architecture, called ResNet, the surrogate model itself is a shallow feed-forward neural net with only one hidden layer.

The codes were developed on the BESOS platform[4] and machine learning models were implemented using Tensorflow via the Keras API [43,44].

In the following, we first introduce our deep learning based feature extraction approach to find $\vec{h}(X_{WTH})$, and the feature engineering approaches we use to benchmark the performance of the deep learning approach. After that we present the case study to empirically show the performance of our approach.

3.1. Feature learning on weather time series data

We use a deep temporal convolutional neural network (TCN) to find $\vec{h}(X_{WTH})$. Two benchmarking studies on machine learning approaches applied to time series modelling showed their outstanding performance [26,27]. Ref. [26] showed that TCNs outperformed recurrent neural networks (RNN), including Long-Short Term Memory (LSTM) networks on a variety of tasks. Ref. [27] found that a ResNet, a residual convolutional neural network [45], outperformed 8 other architectures (incl. fully connected networks, and fully convolutional neural networks) on a variety of univariate and multivariate time series classification tasks.

In both cases, generic convolutional network architectures were applied which can readily be applied to a variety of problems. We use the ResNet architecture [45] to predict a performance target from raw

[4] https://besos.uvic.ca/.

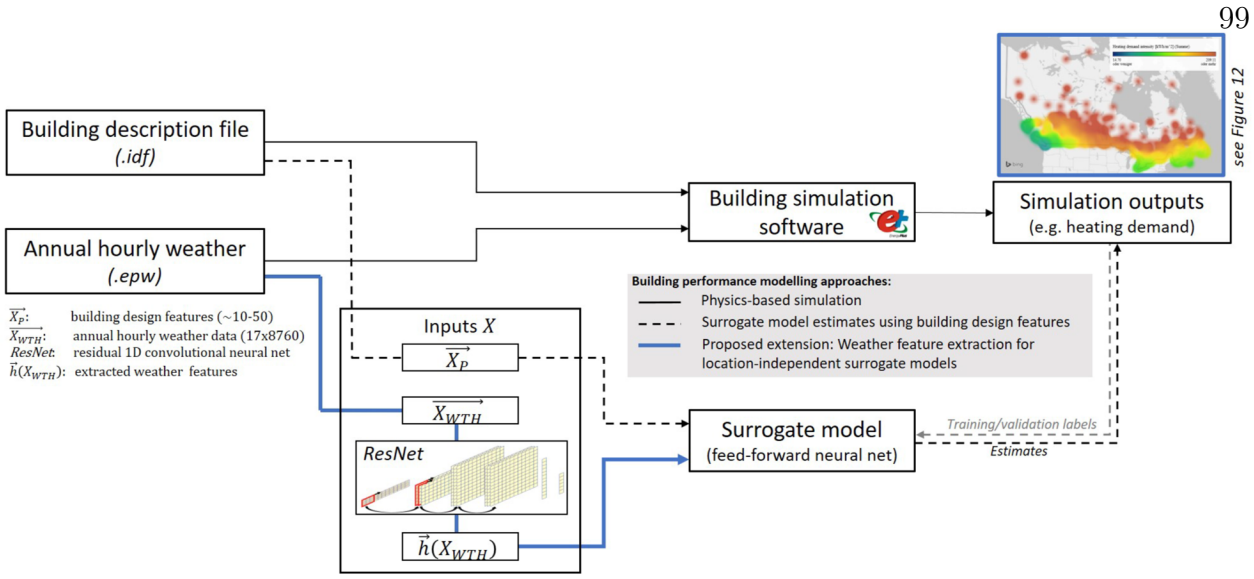


Fig. 4. Our approach for location-independent surrogate models. While conventional surrogate models typically learn from variations in building design parameters only ($\overrightarrow{X_p}$), we use a deep convolutional residual neural network (*ResNet*) allowing to process high-dimensional weather files ($\overrightarrow{X_{WTH}}$) as surrogate model inputs. The evaluation speed of *ResNet* allows us to estimate building performance for thousands of designs at different locations within seconds.

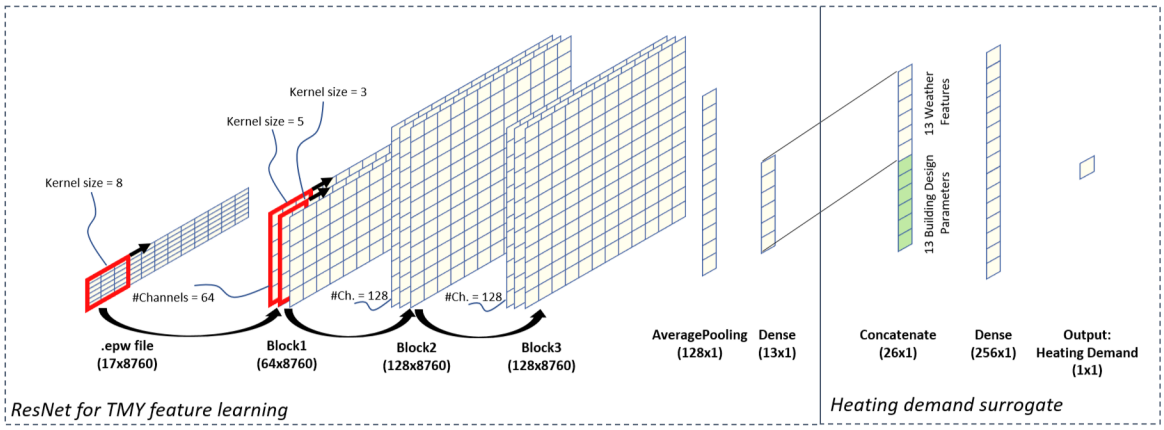


Fig. 5. Architecture of the residual neural network. It encodes high dimensional weather data into low dimensional features which serve as input to a surrogate model [27,45].

annual hourly weather data (see Fig. 5), where the target variable (heating demand, cooling demand, etc.) is easily interchangeable.

3.1.1. Residual Neural Network (ResNet)

The ResNet architecture is a deep convolutional neural network architecture composed of multiple one-dimensional convolutional layers and is shown in Fig. 5. In total it consists of eleven layers. The first nine layers are a sequence of three similar layer-blocks consisting of three one-dimensional convolutions, a batch normalization and a rectified linear unit (ReLU) activation. All those blocks are interconnected by a residual shortcut, connecting the outputs of the last layer of the previous block with the next block (see black arrows) [46]. This prevents a vanishing gradient found when training a deep network architectures using a gradient-based optimizer. Finally, the output of the last convolutional block is averaged along with the time dimension in a global average pooling layer, and compressed to 13 neurons using a fully-connected layer. The number matches the number of building design parameters, fed-in as additional inputs in a subsequent layer.[5]

All convolutions have a stride S of one (step size of the kernel being shifted along the time dimension). Zero padding is applied prior to each convolution on the edges of each the time series, which adds $C_{l-1} \times K_l/2$ zeros to the beginning and end of the sequence data. This guarantees that the size of the time series is preserved across each convolution.[6] The kernel size K shrinks from 8 to 5 and 3 for the first, second and third convolution of each of the three blocks. The number of kernel channels C_l, or the number of rows in each layer output, increases from 17 input channels (i.e. weather variables), to 64 channels in each convolution of the first block, and 128 channels in the two latter blocks.

In Fig. 6 the first convolution following the input layer is shown.

$$E^{(l=1)} = f\left(W^{(l=1)} * E^{(l=0)} + B^{(l=1)}\right) \tag{1}$$

where $E^l \in \mathbb{R}^{C_l \times T}$ is the output of layer l with $T = 8760$ being the total length of the considered time series, W is the tensor of all kernel weights, and B is the vector of all biases.

We apply zero padding, adding $C_{l-1} \times K_l/2$ zero entries to the beginning and end of the input time series matrix X_{WTH}, where the

[5] This number was assessed by looking at the activation of the 13 neurons. Even with 13 neurons only, some of them did not exhibit any activity, see Fig. 9.

[6] Preserving the length of the time series is not necessary to predict annual performance metrics, however, it allows us to use the same architecture for estimating hourly performance metrics $\{y_1, \ldots y_{8760}\}$ for each hour of the weather inputs, see Section 4.3.2.

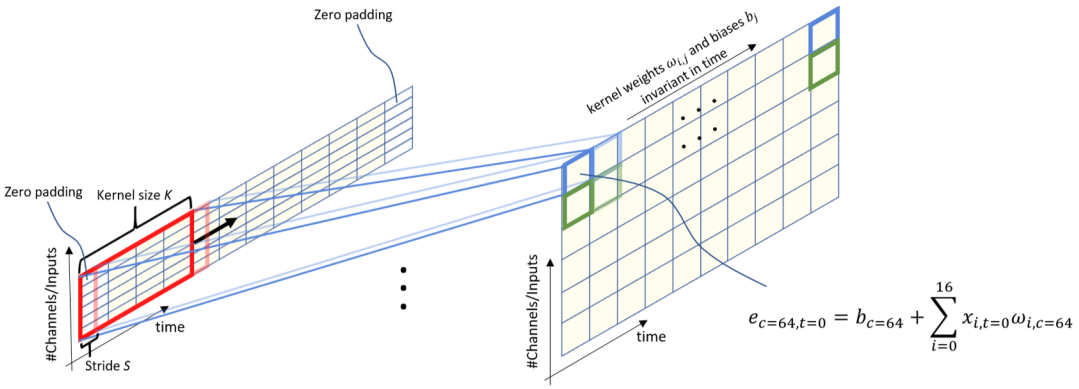


Fig. 6. Details on the parameters of a 1D temporal convolution. A kernel, i.e. a tensor of weights (red window), is applied to sequential data similar to Fig. 1. The kernel weights are invariant when shifted along time. The number of kernels, or channels, corresponds to the number of rows in the outputs of the convolutional layer.

number of input channels is $C_0 = 17$. The kernel size (or size of time window, red box) of the first convolution is $K_1 = 8$ and $C_1 = 64$ different kernels are applied where the blue and green boxes each represent a single kernel shifted along time. Each kernel is shifted a stride of $S = 1$, where the stride S is the same for all layers and kernels. The output of a single neuron is given by

$$e_{c,t} = f\left(b_c + \sum_{i=0}^{C_{l-1}-1} x_{i,t} w_{i,c}\right) \tag{2}$$

with f being the non-linear activation function (here ReLU), b_c being the bias of kernel $c \in C_l$ with being the total number of kernels in layer l, $w_{i,c} \in \mathbb{R}^{K_l \times 1}$ being the vector of weights applied to the input $x_{i,t} \in \mathbb{R}^{1 \times K_l}$, where $t = \{0, ..., T\}$ is the time index.

3.2. Feature engineering approaches for benchmarking

We benchmark the feature learning approach against common features which have been used in building performance surrogate modelling to capture weather impact [36,33]. All approaches are listed in Table 1 and explained in the following.

We assess the quality of the different features $\vec{h}(X_{WTH})$ by feeding them to a surrogate model (Fig. 5, right) with the same network architecture independent of the feature extraction approach. Only the number of features, i.e. inputs to the surrogate, vary. We assume that the more accurate the surrogate model is, the more information is captured by the feature, where the accuracy is quantified by different error metrics.

3.2.1. No weather features

The first feature set serves as a reference case. Although the surrogate is trained on samples from all locations, this set of features does not include any weather information. This allows to point out how much of the variance in the heating demand for various locations is caused by the weather data. We expect the trained surrogate model predicts an "average Canadian heating demand" given a building design.

3.2.2. HDD only features

The second feature set serves as a benchmark for a commonly used engineered feature, the heating degree days (HDD)[37]. HDD are calculated for every location using the following equation:

$$\text{HDD} = \sum_{i=1}^{n} \left(\theta_{ref} - \theta_i\right), \quad \forall\, i\text{:}\ \left(\theta_{base} > \theta_i\right), \tag{3}$$

where θ_{ref} corresponds to a reference temperature at which the building is expected to need no heating, θ_i are the daily means of minimum and maximum outdoor dry-bulb temperature, and n is the number of days in the period, usually a year. HDD values are provided as input along with the 13 building design input.

3.2.3. Engineered features

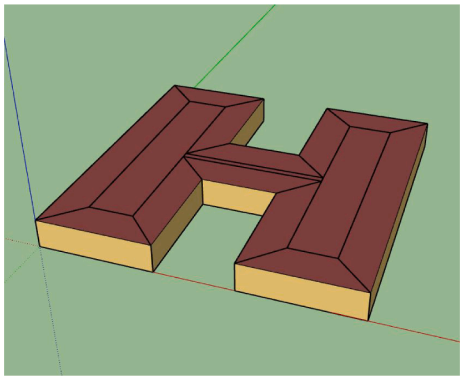
This feature set contains a range of manually picked features. Alongside the heating degree days, it includes longitude, latitude, and elevation as well as the mean and standard deviation over the year of dry-bulb temperature, relative humidity, and global horizontal solar radiation which are calculated using the weather files. Again, the whole set of features is provided as input to the surrogate model along the 13 building design inputs.

3.3. Surrogate model architecture

The surrogate model architecture is the same no matter which feature set is used. For each set of features, it is trained separately and its accuracy recorded. The architecture is a shallow feed-forward neural network with one fully connected layer of 256 neurons, which we regularize using L2-regularization. We use a rectified linear unit (ReLU) activation. This surrogate architecture is based upon previous publications in the field of building energy surrogate modelling with

Table 1
Feature extraction approaches for finding $\vec{h}(X_{WTH})$. All features share the same feed-forward neural network (FFNN) architecture with different number of inputs (weather features combined with building design features), one hidden layer with 256 neurons, and one output.

Name	Weather features $\vec{h}(X_{WTH})$	Network architecture
No weather Data	None	FFNN(13, 256, 1)
HDD only	heating degree days	FFNN(14, 256, 1)
Engineered	heating degree days, longitude, latitude, elevation, mean & stand. dev. of: dry-bulb temperature, relative humidity, solar radiation (hor.)	FFNN(23, 256, 1)
Learned (ResNet)	all annual hourly weather data encoded to 13 learned features	*ResNet* + FFNN(26, 256, 1)



No.	Parameter Name	Unit	Min.	Max.
1	North Axis	°	0	360
2	Wall Insulation Conductivity	$\frac{W}{mK}$	0	0.1
3	Wall Insulation Thickness	m	0.05	0.2
4	Window Solar Transmittance	-	0.1	0.7
5	Window Conductivity	$\frac{W}{mK}$	0.008	0.05
6	Wall Solar Absorptance	-	0.7	0.95
7	Heating Setpoint	°C	18	22
8	Cooling Setpoint	°C	22.01	26
9	Equipment Gains	$\frac{W}{m^2}$	10	12
10	Lights Gains	$\frac{W}{m^2}$	10	12
11	Ventilation Flow Rate	$\frac{m^3}{sPerson}$	0.005	0.01
12	Infiltration Flow Rate	$\frac{m^3}{sm^2}$	0.0001	0.0005
13	Window to Wall Ratio	-	0.01	0.99

Fig. 7. Building template and the varied design parameters. The black contours show the 15 thermal zones varying between $16m^2$ and $180m^2$. The perimiter zones all have a room depth of $5m$.

similar complexity in the design parameters [20] and was refined considering different numbers of hidden layers (1–3) with various numbers of neurons (2^6-2^9). Both parameters had low impact on the final performance, whereas the l2-regularization coefficient had a far larger impact. We picked it in a grid search and 3-fold cross-validation for each of the feature extraction approaches, where the learning rate was fixed.

3.3.1. Error metrics

We base the selection of error metrics on previous work in the field encompassing the coefficient of determination R^2 [20], the normalized Mean Bias Error (nMBE), the Root Mean Squared Error (RMSE) and the Mean Absolute Percentage Error (MAPE) [47]. While R^2, nMBE and MAPE are relative metrics, the RMSE is an absolute metric. Given the large variance in heating demand among the considered design and locations, relative metrics are favoured and we use the Root Mean Squared Percentage Error (RMSPE) instead. All metrics are defined as follows,

$$R^2\left(y, \hat{y}\right) = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{4}$$

$$\text{nMBE}(y, \hat{y}) = \frac{1}{n} \frac{\sum_{i=1}^{n} \left| y_i - \hat{y}_i \right|}{\bar{y}_i}, \tag{5}$$

$$\text{MAPE}\left(y, \hat{y}\right) = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i}, \text{ and} \tag{6}$$

$$\text{RMSPE}\left(y, \hat{y}\right) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{y_i} \right)^2}, \tag{7}$$

where $\hat{y}$ corresponds to the vector of predicted values with length n, y corresponds to the vector of true values, and $\bar{y}$ corresponds to the mean of y values. When the error term, $y_i - \hat{y}_i$, approaches zero, R^2 approaches one, and nMBE, MAPE and RMSPE go to zero.

The four terms provide a diversified insight into the error characteristics. For example, when the points show a small variance, R^2 might indicate a good model performance whereas model bias (systematic over- or underprediction) is large. MAPE and RMSPE differ by their denominator. Due to the quadratic term in RMSPE, few but large sample errors impact the aggregated errors much more than in case of MAPE [48].

3.4. Case study

We apply the methodology in a case study, where the building performance simulation outcomes of design variations of an office building at 569 locations in Canada are emulated.

3.4.1. Building template

The considered building is a small, H-shaped office building with a footprint area of 1300 m^2, designed to host 75 people (Fig. 7). A set of influential parameters and their respective sampling ranges is chosen and complemented after a comparison with commonly employed parameters [9]. We limit the number of variable building design parameters to only 13 parameters and do not include any changes of the geometry (besides window sizes), as we are focussing on the generalization of surrogate models for various locations. However, our method of generalizing a surrogate over different weather data sets can easily be applied to surrogates with more than 13 parameters, including those that generalize geometry which is addressed in[49].

3.4.2. Building performance outputs

We apply and benchmark the feature learning against the feature engineering approaches. Therefore, we calculate the accuracy of the surrogate model, trained separately using all the different sets of features, to estimate annual heating demand. In EnergyPlus, this is represented using an Ideal Air Loads HVAC system object, which quantifies how much energy must be added to each thermal zone of the building in order to meet comfort constraints. Here, we assume zero latent heating loads and 30% latent cooling loads.

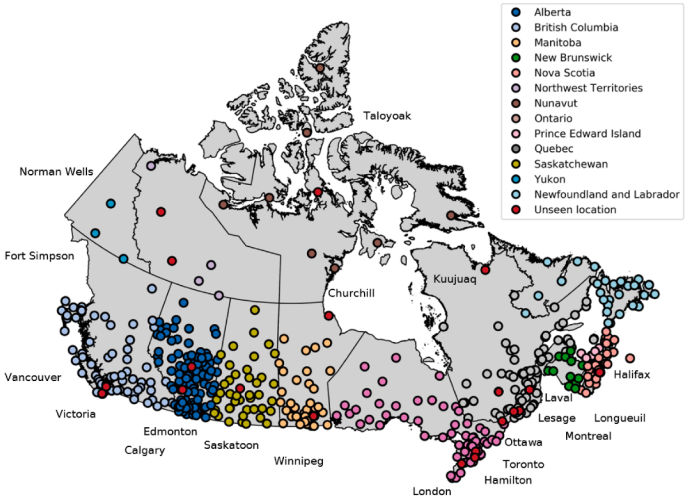
Furthermore, as *ResNet* is designed to be problem agnostic, we also analyse its performance to predict cooling demand (i.e. ideal cooling loads) and *hourly* heating demand.

3.4.3. Weather Data

The weather data used for the simulation consists of 569 *.epw* files (EnergyPlus weather files) which are shown in Fig. 8 [23]. The data is based on several years of observations providing a Typical Meteorological Year (TMY) [50] and consists of 8760 entries, one for every hour in a year, recording 29 weather parameters. The geographic distribution of the files is not uniform but similar to the population density of Canada. The files also include metadata such as province, latitude, longitude, and altitude of the weather station location.

Twelve of the 29 parameters in a weather file had significant amounts of missing values and were discarded. The remaining 17 parameters are shown in Table 8. If they had few missing values (<0.1% of all entries) we filled them using the closest non-missing value (up to six time steps). [7]

[7] Please note, EnergyPlus does not use extraterrestrial horizontal radiation,



No.	Parameter Name
1	Dry-bulb Temperature
2	Dew Point Temperature
3	Relative Humidity
4	Atmospheric Station Pressure
5	Extraterrestrial Horizontal Radiation
6	Horizontal Infrared Radiation Intensity
7	Global Horizontal Radiation
8	Direct Normal Radiation
9	Diffuse Horizontal Radiation
10	Global Horizontal Illuminance
11	Direct Normal Illuminance
12	Diffuse Horizontal Illuminance
13	Wind Direction
14	Wind Speed
15	Total Sky Cover
16	Opaque Sky Cover
17	Snow Depth

Fig. 8. Geographical origins of weather files and considered climate variables used for feature learning. The places used for testing the surrogate model are annotated.

3.4.4. Sampling

We pick simulation samples using Latin Hypercube sampling (LHS) [51]. It is a space-filling approach, where the design space is partitioned into N strata from which one sample each is drawn (N being the number of samples). Here, we sample the building design parameters from uniform distributions with minimum and maximum values shown in Fig. 7. We apply LHS to collect both the training/validation data set ($n = 10,000$) and the test set ($n = 10,000$). This ensures that both sets cover the entire design space. Learning curves were used to find a suitable number of simulation samples.

We picked 20 cities for testing the surrogate model accuracy ($n = 500$ each), referred to as *unseen locations* in Fig. 8, and the rest of the weather data is used for training. Running all $n = 20,000$ simulations took 15 h on 13 2.1 GHz CPUs (approximately 30 s per sample with one CPU).

4. Results

In this section we quantify the accuracy of our approach for location-independent surrogate models using the case study introduced in Section 3.4. The accuracy of the surrogate model having both learned features (using *ResNet*), and manually selected weather features as inputs is compared. As part of that comparison we try to interpret the physical meaning of the learned features using a correlation heat map and feature map visualiziations. Finally, we conducted experiments to show that the proposed location-independent surrogate model architecture is problem agnostic, i.e. that it can be used to emulate ambiguous building performance simulation outcomes. Here we present results of the surrogate model estimating annual cooling demand and sequential hourly heating demand (8760 values). The latter qualifies surrogates for advanced design tasks, as for example heating, ventilation and air conditioning system design [52].

4.1. Benchmarking the feature learning approach

In the following, we benchmark the accuracy of a surrogate model having four different sets of inputs, i.e. one set of learned features and three sets of manually selected features (see Fig. 5). This is performed on the task to emulate annual building heating demand simulations of buildings with various designs at 569 training climates and 20 testing climates.

We compare the performance of *no weather*, *HDD*, *Engineered*, and *Learned* features in Table 2 using the performance metrics introduced above. All results are based on a 3-fold cross-validation and the best model is tested on 500 building designs at 20 unseen locations (climates).

The surrogate model which does not receive any weather features performs by far the worst. We use it to quantify how much of the variation in the target heating demand is caused by variation in weather data.

We find that the other approaches all provide mean absolute percentage errors of less than 10% on the test data. All of them slightly underestimate heating demand in the test data (nMBE<0). The learned features (*ResNet*) allow us to find the most accurate estimates. They almost cut the RMSE error by 50% compared to the second best approach (*Engineered*), which again has almost half the RMSE of a heating-degree-day-only based surrogate model. The MAPE could also be significantly improved (>20%). When comparing MAPE and RMSPE, we notice that the learned features, in particular, lower the RMSPE (3.81%) such that it almost aligns with the MAPE (2.94%). Aligned RMSPE and MAPE indicate that no big outliers in the errors are found (see Section 3.3).

We further look at the geographical variation of the error. Therefore, we break down the error terms for each location in the test set (see Table 3). Each row corresponds to the R^2 score calculated when estimating the heating demand for 500 building designs at each test location. Using the learned features the surrogate model has the highest accuracy for 15/20 cities. However, for most cities the differences in the performance are rather small. We find that *ResNet* in particular helps to improve the accuracy at locations where the other extraction approaches lead to very low accuracy scores, i.e. Victoria, Halifax, Toronto, and Vancouver. This indicates that overall feature learning produces lower error variation throughout Canada, which implies higher generalizability of a single surrogate model with feature learning.

4.2. Physical insight into the learned features

Having shown the competitive performance of learned weather features as inputs to surrogate models, we now try to better understand their physical meaning.

In this study, we have access both the learned features and our

(*footnote continued*)
global horizontal radiation, global horizontal illuminance, direct normal illuminance and diffuse horizontal illuminance as input. Still, they are kept as input to *ResNet*.

Table 2
Surrogate model accuracy with different features as input.

	Training Performance				Testing Performance (unseen designs and locations)			
Feature set	R^2	nMBE	MAPE	RMSPE	R^2	nMBE	MAPE	RMSPE
No Weather Data	0.3223	0.84%	43.37%	81.01%	<0	−13.83%	52.26%	74.95%
HDD Only	0.9931	0.07%	2.25%	3.78%	0.9852	−3.82%	8.33%	13.62%
Engineered	0.9966	−0.10%	3.22%	8.77%	0.9951	−0.96%	3.76%	7.10%
Learned	0.9977	−0.03%	1.93%	2.60%	0.9971	−0.43%	2.94%	3.81%

engineered features. Comparing both, allows us to which extend the learned features overlap with our selection of features. Furthermore, another popular approach for understanding the behaviour of neural networks is to visualize the outputs of the hidden layers, called feature maps [53]. This is shown in Fig. 10.

4.2.1. Comparison to the engineered features

We compare engineered and learned features by feeding all 569 TMY data to the *ResNet* model and storing the associated values of the 13 neurons in the last dense hidden layer of *ResNet* which serve as input to the surrogate model (see Fig. 5). Then, we compare those values to all engineered features in a correlation map (see Fig. 9, left). This shows the correlation coefficients *R* of all learned features and all engineered features for the 569 weather files. By looking at the coefficients, we find that some learned features correlate well with the engineered features, in particular with the heating degree days (e.g. Feature 5). The second highest correlation is found for Feature 1 which is negatively correlated to the standard-deviation of the outdoor-air-temperature. Feature 12 exhibits a strong correlation with solar radiation and Feature 7 again with standard-deviation of temperature.

Interestingly, none of the learned features 0, 2 and 6 correlate with any of the engineered features. This indicates alongside the lower accuracy in predicting heating demand, that our set of engineered features misses some relevant weather impacts on the heating demand.

Features 3, 9, 10, and 11 do not show any activity at all. This shows that the number of encoded features (13) is high enough for the given task and could even be reduced. Minimizing the number of learned features was not considered in this study, but could be done in future work.

4.2.2. Feature map analysis

Another way to show the neural network behaviour is to look at the output of the convolutional kernels for different samples (compare to [24]). This is shown in Fig. 10. We show the final ReLU-activation of each of the three blocks of *ResNet*, as well as the normalized inputs and the geographic location of each weather file sample. Note that the building design is irrelevant as the building design parameters are inserted into the network after *ResNet*.

First, we look at the three columns showing plots of the neuron activity for the three *ResNet* blocks. Each row along the y-axis represents one kernel of a convolutional layer and the x-axis represents the time step at which the kernel window was applied. Looking at the second row from the top, the neuron activity for a weather file from a location in the far north in Canada with extremely low temperatures is shown. We find the three feature maps are very dark, meaning a lot of neurons are activated. We find that the activity fluctuates throughout the year with much higher neuron activity in the winter months than in summer months. This is different for the weather file sample at the top, where the neuron activity for a weather file from the West coast of Canada is shown. This part of Canada is known for mild climates. Generally, the neural activity is low and no seasonal behaviour in the neurons is found. Row 3 and 4 show locations from continental Canada and the East coast. They represent locations with average heating demand in Canada and the neuron activity is somewhat in between the

Table 3
R^2-score for all locations in the test data. On the left, the results of benchmarking the surrogate model trained on different sets of features are shown. This was done using the annual heating demand as the target variable. On the right, the performance of feature learning using *ResNet* on other building performance variables, i.e. annual cooling demand and hourly heating demand, is presented to highlight that *ResNet* is problem agnostic. Cities marked with an asterisk have a distance of more than 100 km to the closest weather file location in the training data.

Features:	No Weather	HDD	Engineered	ResNet		
Target variable:	Heating	Heating	Heating	Heating	Cooling	Hourly heating
Locations (HDD):				see Section 4.1 ←	→ see Section 4.3	
Winnipeg (5860)	.3197	.9143	.9923	**.9953**	.9932	.9416
Ottawa (4715)	.8876	.8826	.9898	**.9928**	.9923	.9316
Longueuil (4640)	.8398	.9115	.9912	**.9938**	.9934	.9245
Toronto (4170)	.2328	.9093	.9628	**.9920**	.9898	.9073
Taloyoak∗ (11500)	−6.1027	.9701	**.9873**	.9838	.8782	.8816
Lesage (5035)	.9016	.9477	**.9926**	.9839	.9959	.9206
Montreal (4950)	.9020	.9440	**.9915**	.9902	.9931	.9305
Norman Wells∗ (8150)	−2.6003	.9299	**.9917**	.9856	.9722	.9127
Edmonton (5660)	.8245	.9774	.9926	**.9943**	.9881	.9253
Victoria (2700)	−40.2160	.4630	.8659	**.9111**	.9905	.7355
Hamilton (3850)	-.7811	.9152	.9824	**.9940**	.9945	.9048
Halifax (3620)	−3.2721	.9735	.9616	**.9906**	.9927	.8809
London (4140)	.0346	.9223	.9806	**.9936**	.9896	.9134
Laval (4750)	.8337	.9642	.9909	**.9948**	.9939	.9255
Calgary (5210)	.8707	.9330	.9875	**.9933**	.9679	.9260
Saskatoon (5860)	.4315	.9580	**.9936**	.9848	.9879	.9304
Churchill∗ (9150)	−2.9467	.9822	.9903	**.9928**	.9574	.9107
Kuujuaq∗ (8570)	−2.2392	.9744	.9918	**.9938**	.9343	.9060
Vancouver (3060)	−10.0011	.9246	.9724	**.9872**	.9918	.7852
Fort Simpson∗ (7560)	−1.8688	.9336	.9922	**.9939**	.9836	.9348

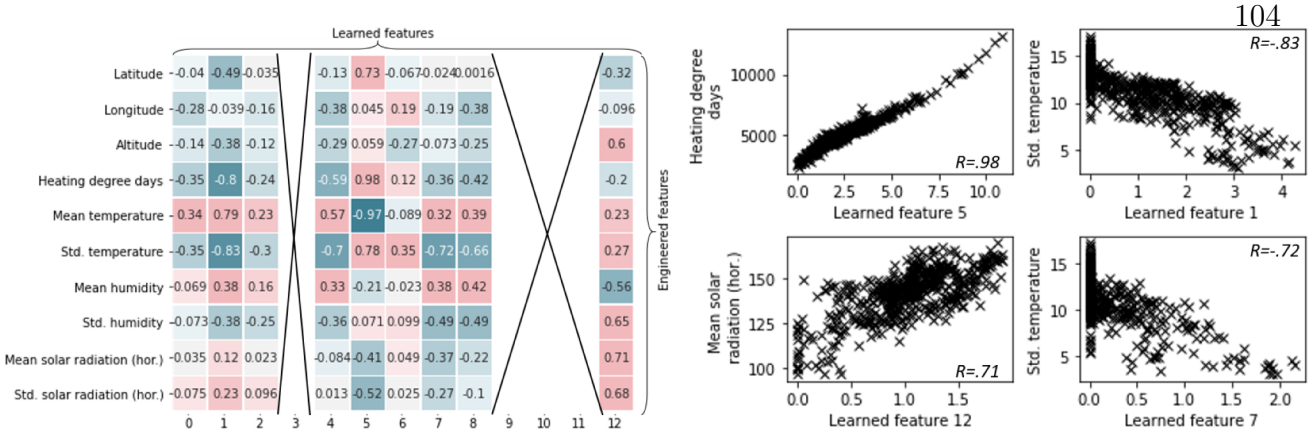


Fig. 9. Comparison of learned and engineered features. The heatmap on the left displays the correlation coefficient between the learned and the engineered features. Features with the highest correlation (positive or negative) are displayed in scatter plots the right.

first two locations.

Now we compare the neural activity for the different blocks. We find that in the first block the activity is rather blurry with less distinctive activity for the different seasons. Also, the 64 channels seem to be diverse in their behaviour while in block 2 and 3 many channels seem to correlate. In the latter blocks, two types of channels can be found. While most channels spike during winter months, some rather spike during summer months. No channels are found with constant behaviour throughout the year.

Another interesting aspect of the feature maps is the temporal resolution of the neural activity. Not only, seasonal fluctuations in activity are found but also some neurons spike only for certain hours of the year. This is specifically found in the last block, where individual neurons spike throughout the year. This motivated us to train the given *ResNet* architecture also to predict the heating demand with an hourly resolution, which we introduce in the following section.

4.3. Feature learning results for other performance objectives

In the following two experiments, we show that the presented approach can be used in the same way as before to also estimate cooling demand and hourly heating demand. This is a first step to show that our method based on *ResNet* is problem-agnostic, and simulations of other performance metrics like overheating hours, air quality, etc. for multiple locations can be emulated in future.

4.3.1. Cooling demand

To train the model on estimating annual cooling demand, the entire model architecture (*ResNet* and surrogate model) remains the same and we only exchange the target variable. After training, we reach a similar prediction accuracy on the test (training) data of $R^2 = 0.991$ and nMBE= 0.241% ($R^2 = 0.995$, nMBE= 0.041%) as we did for estimating heating demand. As in some Canadian climates the cooling demand approaches zero, our normalized accuracy metrics (*MAPE* and *RMSPE*) become very large due to division by very small values and were omitted in this section.

We also report the R^2 score for each location in the test data (see Table 3). The accuracy scores are computed based on 500 different building designs simulated for each location. The cooling demand accuracy scores behave similarly to the heating demand scores. Low scores are found for places with low cooling demand in the North of Canada (HDDs>8000, e.g. Taloyoak, Kuujuaq, Churchill). In case of heating demand, the surrogate model accuracy is lower for Victoria and Vancouver where heating demand is relatively low (HDDs≈3000). This shows that the surrogate model struggles to capture diminishing effects of changes in building design on heating demand (cooling demand) in hot climates (cold climates).

4.3.2. Hourly heating demand

In this section, we scrutinize the ResNet-based surrogate modelling performance in estimating heating demand with higher temporal resolution. Hourly heating demand estimates can be very helpful for many applications. For example, it allows fast, optimized heating system layout on building and district level [52], or to analyse demand response potential [54].

While other approaches for heating demand estimation are specifically designed for aggregated annual or monthly energy demand prediction, the invariant kernel size with 8760 neurons in each channel allows us to estimate 8760 hourly heating demand values as outputs, i.e. we produce an output sequence given an input sequence. CNNs have recently been shown to reach high accuracy on sequence-to-sequence modelling tasks and outperformed recurrent neural networks on tasks where the format and length of the sequential input data (here the weather file) do not change [55].

Shifting from one output to a sequence of 8760 outputs, we slightly adjust our model's architecture. We remove the pooling layer and replace the one-dimensional concatenation layer with a concatenation of the layer with 128x8760 encoded weather features and the 13 building design parameters (at each time step). Additionally, we replace the feed-forward surrogate model architecture, with a convolutional neural network. Instead of two fully-connected layers, we use two CNN blocks similar to the ones of ResNet. This provides a final network of 12 layers. Note, that this neural network architecture and its parameters were not optimized but should rather serve as a first proof of how suitable the proposed approach is for sequence-to-sequence modelling of building time series.

Training was performed in 500 epochs. We observe an overall R^2 (nMBE) score on the testing data of 0.923 (nMBE= −7.6%) and the accuracy score at each of the testing locations varies between 0.7355 in Victoria and 0.9416 in Winnipeg (see Table 3, right). In particular, at few locations with rather mild heating demands (Victoria, Vancouver), hourly surrogate estimates were found to differ significantly from the simulation outcomes. More model refinement or a more balanced climate data set with an equal share of mild and cold climate weather files could solve that.

As an example, we show the surrogate model estimates for one location for one year and one week in Fig. 11. The surrogate model is capable to predict the seasonalities in the demand and also correctly identifies most of the peaks as shown in the second plot. However, we find that in summer, when heating demand converges to zero, the surrogate fails to correctly predict small sudden increases in heating demand. Given the low absolute heating demand during summer, normalized error metrics (MAPE, RMSPE) become very large and therefore, we omitted when reporting surrogate accuracies in this section.

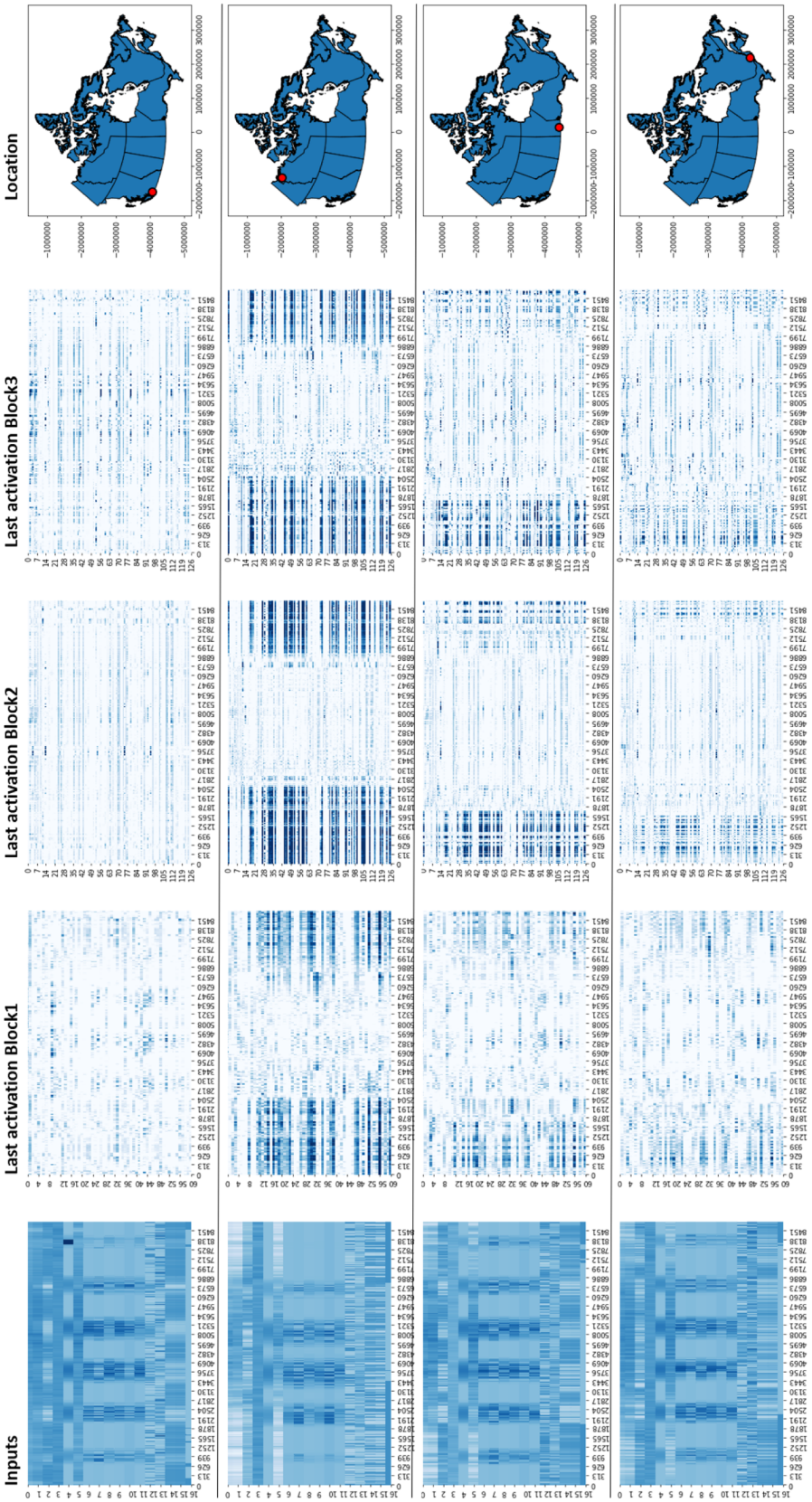


Fig. 10. Inside the convolutional neural network. The outputs of the last layer in each block of *ResNet* when applied to weather data from locations with mild and extreme climates. Each row in each of the rectangular heat maps is the output for one kernel applied to each time step of the time series.

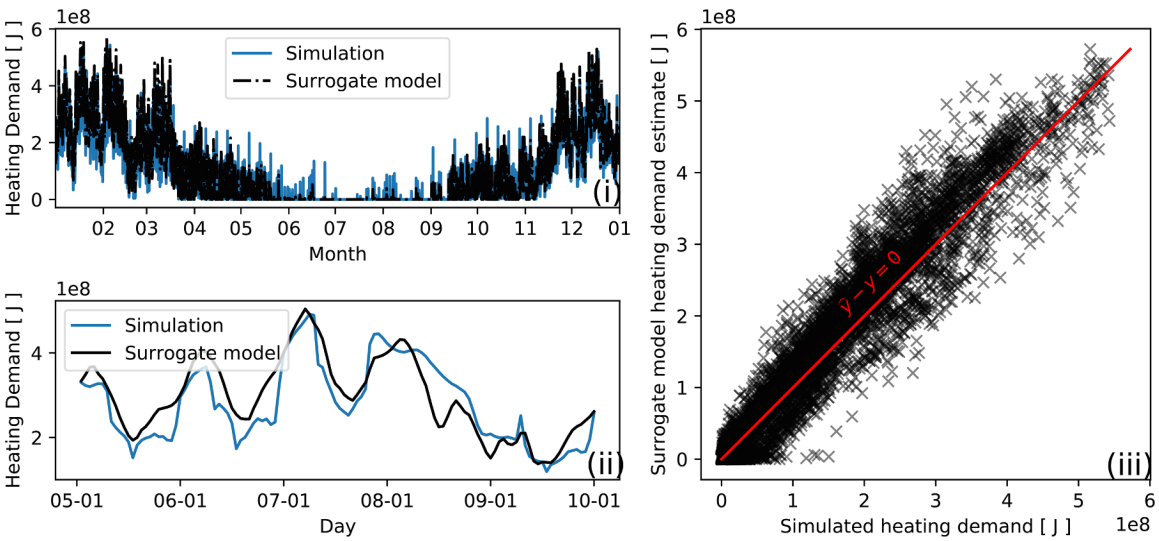


Fig. 11. Comparison of simulation and surrogate model estimates the TMY of Winnipeg. *(i)* shows all 8760 hourly estimates and true values in a sequential format, *(ii)* shows one week of data, and *(iii)* compares 8760 surrogate estimates and simulation outcomes in a scatter plot, where a perfect fit would put all points on the red line.

5. Discussion

In the previous section, we derived location-independent surrogates which emulate simulated energy performance of buildings with various designs at 569 locations in Canada. In the following few paragraphs, we discuss the accuracy of the derived surrogate models with different features as inputs and different performance metrics as outputs, we further elaborate on a key aspect of surrogate modelling, the achieved reduction in computational cost by using a surrogate model over running building simulations. Last, we consider how the approach can serve as a basis for a much wider application of surrogate models to assess building energy performance.

5.1. Accuracy

We reported surrogate accuracy scores for surrogates to estimate (i) annual heating demand (with engineered and learned features), (ii) annual cooling demand and (iii) hourly heating demand.

First we could show that a surrogate provided with learned features performs better than the same surrogate provided with a set of manually, engineered features (see Table 3). This lets us conclude that automated, feature learning cannot only compete with manual feature selection but also outperform it. Hence, less manual feature selection and higher surrogate accuracy go hand in hand.

We reached a mean absolute percentage error of 2.94% when estimating heating demand for new designs and climates. This corresponds to an R^2 score of 0.997 on the test data consisting of unseen building designs at unseen locations. That score is in line with other publications where the building location was not varied [20,9] and outperforms a study [36] where different climates but another performance metric, overheating hours, was considered ($R^2 = 0.971$). As a conclusion, our method of weather feature learning allows us to use surrogate models at various locations without compromising surrogate model accuracy.

To better generalize the findings, we also used the feature learning

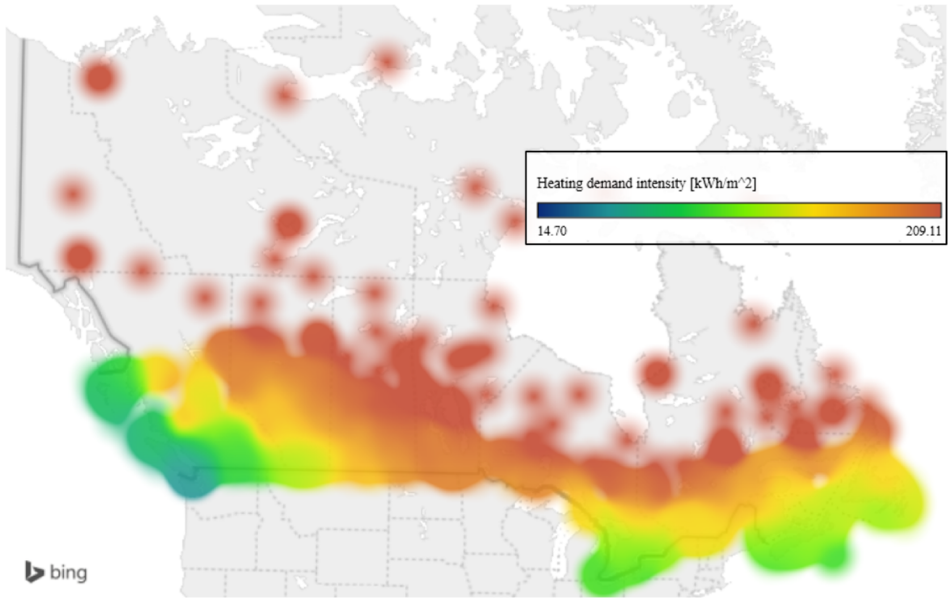


Fig. 12. Surrogate model heating demand estimates for Canada. The runtime to evaluate the performance of one specific building design sample at the 569 locations of the case study takes approximately 2 s.

approach to predict a different simulation output, i.e. cooling demand ($R^2 = 0.987$), and changed the temporal resolution (from annual to hourly, $R^2 = 0.923$). The former implies that the approach scales to various aggregated annual performance metrics. Estimating hourly performance inevitably lead lower performance but it still exceeds the reported number in the only other study where hourly thermal demand for different climates was estimated ($R^2 = 0.85$) [33]. In their case manual feature selection was done.

5.2. Computational cost for evaluation and training

Computational cost of surrogate modelling must be split into the cost of evaluating an already trained surrogate model, and the training time itself. The former is far more important, as it is directly linked to the time a building designer has to wait for performance estimates for her proposed design. The latter can even be regarded negligible, as a highly generalized surrogate model is trained once and afterwards only used to produce performance estimates. Nonetheless, training time is provided for completeness.

5.2.1. Evaluation

In our modelling approach, evaluating the large *ResNet* to extract weather features takes much longer than evaluating the surrogate model itself. Extracting the features of the 569 weather files in the case study and evaluating one building design per location (see Fig. 12) takes around 2.0 s with the TESLA K80 GPU (4992 cores @ 1253 MHz, 480 GB/s, 24 GB VRAM)[8] and around 11.5 s using CPU only (6 2.1 GHz CPUs, 24 GB RAM), whereas evaluating the surrogate at one location (i.e. only one *ResNet* evaluation) for 569 different building designs takes only around 70 ms with GPU and 80 ms using CPU.

The latter shows that the cost of weather feature extraction are insignificant in most building design analyses as usually designs at one specific location are evaluated. The resulting speed allows us to interactively design a building with close to instantaneous performance feedback. Comparing the runtime to building simulation software (here EnergyPlus, around 50 min for 569 runs with 6 2.1 GHz CPUs) we achieve time savings of $4.3 * 10^4$.

5.2.2. Training

Automatically learning features with a temporal convolutional neural network comes at high computational cost. The presented TCN architecture has more than 500, 000 parameters allowing to process 150,000 inputs from 17 time series variables. Training that network required 8 h on a server with a TESLA K80 GPU.

Without feature learning but by using engineered features instead, the heating demand surrogate model is trained in about 4 min while the accuracy only suffers a little (R^2 similar, *MAPE* from 2.95% to 3.74%, *RMSPE* from 4.51% to 10.23%). Looking at the worst performing climate, the errors do increase significantly (Victoria, e.g. 14.20% to *RMSPE* 35.12%). Also, the time to pick and extract engineered features was not considered, where this fully avoided in case of automated feature learning.

5.3. Scalability of the approach

Surrogate models are being used to accelerate and enable large scale building design space exploration. They already have shown their great use to help building designers and architects on building level [10,11] and city level [33]. However, in most of the studies the surrogates were trained for a specific building project. Offering generalized, already trained models off-the-shelve, will make surrogates more accessible to architects and designers without extensive machine learning domain knowledge.

[8] Note, that cheaper and faster GPUs are widely available.

Our approach covers one option to improve the level of generalization of surrogates. Using a CNN, we increased the surrogates' geographical scope from a single location to multiple ones. When we consider the following list of next steps, the use of our method for the building domain can be significant:

- More surrogate inputs:
  In this study, we derived a surrogate model architecture that can handle a combined set of diverse inputs, i.e. large, multivariate time series data and static building design parameters. Both types of inputs can be augmented: We limited the number of design parameters to thirteen which for example do not allow to adjust the building's geometry. However, in other research it was already shown that neural network surrogate models can handle much larger number of parameters well, which should also apply to our architecture. Furthermore, the set of time series processed by the CNN could potentially integrate other dynamic factors impacting building energy demand like occupancy profiles or internal gain profiles.
  Apart from that, neural networks have proven to be modular. In this study, we use a deep CNN to extract weather features and concatenate them with building design featuers. Similarly, other authors compartementalized neural networks into functional units [31]. We foresee that our work can be combined with similar work where surrogate models generalized over various geometries [33,49] or mechanical system setups [35].
- More climates:
  In this study, we considered a large variety of climates found in Canada, the second biggest country in the world. We see no obstacles that our approach will allow to develop a surrogate model that spans entire continents or even the entire globe.
- Larger variety of performance metrics:
  A key value of feature learning is that we can automate the feature extraction step. We showcased that we can automatically learn features to estimate various building performance metrics (heating demand, cooling demand) with different temporal resolution (annual and hourly). This automation in feature extraction may help to scale location-independent surrogates to a holistic set of building performance outputs without any additional manual work. In future work, we will quantify the performance a location-independent surrogate with other targets like occupant comfort or natural ventilation performance [36].

6. Conclusion

In this paper we contribute with a method that increases the geographical scope of a single building energy surrogate model to arbitrary many locations. We are among the firsts in the domain to use a deep convolutional neural network to extract features from multivariate, hourly weather data, which capture the impact of location-specific climate on building energy performance.

In a case study with 569 weather files from Canada, we could show that the feature learning approach outperformed a manual selection of weather features (variables) to estimate annual heating demand. A mean percentage error to the physics-based simulation software of less than 3% on a test data set was reached. The set of manually selected features covers common ones found in literature including heating degree days, mean temperature, standard deviation of temperature, average humidity and other variables. We collected the values of the features produced by the convolutional network for various climates and found some of them correlating to the engineered features which indicates that the network learned physically meaningful features. On the other hand, other learned features showed weak correlation to manually selected ones, implying that they captured information going beyond the manually selected ones.

Finally, our experiments suggest that our location-independent surrogate models are problem-agnostic. We could confirm the high

accuracy when estimating not only annual heating demand but also annual cooling demand, and sequential hourly heating demand estimates.

We see a large potential of surrogate models which generalize over multiple climates, building systems and geometries to play an elemental role in building design processes in future. The generalization will allow building designers and architects without extensive machine-learning knowledge to use them off-the-shelve for fast performance feedback on their design ideas.

7. Code and Data availability

The entire source code of this work, the EnergyPlus description file (*.idf*) of the building template, and instructions on how to download the required weather files are provided in a GitLab repository.[9]

CRediT authorship contribution statement

Paul Westermann: Methodology, Software. **Matthias Welzel:** Methodology, Software. **Ralph Evins:** Supervision, Resources, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by grant funding from CANARIE via the BESOS project (CANARIE RS-327).

References

[1] C.D. John Dulac, Thibaut Abergel, Tracking buildings, Tech. rep., Internation Energy Agency (2019). URL https://www.iea.org/reports/tracking-buildings.
[2] B. Energy Step Code Council, Bc energy step code, a best practices guide for local governments, Tech. rep., Energy Step Code Council, Building and Safety Standards Branch (Jul. 2019).
[3] Crawley DB, Pedersen CO, Lawrie LK, Winkelmann FC. Energyplus: energy simulation program. ASHRAE J 2000;42(4):49.
[4] IES Virtual Environment (2020). URL https://www.iesve.com/.
[5] Hensen JL, Lamberts R. Building performance simulation for design and operation. Routledge; 2012.
[6] Attia S, Gratia E, De Herde A, Hensen JL. Simulation-based decision support tool for early stages of zero-energy building design. Energy Build 2012;49:2–15.
[7] Clarke JA. Energy simulation in building design. Routledge; 2001.
[8] Wang GG, Shan S. Review of metamodeling techniques in support of engineering design optimization. J Mech Des 2007;129(4):370–80.
[9] Westermann P, Evins R. Surrogate modelling for sustainable building design – a review. Energy Build 2019;198:170–86. https://doi.org/10.1016/j.enbuild.2019.05.057.
[10] Brown NC. Design performance and designer preference in an interactive, data-driven conceptual building design scenario. Des Stud 2020.
[11] Hester J, Gregory J, Kirchain R. Sequential early-design guidance for residential single-family buildings using a probabilistic metamodel of energy consumption. Energy Build 2017;134:202–11. https://doi.org/10.1016/j.enbuild.2016.10.047. URL < Go to ISI > ://WOS:000390624800018.
[12] Eisenhower B, O'Neill Z, Fonoberov VA, Mezić I. Uncertainty and sensitivity decomposition of building energy models. J Build Perform Simul 2012;5(3):171–84.
[13] Rivalin L, Stabat P, Marchio D, Caciolo M, Hopquin F. A comparison of methods for uncertainty and sensitivity analysis applied to the energy performance of new commercial buildings. Energy Build 2018;166:489–504.
[14] Tian W, Choudhary R, Augenbroe G, Lee SH. Importance analysis and meta-model construction with correlated variables in evaluation of thermal performance of campus buildings. Build Environ 2015;92:61–74. https://doi.org/10.1016/j.buildenv.2015.04.021. URL < Go to ISI > ://WOS:000358807800007.
[15] Tsanas A, Xifara A. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy Build 2012;49:560–7. https://doi.org/10.1016/j.enbuild.2012.03.003. URL < Go to ISI > ://WOS:000305875500065.
[16] Eisenhower B, O'Neill Z, Narayanan S, Fonoberov VA, Mezic I. A methodology for meta-model based optimization in building energy models. Energy Build 2012;47:292–301. https://doi.org/10.1016/j.enbuild.2011.12.001. URL < Go to ISI > ://WOS:000301989800034.
[17] Magnier L, Haghighat F. Multiobjective optimization of building design using trnsys simulations, genetic algorithm, and artificial neural network. Build Environ 2010;45(3):739–46. https://doi.org/10.1016/j.buildenv.2009.08.016. URL < Go to ISI > ://WOS:000272307700025.
[18] Prada A, Gasparella A, Baggio P. On the performance of meta-models in building design optimization. Appl Energy 2018;225:814–26.
[19] Nagpal S, Mueller C, Aijazi A, Reinhart CF. A methodology for auto-calibrating urban building energy models using surrogate modeling techniques. J Build Perform Simul 2019;12(1):1–16.
[20] Ostergard T, Jensen RL, Maagaard SE. A comparison of six metamodeling techniques applied to building performance simulations. Appl Energy 2018;211:89–103. https://doi.org/10.1016/j.apenergy.2017.10.102. URL < Go to ISI > ://WOS:000425075600008.
[21] Van Gelder L, Das P, Janssen H, Roels S. Comparative study of metamodelling techniques in building energy simulation: Guidelines for practitioners. Simul Model Pract Theory 2014;49:245–57. https://doi.org/10.1016/j.simpat.2014.10.004. URL < Go to ISI > ://WOS:000345153100018.
[22] Geyer P, Schlueter A. Automated metamodel generation for design space exploration and decision-making - a novel method supporting performance-oriented building design and retrofitting. Appl Energy 2014;119:537–56. https://doi.org/10.1016/j.apenergy.2013.12.064. URL < Go to ISI > ://WOS:000333506900050.
[23] L.K. Lawrie, D.B. Crawley, Development of global typical meteorological years (tmyx). (Jun. 2019). URL http://climate.onebuilding.org.
[24] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.
[25] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016.
[26] S. Bai, J.Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, arXiv preprint arXiv:1803.01271.
[27] Fawaz HI, Forestier G, Weber J, Idoumghar L, Muller P-A. Deep learning for time series classification: a review. Data Min Knowl Disc 2019;33(4):917–63.
[28] Forrester A, Keane A, et al. Engineering design via surrogate modelling: a practical guide. John Wiley & Sons; 2008.
[29] Garud SS, Karimi IA, Kraft M. Design of computer experiments: A review. Comput Chem Eng 2017;106:71–95.
[30] Trčka M, Hensen JL, Wetter M. Co-simulation of innovative integrated hvac systems in buildings. J Build Perform Simul 2009;2(3):209–30.
[31] Geyer P, Singaravel S. Component-based building performance prediction using systems engineering and machine learning. Appl Energy 2017;228:1439–53.
[32] Gratia E, De Herde A. A simple design tool for the thermal study of an office building. Energy Build 2002;34(3):279–89.
[33] J. Vazquez-Canteli, A.D. Demir, J. Brown, Z. Nagy, Deep neural networks as surrogate models for urban energy simulations, in: Journal of Physics: Conference Series, Vol. 1343, IOP Publishing, 2019, p. 012002.
[34] Heo Y, Choudhary R, Augenbroe G. Calibration of building energy models for retrofit analysis under uncertainty. Energy Build 2012;47:550–60.
[35] Korolija I, Zhang Y, Marjanovic-Halburd L, Hanby VI. Regression models for predicting uk office building energy consumption from heating and cooling demands. Energy Build 2013;59:214–27.
[36] Rackes A, Melo AP, Lamberts R. Naturally comfortable and sustainable: Informed design guidance and performance labeling for passive commercial buildings in hot climates. Appl Energy 2016;174:256–74. https://doi.org/10.1016/j.apenergy.2016.04.081. URL < Go to ISI > ://WOS:000377728700022.
[37] Brown MA, Cox M, Staver B, Baer P. Modeling climate-driven changes in us buildings energy demand. Clim Change 2016;134(1–2):29–44.
[38] Hygh JS, DeCarolis JF, Hill DB, Ranjithan SR. Multivariate regression as an energy assessment tool in early building design. Build Environ 2012;57:165–75. https://doi.org/10.1016/j.buildenv.2012.04.021. URL < Go to ISI > ://WOS:000307618900017.
[39] Jaffal I, Inard C. A metamodel for building energy performance. Energy Build 2017;151:501–10. https://doi.org/10.1016/j.enbuild.2017.06.072. URL < Go to ISI > ://WOS:000410010400044.
[40] Lam JC, Wan KK, Liu D, Tsang C. Multiple regression models for energy use in air-conditioned office buildings in different climates. Energy Convers Manage 2010;51(12):2692–7.
[41] Romani Z, Draoui A, Allard F. Metamodeling the heating and cooling energy needs and simultaneous building envelope optimization for low energy building design in morocco. Energy Build 2015;102:139–48. https://doi.org/10.1016/j.enbuild.2015.04.014. URL < Go to ISI > ://WOS:000358458100012.
[42] Y. Bengio, Deep learning of representations for unsupervised and transfer learning, in: Proceedings of ICML workshop on unsupervised and transfer learning, 2012, pp. 17–36.
[43] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: a system for large-scale machine learning., in: OSDI, Vol. 16, 2016, pp. 265–283.
[44] F. Chollet, et al., Keras (2015).
[45] Z. Wang, W. Yan, T. Oates, Time series classification from scratch with deep neural networks: A strong baseline, in: 2017 international joint conference on neural networks (IJCNN), IEEE, 2017, pp. 1578–1585.
[46] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition, in. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770–8.
[47] Edwards RE, New J, Parker LE, Cui B, Dong J. Constructing large scale surrogate

[9] https://gitlab.com/energyincities/building_surrogate_modelling.

models from big data and artificial intelligence. Appl Energy 2017;202:685–99. https://doi.org/10.1016/j.apenergy.2017.05.155. URL < Go to ISI > :// WOS:000407188500055.
[48] Chai T, Draxler RR. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. Geoscient Model Develop 2014;7(3):1247–50.
[49] Singaravel S, Suykens J, Geyer P. Deep convolutional learning for general early design stage prediction models. Adv Eng Inform 2019;42:100982.
[50] I.J. Hall, R. Prairie, H. Anderson, E. Boes, Generation of a typical meteorological year, Tech. rep., Sandia Labs., Albuquerque, NM (USA) (1978).
[51] Stein M. Large sample properties of simulations using latin hypercube sampling. Technometrics 1987;29(2):143–51.
[52] Evins R, Orehounig K, Dorer V, Carmeliet J. New formulations of the 'energy hub'model to address operational constraints. Energy 2014;73:387–98.
[53] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson, Understanding neural networks through deep visualization, arXiv preprint arXiv:1506.06579.
[54] Bianchini G, Casini M, Vicino A, Zarrilli D. Demand-response in building heating systems: A model predictive control approach. Appl Energy 2016;168:159–70.
[55] Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN. Convolutional sequence to sequence learning. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org; 2017. p. 1243–52.

Epilogue

Climate-independent surrogate models are a significant step towards more general surrogate models that can be applied to a multitude of design studies. It is very relevant for applications like the NetZero Navigator (see Section 3), where a surrogate model is provided to building designers to analyse their own design problems quickly. By using the climate-independent surrogate they can provide the weather file from their own location and immediately receive performance estimates.

A current drawback of the surrogate is that it cannot model other dynamic impacts like occupant behaviour or the impact of the built environment. This includes that shading effects or wind flow blocking from neighbouring buildings. Unpublished work has shown that image generation algorithms allow to generate fast daylighting and fluid-dynamics maps for urban environments.[1]

A promising outlook of the research is that temporal convolutional networks allow a surrogate model to learn various dynamic impacts on the target variable. We foresee that analogously to weather data other time series like occupancy profiles or appliance load profiles can be incorporated into surrogate model approximations.

[1] http://cities.ait.ac.at/site/index.php/2019/10/31/cil-opening-design-space-exploration/

Part II

Surrogate modelling for building calibration

Chapter 6

Surrogate-based model calibration

Surrogate model-based calibration of building energy models is a way to efficiently connect the observed world with physics-based building energy models. The calibrated energy models allow us to model the effect of various retrofit options for existing buildings. Surrogate models are promising in that context as they may increase the speed of calibration processes, which enables to assess retrofit potentials for entire building stocks.

However, model calibration faces challenges when applied to large number of buildings. This involves that a suitable base model, which is calibrated with measured data, needs to be found. Often, researchers derive archetype models whose characteristics are retrieved from building stock databases [25]. However, these databases may lack important information, like the buildings' heating system, and vary from one region to another. Instead, unsupervised learning techniques are receiving growing attention to automatically extract discrete building characteristics [17][22].

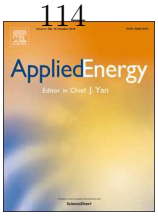
In the following study, we developed an unsupervised method to extract the building type and heating system type using smart meter data only.

Applied Energy 264 (2020) 114715

Contents lists available at ScienceDirect

Applied Energy

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



Unsupervised learning of energy signatures to identify the heating system and building type using smart meter data

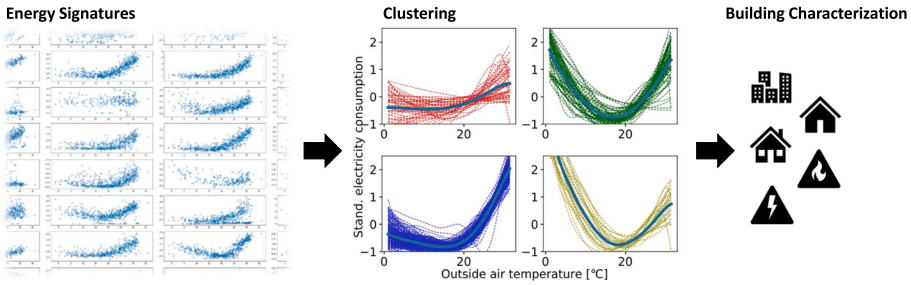
Paul Westermann[a,b,*], Chirag Deb[b], Arno Schlueter[b], Ralph Evins[a]

[a] Energy in Cities Group, Department of Civil Engineering, University of Victoria, Canada
[b] Architecture and Building Systems, Institute of Technology in Architecture, ETH Zurich, Switzerland

HIGHLIGHTS

- Automated method to identify building characteristics using smart meter data.
- Energy signatures for hundreds of building generated and clustered by shape.
- Infer the heating system and the building type from the energy signature shape.
- Two case studies conducted totalling 889 buildings.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Smart meter data
Energy signatures
Unsupervised learning
Dynamic time warping
Clustering
Data mining
Machine learning

ABSTRACT

A high-quality building energy retrofit analysis requires knowledge of building characteristics like the type of installed heating system. This means auditing the building in person or conducting a detailed survey, which is not readily scalable for many buildings.

This paper presents a data-driven methodology to identify building characteristics from raw smart meter data sets to allow large scale, high-quality building retrofit analysis. We use the concept of energy signatures, a scatter plot with outside air temperature on the x-axis and electricity consumption on the y-axis, which condenses each building's electricity use into one highly informative graph. Using a Support-Vector Regression model we extract the shape of each signature and cluster them subsequently. Dynamic time warping is used to align the signature shapes of all buildings. In two case studies, consisting of smart meter data sets from 408 and 480 buildings respectively, we show that our clusters correlated well to the heating system type and the building type by comparing to building-level metadata or demographic data.

1. Introduction

Retrofitting the existing building stock is a key challenge to enable a global clean energy transition as buildings account for 25% of global carbon emissions [1]. The International Energy Agency (IEA) recently pointed out that the transformation process of the building sector is slow and lags behind the carbon-reduction targets as defined in the Paris Agreement [2].

One important reason may be the need for customized building-level performance analysis to assess the cost-effectiveness of carbon reduction or energy conservation measures.

Traditional retrofit decision making is to a large extent based on

Abbreviations: DTW, dynamic time warping; HVAC, heating, ventilation and air conditioning; OAT, outside-air-temperature; ES, Energy Signature; M&V, measurement and verification; SVR, support-vector regression

* Corresponding author at: Energy in Cities Group, Department of Civil Engineering, University of Victoria, Canada.
E-mail addresses: pwestermann@uvic.ca (P. Westermann), deb@arch.ethz.ch (C. Deb), schlueter@arch.ethz.ch (A. Schlueter), revins@uvic.ca (R. Evins).

https://doi.org/10.1016/j.apenergy.2020.114715
Received 5 November 2019; Received in revised form 21 January 2020; Accepted 21 February 2020

expert knowledge and involves time-intensive steps including on-site building audits to collect building characteristics and possibly, custom building retrofit performance analysis to quantify the potential of each retrofit measure (e.g. [3]). This is hardly scalable to entire building stocks and hence there is a rapidly growing field of research attempting to automate the process of retrofit analysis [4]. This involves two fundamental tasks directly related to traditional retrofit design.

(a) Automation of the collection or estimation of building characteristics from available data sources.
(b) Automation of the retrofit performance analysis to identify viable retrofit measures for a specific building.

Many methods have been developed to automate the retrofit performance analysis (b) and they are further improving with the rise of machine learning methods [5,6]. However, many of them share the drawback that they *"require complex characteristic data about each building such as geometric dimensions, building materials, the age and type of mechanical systems, and other metadata to execute the process"* [7]. Clearly, automating the collection of building characteristics (a) is a bottleneck for large scale building retrofit analysis.

This study proposes a novel, automated method to find building characteristics using only raw smart meter and high-level outside air temperature data. We plot the energy signatures for each building in a building stock, group them based on shape and infer the building type and heating system. Energy signatures have widely been used for whole building parameter estimation (e.g. [8]). However, we employ them as a way to capture each building's thermal characteristics in an automated manner, which forms the basis of this paper.

The paper is structured as follows. In Section 2, we give an overview of existing methods for automated building characterization and introduce the concept of energy signatures. In Section 3, we present our method to cluster buildings based on their energy signature shape including the data preprocessing steps, the energy signature shape extraction, and the clustering approach. Next, we introduce two smart meter datasets for residential buildings in Austin (Texas) [9] and Vancouver Island (British Columbia) which we use as test cases for our method. The first case study comprises buildings with varying type and different heating systems installed, while the second case study includes only electrically heated single/duplex buildings (heat pumps, resistance heaters). The results of the two case studies are given in Section 4. Based on these, we discuss our new method and highlight its potential application for detailed building retrofit analysis, for building benchmarking and as a data source for policy design (Sections 5, 6).

2. Background

2.1. Automated building characterization using smart meter data

To be scalable, a method for automated building characterization requires widely available data sources. With the recent increase in the distribution of sensors in buildings, large amounts of raw data are accumulating, and have already been leveraged to extract information on buildings [7]. Alongside digital thermostats collecting temperature data, advanced electricity metering infrastructure, often called smart meters, is forecasted to be installed in one billion buildings by the end of 2020 [10].

Smart meter data is an attractive raw data source as the data acquisition is highly standardized [11]. That advantage puts smart meters apart from other sources like customized building-specific sensor networks where the number, location and types of sensors may vary [12]. Any data mining method relying only on smart meter data only is scalable to all the buildings where smart meters are installed, a large and growing cohort.

In Fig. 1, we summarize existing methods to extract characteristics of buildings from smart meter data, and split them into two groups. Either the temporal patterns of the electricity use are analysed (*time-of-use*), or the electricity use at certain weather conditions is examined. Given the importance of outdoor air temperature to the thermal performance of buildings, we call this approach *temperature-at-use*.

2.1.1. Time-of-use methods

Time-of-energy-use patterns in buildings have been used for extracting socio-economic characteristics [13]. Different daily load profiles indicate the building's primary use type, for example residential, educational and government buildings [14], or laboritories and dormitories [15]. Furthermore, in residential buildings factors like employment status, number of bedrooms, age of occupants, household composition, social class, water heating type and cooking type also have significant influence on time-of-use behaviour [16,17]. Based on those findings, customers can be classified into certain groups of socio-economic characteristics given their time-of-use profile [18].

Another related field is smart meter based load disaggregation, however, it requires higher frequency data (e.g. 5 min) compared to the data available in this study (hourly) [19].

2.1.2. Temperature-at-use methods

Temporal patterns are of limited use to analyse thermal buildings characteristics. Instead, knowing the weather conditions causing a certain building energy consumption is more relevant, and is often leveraged for data-driven performance analysis of existing buildings. Many approaches are thoroughly reviewed in [20] and may be split into quantitative and qualitative methods.

In *quantitative methods*, the temperature and energy use data are used to calibrate some parameters of a physics-based or semi-physics-based building energy model [21]. In the most complex case, a dynamic model is calibrated with a varying number of parameters [22]. Both model selection and data quality are important to obtain stable parameter estimates, and this may include a lot of manual work [23]. The accuracy of the calibration process improves the more information on the building is available, as shown in a case study in [24]. Other calibration studies also rely on extensive knowledge on one specific building [3]. This highlights the motivation of this paper: to augment the set of automatically retrievable building characteristics to enable large-scale building performance analysis.

In comparison to dynamic models, steady-state models tend to be more stable and require lower data quality. One common approach is to average smart meter recordings to daily or monthly values [25] to eliminate transient and higher-order terms from the building energy equations [26]. It is usually applied to estimate whole building parameters, for example the building's base-load, the heat loss rate and the change-point temperature [27], or the cooling rate [28]. The approach relies on piecewise linear regression models often using a univariate model with the outside air temperature as the independent variable and the energy consumption as dependent variable [29]. One can visualize this approach using a simple scatter plot with temperature on the x-axis and energy consumption on the y-axis (s. Fig. 2). This plot is often referred to as the *energy signature* of a building. In the case of a univariate model, multiple linear segments are fitted to the cloud of points. Variables other than outside air temperatures like solar gains, thermal mass [30] and occupancy [31] can be incorporated. As an example, steady-state approaches have been used to assess the success of retrofits within a measurement and veri?cation (M&V) scheme. Measurements after the retrofit are compared to the regression model predictions, which were fitted to data collected before the retrofit[32,31]. Although the approach already is highly simplified, some knowledge on each building is still required. For example, the heat loss estimate is affected by the heating system efficiency, and a lack of knowledge of either the

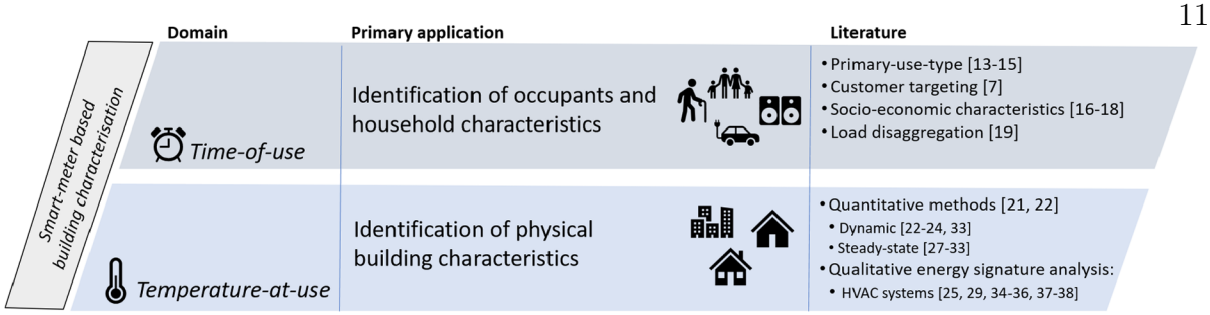


Fig. 1. Examples of smart meter building analyses grouped into two categories.

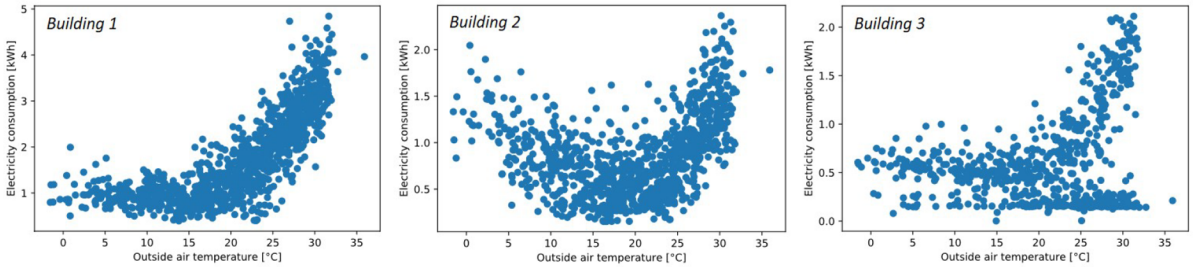


Fig. 2. Three energy signature samples from a building stock with hundreds of buildings [9]. The observed daily mean electricity consumption and outside air temperature are compared.

envelope or the heating system may lead to wrong conclusions [33].

While the previous literature addresses quantitative whole building analysis for the most part, it has been observed that energy signatures are shaped by *qualitative* characteristics of a building like the type of HVAC system installed. For example, when fitting a piecewise linear regression model to the energy signature, [29] stated that having five regression segments is most suitable for *"buildings with electric heat pumps or both electric chillers and electric resistance heating"*. Similarly, [34] outlines that the curvature of energy signatures provides information on whether multiple heating systems are installed, or [35,36] reported how the signature shape of a heat pump differs from other heating systems. Similarly, HVAC engineers have been comparing measured signatures to ones generated with a simulation model [37,38]. The difference between the signatures gave details on where the simulation model is wrong, or if the installed HVAC system is defective. While these authors only considered individual buildings, [25] showed how energy signatures could potentially separate thousands of electrically heated from non-electrically heated buildings, although validation data was missing for their study.

2.2. Reading the energy signature for building characterization

Reviewing the initial studies, we envision the energy signature as a highly informative tool to retrieve qualitative building characteristics. In this study, we develop a method to analyse prevailing energy signature shapes in a building stock to infer building type and heating system type for hundreds of buildings. In our model-free, non-linear approach we capture the shape of an energy signature in a building stock and use the shape as the input to a cluster analysis. This step extracts the typical signatures found in a building stock and allows us to group buildings. Available meta data on the buildings is compared to the clusters, and we can show that the method groups buildings of the same type (apartment or single/duplex building) and with similar HVAC systems. This study relies on an unsupervised clustering approach, but in future could form the basis for a supervised learning approach to predict building characteristics based on energy signatures.

3. Methodology

An overview of the methodology is given in Fig. 3. After data preparation and preprocessing we apply two machine learning steps.

First, we normalize the electricity use data of each building using the z-score to enable the comparison of energy signatures collected from buildings with varying magnitude in their electricity consumption. Then we fit a univariate, non-linear regression model to estimate electricity use based on outside-air-temperature (OAT) and evaluate it at 0.5 °C increments. The regression model fitting and subsequent evaluation serve as a means of feature extraction and dimensionality reduction to capture the shape of an energy signature in a one-column vector.

Finally, the shape vectors of all buildings are grouped into C clusters to represent dominant signature shapes among the building stock.

3.1. Data

We apply the method on data sets of two building stocks. One data set was collected in Austin, TX, USA, and is cooling dominated, while the other data set is from Vancouver Island, BC, Canada, and is heating dominated. Further details on each data set are given in the associated section.

Apart from the electricity use data, outside air temperature recordings were obtained from a near-by meteorological data sources and are part of the data set (case study I) or retrieved from [39] (case study II). No other heat gains are considered, in particular, solar gains are ignored which may lower the accuracy of the electricity consumption regression model. As we are mostly concerned with a qualitative view of the univariate energy signature, we ignore this for now.

Both data sets include residential buildings only, but we expect the method to be applicable to buildings with different primary space usage.

3.2. Preprocessing

We first preprocess the data by resampling hourly recordings to daily mean values. This averages out dynamic effects and daily

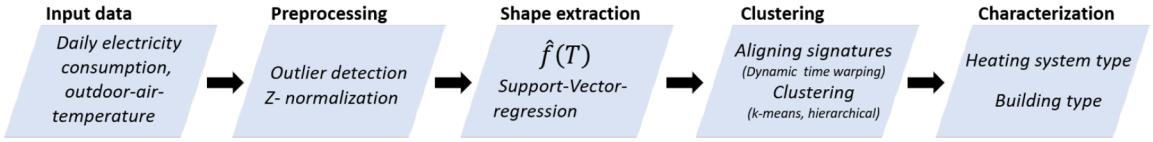


Fig. 3. Overview of the methodology.

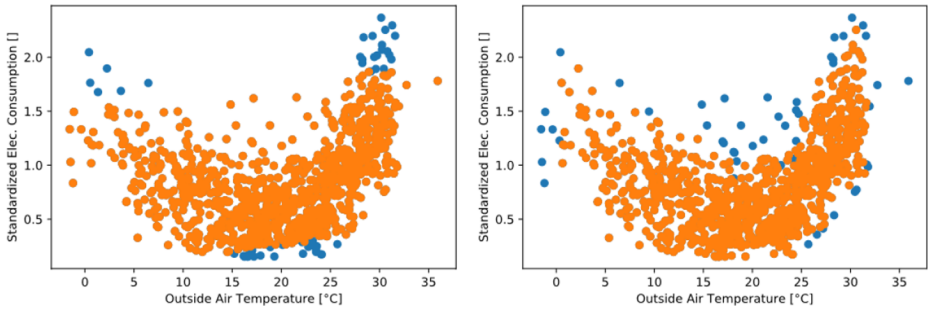


Fig. 4. Outlier removal for example *Building 2* from Fig. 2. **Left:** 5% outlier removal using the Mahalanobis distance. **Right:** 5% outlier removal using the local outlier factor score.

occupant related variation in the electricity consumption, and is commonly done for steady-state, OAT-based building energy models [20].[1] We exclude buildings with less than one year of recordings and delete individual days if zero electricity consumption was observed.

The latter may help to identify non-occupied periods in a building, but an absence detection algorithm was not applied (e.g. [40]).

3.2.1. Outlier filtering

The identification of outliers in an energy signature is a bivariate problem where both OAT and energy consumption are taking into account. Multiple different multivariate outlier detection methods exist. As the energy signature may be multi-modal and non-linear (e.g. Fig. 4), popular methods like Mahalanobis distance [41] or minimum covariance determination [42], which assume the data to originate from a multivariate Gaussian distribution, are not suitable.

Instead, we quantify the local outlier factor of each daily recording [43]. This involves the calculation of the density of neighbours around one sample point. It is called a local method as it compares the density only among the nearest neighbours. Samples with a density lower than a certain threshold are classified as outliers.

The method was applied using the ScikitLearn toolbox [44]. It involves the selection of neighbourhood-selection algorithms (here brute force), the distance metrics (here Euclidean distance), the size of the neighbourhood and the number of outliers to be deleted. The size of a neighbourhood was chosen to be 5 observations (i.e. 5 days) after conducting an exhaustive grid search between 3 and 25 days. The bottom 1% of the observations, sorted by density, are deleted for each building.

The process is compared to Mahalanobis based outlier detection in Fig. 4. The covariance-based method uses a multivariate Gaussian distribution (2-D, ellipse) and classifies the points most distant from the center as outliers. This cuts off high and low values of this bi-modal energy signature (Fig. 4, left), and samples with abnormally high electricity consumption between 10 °C and 25 °C are not tagged as outliers. When using local outlier factors, these two effects are avoided.

[1] In some of the buildings considered, autocorrelation in the regression model residuals could still be observed, possibly because we ignore solar irradiation and day of the week information. For now, this was not investigated further as we aim to understand the information encoded in the shape of a univariate energy signature.

3.2.2. Standardization

Buildings vary in size and number of occupants, so standardization is required to compare their energy signatures. Often, *intensity* metrics are used where building energy consumption is normalized by floor area. However, here the floor area is unknown, and also this can lead to skewed results as it does not control for varying numbers of occupants or incorrectly reported floor area (e.g. if conditioned space deviates from the overall floor area).

Instead, we use a purely statistical z-score standardization [45].

$$z_{r,n} = \frac{e_{r,n} - \mu_n}{\sigma_n} \tag{1}$$

This transforms each building's electricity consumption data to have zero mean and a standard deviation of 1, where $e_{r,n}$ is the r-th of R daily energy recordings of building $n \in \{1,...,N\}$ with N being the total number of all buildings. μ_n is the mean daily energy consumption, and σ_n the standard deviation. The same method was applied in [14,46] where standardization was required to cluster typical load profiles of buildings.

3.3. Non-linear model fitting

To enable clustering of signatures based on their shape, suitable features representing the shape of each signature are found.

We approach this by fitting a univariate regression model $\hat{f}_n(T_{OAT})$

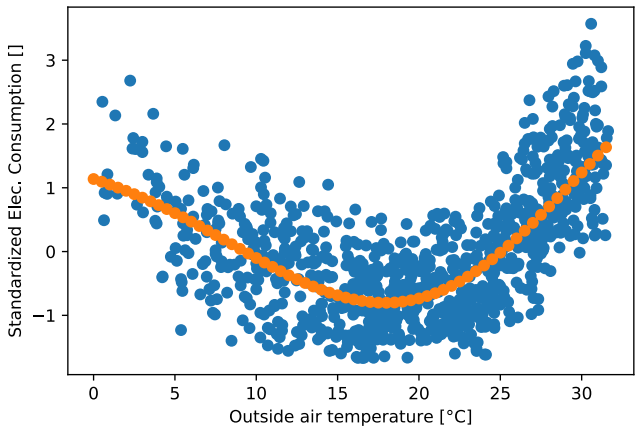


Fig. 5. Support-vector regression model of electricity consumption evaluated at 0.5 °C increments for *Building 2* of Fig. 2.

to the Energy Signature (ES) of each building n and evaluate that model at 0.5°C increments (see Fig. 5). The model evaluations are stored in an $Ix1$ vector $\overrightarrow{E_n}$, where I is the number of 0.5°C increments and depends on the maximum and minimum outside air temperature recording ($T_{OAT,max}$, $T_{OAT,min}$). Next, we convert two-dimensional time series data (outside-air-temperate, electricity consumption) with varying number of recordings (possibly missing data in each building) into one vector with the same number of entries for each building. We term this vector $\overrightarrow{E_n}$ the *profile* of the energy signature.

$$\overrightarrow{E_n} = [\hat{f}_n(T_{OAT,min}),\ ...,\hat{f}_n(T_{OAT,max})] \tag{2}$$

We use algorithmic modelling to find the non-linear univariate fit. Thus, we circumvent picking the number and location of breakpoints as required in the change-point regression approach [32].

We fit a support-vector regression (SVR) model with a polynomial kernel which takes around 2 s for each building on a common personal computer. We optimize the regularizer γ in a grid search for each building using 5-fold cross-validation.[2]

3.4. Clustering

The goal of clustering is to find C typical signature profiles among a large set of different profiles.

In the following section we introduce two similarity scores, the Euclidean distance and dynamic time warping paired with the Euclidean distance, and two clustering algorithms, k-means and hierarchical clustering, which were used in this study. Finally, we explain the Silhouette score which serves as metric to assess the quality of the chosen approaches.

3.4.1. Similarity of two energy signatures

We quantify the similarity of two signature profiles by comparing the electricity consumption at a certain temperature. In the simplest case we compute the squared Euclidean distance between two profiles of building n and p.

$$D = \sum_{i=1}^{I} (E_{n,i} - E_{p,i})^2 \tag{3}$$

However, this may be problematic. In Fig. 6), the general heating and cooling behaviour is the same in both buildings in both plots. They only differ in the balance point, which is the temperature at which either the heating or cooling system is switched on. However, the Euclidean distance for this case would be large. In the simplest case (electric heating or cooling only) a simple shift of the two curves would solve the problem (Fig. 6, left). However, if both buildings depict electric heating and cooling behaviour, a bi-directional shift will be required. A highly efficient building will have a lower heating balance point (the temperature where heating is switched on), and higher cooling balance point (the temperature where cooling is switched on, compared to a less insulated building. To account for this difference in balance points, the alignment of the two profiles needs a bidirectional warping of one signature to the other.

We approach this problem by warping the profiles in a non-linear fashion using the dynamic time warping (DTW) algorithm. The goal of DTW is to find the minimum distance of two sequences. Therefore, pairwise distance of all possible combinations of points in the sequence is computed. Those distances are stored in a *IxI* matrix. The shortest path, i.e. the path leading to the minimum summed distance, through that matrix is found using dynamic programming. The process is shown in Fig. 7 where the black lines indicate which profile vector entries are compared to each other. The length of the black lines tell us by how much the profiles are warped towards each other to achieve minimum distance. The process allows us to preserve the shape of the energy signature profiles (i.e. the entries of the profile vector remain), and at the same time it allows us to find energy signatures with similar shape even if they are offset.

All DTW-minimized pairwise distances between all N buildings are stored in a *NxN* matrix. This matrix is provided as input to the k-means and hierarchical clustering algorithms introduced below.

By finding the alignment of both profiles with minimum distance, two signatures with similar profile shapes are expected to be sorted to the same cluster. The particulars of the DTW algorithm leads to some effects to be considered when applied to energy signature comparison:

- Consumption at one temperature may be matched to the energy consumption at multiple temperatures of the other signature (see Fig. 7 top left).
- DTW is direction dependent (sequences sorted either by increasing or decreasing temperature values). This is due to its origin in time series analysis where it must be ensured that matching backwards in time is not feasible.
- The lowest temperatures and highest temperatures of two profiles are always aligned. Consequently, at the edges of the profiles the two profiles to not get shifted at all.

The above issues, do introduce some error when aligning the energy signature profiles. Especially, the alignment of the profile edges is problematic and leads to overestimating the distance among profiles (see 7 at 0 °C). As a consequence, we not only using apply dynamic time warping, but also compare it to common euclidean distance based clustering to show that the benefits of DTW outweigh drawbacks. In future, better alignment strategies can be developed.

3.5. Clustering algorithms

We use k-means and hierarchical agglomerative clustering to group similar energy signatures. We combine both methods with DTW and quantify the quality of the determined clusters using the Silhouette score, which is explained in a later section.

3.5.1. K-means

In k-means clustering the number of clusters C is predefined [47]. C centroids are randomly initiated and each profile is assigned to the nearest centroid (measured by the squared Euclidean distance). All profiles belonging to one centroid form a cluster. Afterwards, the centroid is computed as the mean of all profiles in the cluster, and cluster assignment is redone. The process is repeated until cluster assignment converges.

K-means is specifically designed for Euclidean-distance minimization and modification of the similarity metrics used here may prevent the algorithm from converging. For clustering based on the DTW-distance metric we use a slight variation, differentiable soft-DTW metric as k-means demands differentiability for optimisation. SDTW considers all possible pairwise-alignments weighted by their probability under the Gibbs distribution while DTW considers only the best one [48].

For k-means we used the SKlearn implementation [44] and for k-means using the SDTW similarity metric we used the TSLearn toolbox [49].

3.5.2. Hierarchical agglomerative clustering

Hierarchical agglomerative clustering is a bottom-up approach, where each sample is initialized as one cluster and clusters are merged pairwise as the process continues. In each iteration, two clusters are merged following a certain criterion [50]. We use the *ward* criterion which minimizes the total within-cluster variance quantified using the preferred similarity metric (see Section 3.4).

[2] Model parameters: SVR(C: 0.5, *cachesize*: 200, *coef*0: 0.0, *degree*: 3, *epsilon*: 0.1, *gamma*: *np*. *logspace*(−5, −2, 10), *kernel*: *rbf*, *max*$_i$*ter*: −1, *shrinking*: *True*, *tol*: 0.001).

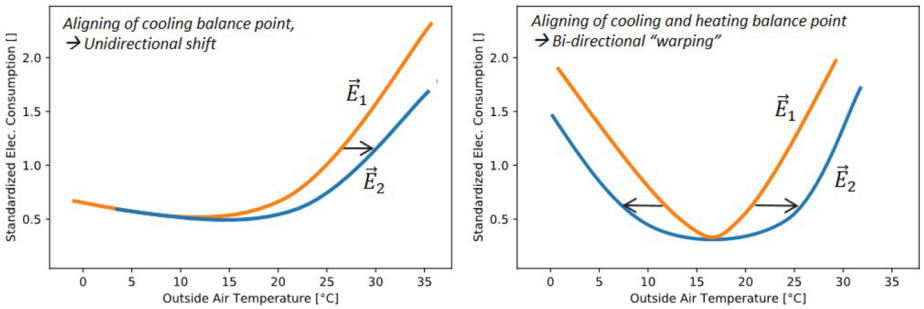


Fig. 6. Alignment of two energy signature profiles to quantify the similarity of their shape.

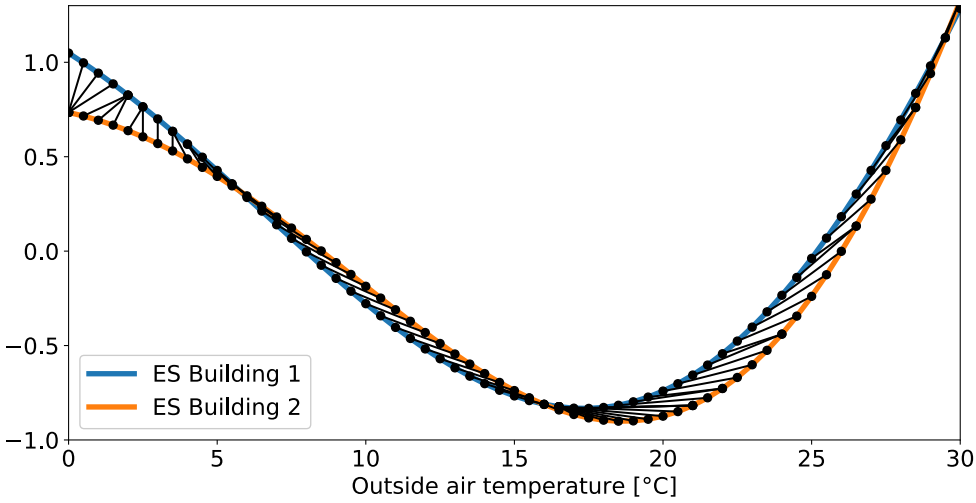


Fig. 7. Pairwise alignment of energy signatures using dynamic time warping. Note that the black lines indicate which points are aligned to each other. They do not indicate the distance score between those two points which is purely based on the vertical distance between two aligned points.

In comparison to k-means clustering, hierarchical clustering does not require the similarity metric to be differentiable. Hence we can use the DTW metric without the need to compute the differentiable SDTW score (see above).

3.6. Validation of clusters

We assess the quality of the resulting clusters using two metrics. For both metrics, we compute the sample Silhouette score $s(n)$ for each building n [51].

$$s(n) = \begin{cases} 1 - a(n)/b(n) & \text{if } a(n) < b(n) \\ 0 & \text{if } a(n) = b(n) \\ b(n)/a(n) - 1 & \text{if } a(n) > b(n) \end{cases}$$

where a(n) is the average similarity of one energy signature to the others within the same clusters and b(n) is the average similarity to all the signatures of the closest cluster. To quantify the overall performance of the clustering algorithm and similarity metric, we take the average of all samples $n \in \{1, ..., N\}$

$$S_{mean} = \frac{1}{N} \sum s(n) \tag{4}$$

The second metric aims to identify small clusters, which is required if the characteristics of buildings in a stock are imbalanced. A small cluster with a high average sample Silhouette score may be outweighed by a small decrease in Silhouette scores of all other energy signatures. To take this into account, we not only look at the overall average sample Silhouette score but also at the average cluster Silhouette score S_c of each cluster $c \in \{1, ..., C\}$.

$$S_c = \frac{1}{N_c} \sum s(n_c) \tag{5}$$

4. Results

In the following section we apply our method to two case studies. First, we look at a building stock in a climate with significant heating and cooling demand which is composed of multiple building types with varying heating systems. In the second case study, we investigate how energy signatures of only electrically heated single/duplex buildings can be clustered into different groups.

4.1. Case study I: heating and cooling climate

We analyse 3 years (2014–2017) of smart meter and outside air temperature recordings of residential customers in Austin, TX, from the Pecan Street data set [9]. In the period considered, temperatures range from $-8\,°C$ to $43\,°C$ with a mean of $20\,°C$ leading to significant heating and cooling demand.

The data encompasses 409 buildings including single-family houses (296), townhouses (17), and apartments (96). Apart from the smart meter data, 191 building owners took part in a survey providing information on the building system installed. Based on this survey, all buildings feature air-conditioning systems with 1–3 stages. Heating systems include gas furnaces (155), electric furnaces (16) and heat pumps (20).

As we can see the data set is imbalanced, with the bulk of the smart meter recordings coming from single-family houses, and most of the surveys conducted on buildings with gas furnaces. Therefore, we compare k-means clustering, which is known to produce clusters of

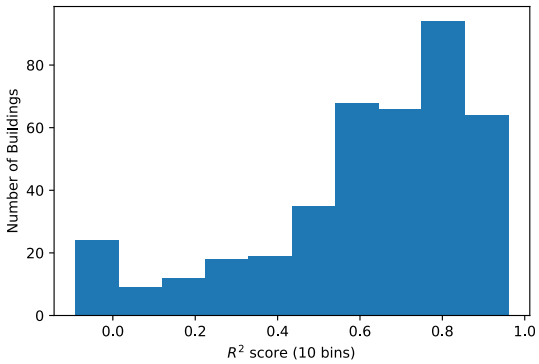


Fig. 8. Accuracy of univariate regression models of 409 buildings.

uniform size, with hierarchical clustering.

4.1.1. Fitting the univariate regression model

We fit 409 univariate SVR models with mean daily OAT as the input and electricity demand (z-standardized) as the output. In each training run, we pick the regularization hyperparameter γ in a grid search and use 5-fold cross validation. We test the accuracy of each model on a separate building-specific test set consisting of the electricity demand of randomly selected days.

The performance of each model on the separate test set is shown in Fig. 8. The average R^2 score is 0.61. Low R^2 values indicate that the majority of the variance in the electricity consumption cannot be explained by the outside temperature. We show below that this holds in particular for individual apartments which are less exposed to outside climate than free standing buildings.

After model-fitting, we evaluate all models in 0.5 °C increments between 1 and 32.5 °C. The limits correspond to the 1- and 99-percentiles of the temperature recordings.

4.1.2. Choosing the similarity metric and clustering algorithm

We conduct hierarchical agglomerative and k-means clustering with and without DTW (see Section 3.4). The performance of these four clustering approaches is shown in Fig. 9. The mean Silhouette score is shown for varying number of clusters. While hierarchical agglomerative clustering is deterministic and terminates with the same clusters after each run, k-means clustering depends on the random initialization of the cluster centroids before cluster fitting. Therefore, ten k-means repetitions were run and the best result was stored.

K-means clustering outperforms hierarchical clustering for low numbers of clusters. For both clustering algorithms we find that the use of DTW leads to an increase in the Silhouette scores and supports our idea to align energy signatures (Fig. 6). In this case study, DTW performed particularly well with k-means clustering. However, in the second case study both clustering algorithms produced higher Silhouette scores when DTW was used (see Fig. 14).

4.1.3. Selection of the number of clusters

Both the k-means algorithm and hierarchical clustering require us to choose the number of clusters to search for. Our choice is based on the average Silhouette scores (Eq. (4)) we compute after each run. The higher the score the more unambiguously each energy signature is assigned to one cluster. [51] suggests average Silhouette scores of larger than 0.5 are strong evidence for an underlying structure in the data, and average scores below 0.25 indicate that no structure is found.

Fig. 9 shows that k-means with DTW provides an average Silhouette scores larger than 0.5 for 4 clusters or less. As the goal of this analysis is to find as many different energy signature types possible, we proceed with four clusters. Another interesting aspect is to look at the average Silhouette scores of each individual cluster. Some clusters perform better than others. In the case of $n_{clusters} = 4$ we find the scores to vary between $S_{C1} = 0.41$ and $S_{C2} = 0.68$ (see Table 2) indicating that especially cluster 3 is very distinct.

4.1.4. The four energy signature clusters

The four clusters are displayed in Fig. 10. The grouping provides a rapid overview of the building stock. The bulk (62%) of all buildings have profiles similar to cluster 3, 18% similar to cluster 2, and 12% and 7% are similar to cluster 1 and 4. On the left of Fig. 10, we display the sample Silhouette score for each building. It is skewed with less than 20% of the energy signatures with a score smaller than 0.25 (see Section 4.1.3). Those signatures do not align well with any cluster, but rather lie in between two or more clusters. This highlights the advantage of a sample-specific quality score which allows us to identify buildings whose cluster assignment is uncertain. They are filtered out as described below.

Description of shapes. The different cluster shapes depict varying sensitivity towards changes in OAT at low and high temperatures.

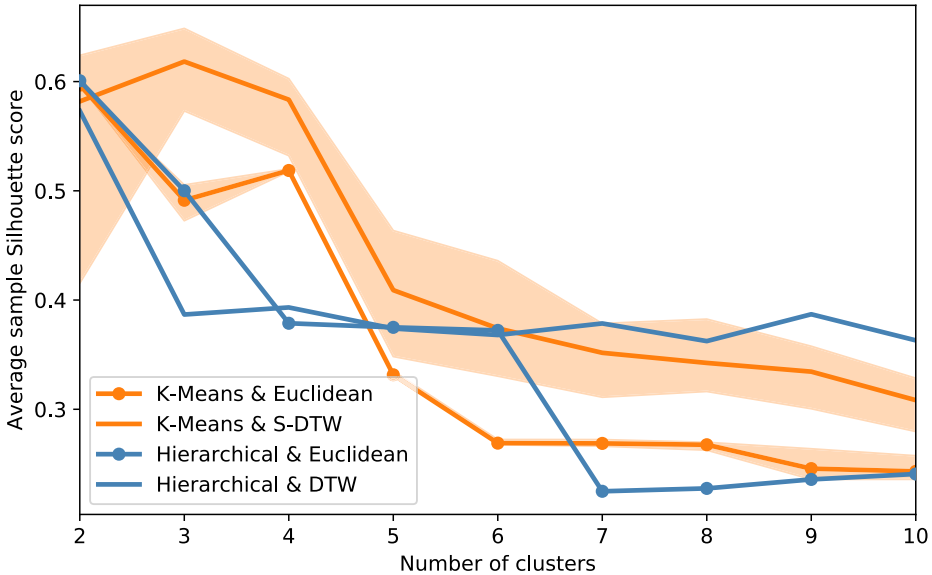


Fig. 9. Average sample Silhouette score for different clustering options. In case of k-means clustering, the performance band for ten different centroid initialization is shown. Agglomerative hierarchical clustering does not depend on the initialization and hence each run terminates with the same results.

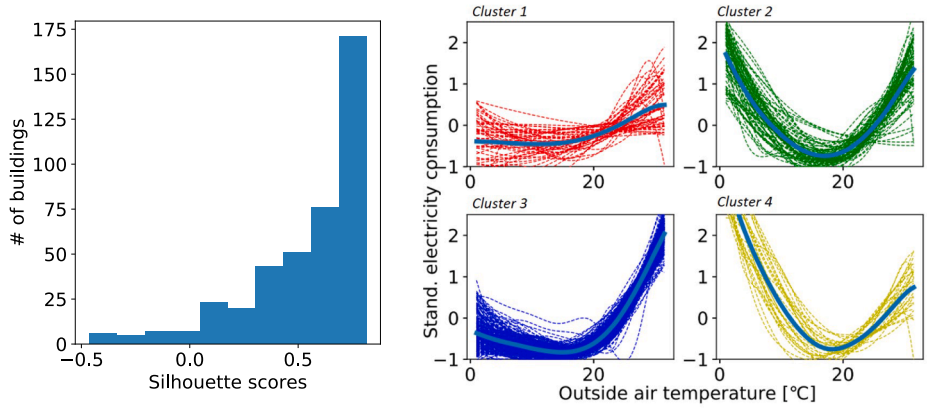


Fig. 10. The four determined energy signature clusters (right) and the distribution of sample Silhouette scores (left).

Table 1
Overview of the building stock data of the two case studies.

	Location	# of buildings	Building types	Heating system	Cooling system
Case study I [9]	Austin, Texas	409* (3-y, hourly)	Single/duplex, apartment, townhouse	Gas furnace, electric furnace, heat pump	(Mini-) split, heat pump
Case study II (undisclosed)	Vancouver Island, British Columbia	480 (2-y, hourly)	Single/duplex	Electric (type unknown)	Electric (type unknown)

* Metadata on heating/cooling systems of 191 buildings is available.

Table 2
Summary of average cluster Silhouette scores for different numbers of clusters. The results for the clustering algorithm providing the highest mean Silhouette score are shown.

$n_{Clusters}$	$C1$	$C2$	$C3$	$C4$	$C5$	$C6$	$C7$	$C8$	$C9$	$C10$
2	0.65	0.55								
3	0.71	0.51	0.51							
4	0.49	0.41	0.68	0.43						
5	0.52	0.33	0.40	0.36	0.47					
6	0.56	0.20	0.39	0.34	0.52	0.22				
10	0.22	0.45	0.64	0.39	0.35	0.27	0.13	0.14	0.42	0.18

In cluster 1 the sensitivity is small in comparison to the other clusters. The cluster centroid varies only between −1 and 1.5 whereas in the other clusters it reaches much higher values at extreme temperatures[3] This observation is also quantified by the mean R-squared score $R^2_{mean,1} = 0.19$, which indicates that on average only 19% of the variance in the consumption data of buildings in cluster 1 is explained by an SVR model with OAT as the exogenous variable. In all other clusters, a larger fraction of variance is explained ($R^2_{mean,2} = 0.49$, $R^2_{mean,3} = 0.73$ and $R^2_{mean,4} = 0.61$).

In comparison to clusters 2 and 4, cluster 3 shows a small temperature correlation at *low* temperatures leading to an inverted L-shape. The shape of profiles in cluster 2 is more symmetric with similar electricity demand at low and high OAT. Cluster 4 involves signatures with very high heating demand outweighing the maximum cooling demand significantly. (See Table 1).

Another way to differentiate the clusters is their mean cluster Silhouette score (see Table 2). Based on that score, Cluster 3 has the highest quality (0.68), while the other clusters reach lower scores between 0.4 and 0.5 (see Section 3.4).

Interpretation. We compare the clusters to available metadata (Figs. 11, 12)) to validate that the clusters indicate certain building

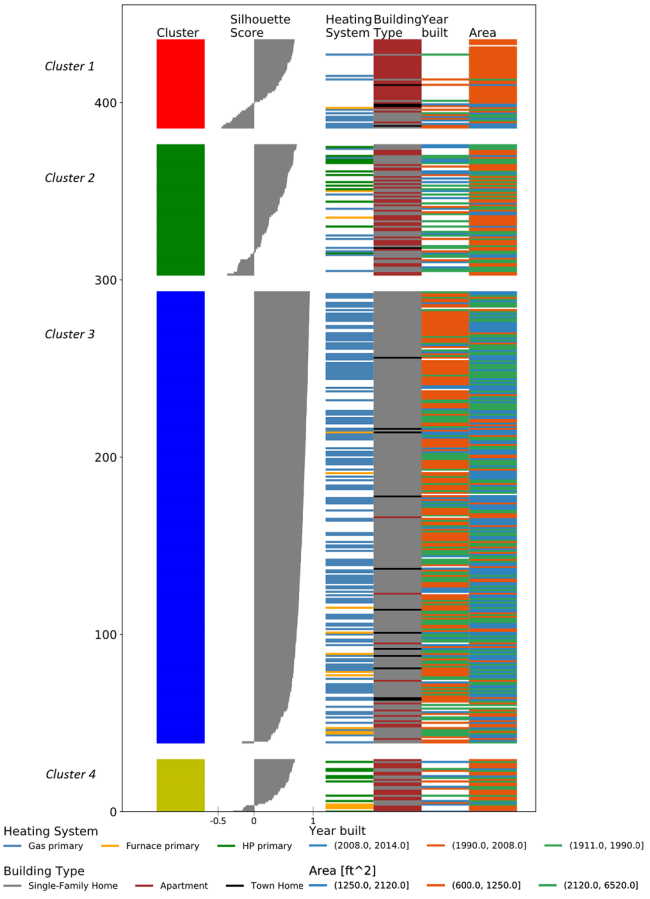


Fig. 11. Comparison of cluster to associated building properties (heating system type, building type, building age and footprint) sorted by the sample Silhouette score (graph inspired by [14]).

[3] Note that the *z-standardization* transforms the data to have zero mean and a standard-deviation of 1.

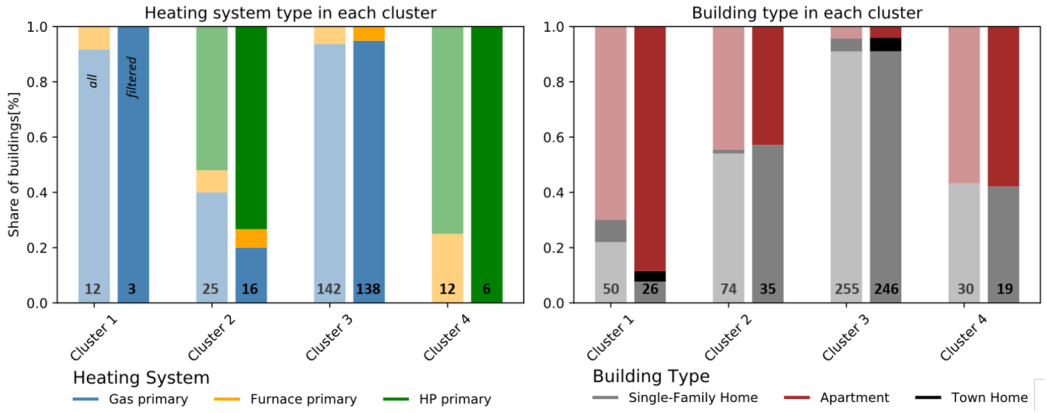


Fig. 12. Comparison of clusters to heating system type and building type. The left bar shows all buildings where metadata is available (number provided at the bottom of the bar), and the right bar after filtering out buildings with low sample Silhouette score.

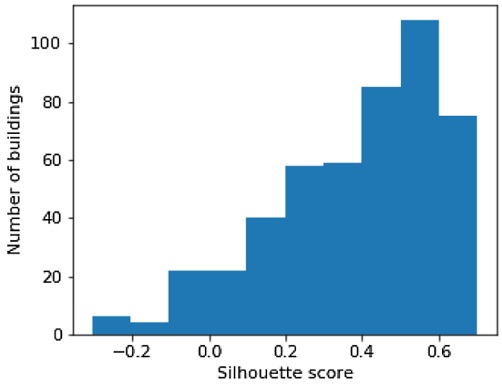


Fig. 13. Silhouette scores of the buildings in the second case study.

characteristics. The metadata were partly collected when installing the metering equipment (building type, available on all buildings), and in a separate survey which provides information on the installed heating system (191 buildings). As mentioned above, we filtered out buildings with Silhouette scores lower than 0.25. This lead to a higher correlation of the clusters with certain building characteristics. For example, after filtering cluster 1 consists of more than 90% apartment buildings (Fig. 12, right bar) instead of 70% without filtering.

When comparing the metadata to the clusters, we find that building characteristics agree with the physical interpretation of the energy signature shapes.

- Cluster 1 has a very flat shape indicating low temperature sensitivity. Looking at the metadata, it is dominated by apartment buildings (90%, Fig. 12, right). This may be because apartments have fewer external walls and hence less temperature-dependent loads. Besides that, heat supply may be metered separately from the electricity meter. For buildings in cluster 1, only a few surveys covering the heating system (Fig. 12, left)) are found. This prohibits a final conclusion on the prevailing heating system.
- Clusters 2 and 4 are U-shaped having a strong temperature dependence. Consequently, the clusters primarily consist of buildings with an electric heating systems installed. After filtering, cluster 4 is composed of buildings with heat pumps only and in cluster 2 more than 80% of the buildings generate heat either with an electric furnace or heat pump.
  However, prior to filtering a significant fraction of the buildings in cluster 2 (≈ 40%) are equipped with gas furnaces. The explanation may be that most of the buildings in the dataset have a ducted central heating system [9]. Logically, the buildings may have significant electric demand from fans to circulate the air through the building. Furthermore, some auxiliary heating or cooling systems not included in the survey may be present. The latter observation shows that buildings with heterogeneous heat sources may be misclassified by our method. Nonetheless, using the sample Silhouette score those uncertain cases will be identified and filtered out.
- Lastly, the L-shaped cluster 3 captures buildings with low electricity consumption for heating and high electricity consumption for cooling. The low heating demand shows that at low temperatures heating demand is covered by non-electric energy sources; the high cooling rate underlines that those buildings have high heat gains in summer adding a significant cooling load relative to the base load of the buildings. The comparison to the metadata is in line with the physical interpretation of the shape. More than 95% have non-electric gas furnaces installed and more than 90% of the buildings are free standing buildings (single/duplex) with possibly higher heat gains than non-free standing buildings. Unless there are high solar gains, apartments or townhouses are better insulated by surrounding structures.

Another interesting aspect is that, although we found two clusters dominated by electric heating systems (clusters 2 and 4), they do not separate buildings with a heat pump from those with an electric furnace. When looking at Table 3, the ratio of electric cooling and heating efficiency is much higher for resistance heaters than for more efficient heat pumps. We would expect that the energy signature profiles of buildings with resistance heaters would have much higher heating compared to cooling demand (i.e. shaped like a hockey stick, similar to cluster 4), than buildings with a heat pump, where the profile should be rather ”U”-shaped with rather similar levels of heating and cooling demand (similar to cluster 2). This could not be confirmed in this case study in particular due to a lack of metadata (only 16 buildings with

Table 3
Ratio of $COP_{cooling}$ and $COP_{heating}$ can serve as a proxy for how much heating demand surpasses cooling demand per change in OAT. This can have direct implications for the shape of the signature. For example, in a system with an electric resistance heater and conventional AC unit, one would expect a non-symmetric shape of the ES with heating demand increasing 4.1–5.9 times faster than cooling demand. **Note:** The COP value ranges were calculated based on *SEER* and *HSPF* ranges found in [52]).

Ratio $\frac{COP_{cooling}}{COP_{heating}}$	η, $COP_{heating}$	AC	Central AC (Split)	Heat pump (cooling)
$COP_{cooling}$		4.1–5.9	4.1–7.6	4.1–8.9
Fuel based furnace	n/a			
Elec. resistance heater	1	4.1–5.9	4.1–7.6	4.1–8.9
Heat pump (heating)	2.4–4.0	1.0–3.1	1.0–2.4	1.0–3.7

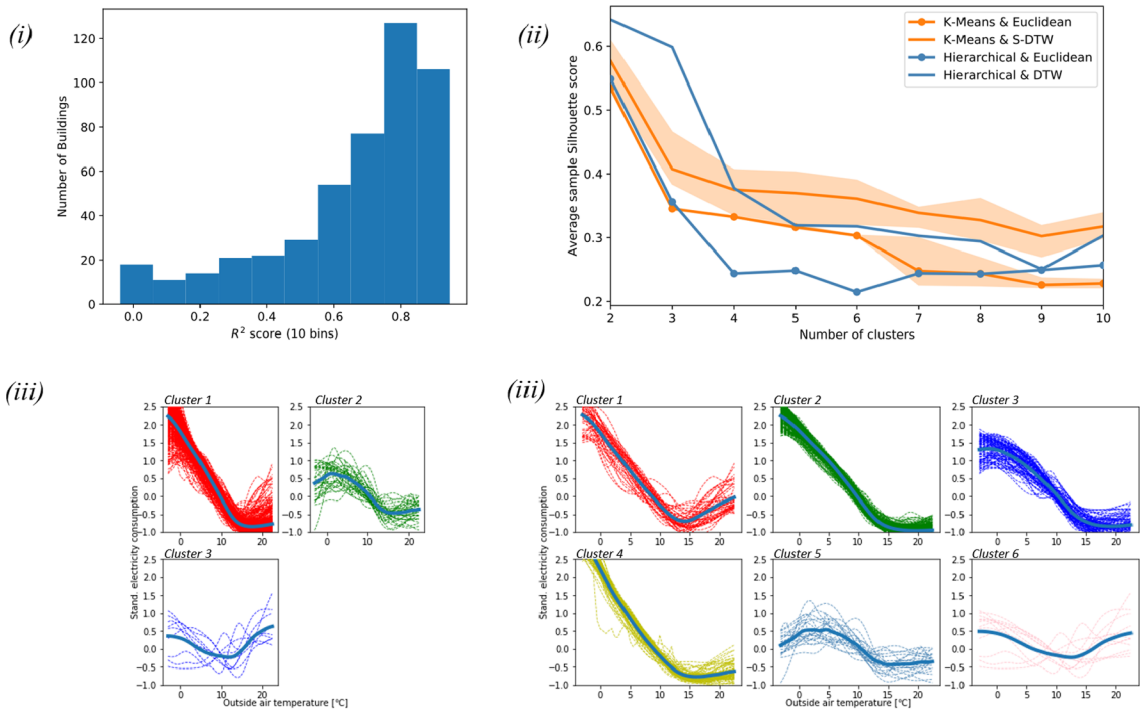


Fig. 14. Summary case study II. *(i)* shows the R^2 score of the univariate SVR regression models. *(ii)* shows the S_{mean} decay for increasing number of clusters. Three clusters lead to a high mean $S_{mean} > 0.5$ and are depicted in *(iii)*. After that the score remains at a level of $S_{mean} \approx 0.4$ with a drop after more than 6 clusters. The determined six clusters are shown in *(iiii)*.

electric furnaces were present in the survey).

We conclude that the first case study has shown, that clustering energy signatures allows us to distinguish electrically heated from non-electrically heated buildings, and to group buildings of similar building type (cluster 1, 3). Although easy to achieve, the former finding is essential to design decarbonization strategies for a building stock. Given the scarcity in the data on electrically heated buildings in this case study and motivated by the physical differences between resistance heaters from heat pumps (Table 3), the following section investigates how clustering may be used to group buildings with different electric heating systems.

4.2. Case Study II: Heating dominated climate, electrically heated single-family buildings only

Case study II aims to better understand the performance of energy signature clustering in identifying the type of *electric* heating system installed. Therefore, we use the algorithm on data from only electrically heated single family or duplex buildings (480 buildings, 2 years). The data consists of daily undisclosed smart meter recordings from Vancouver Island, British Columbia, whose climate is heating dominated. Here no building-level metadata on the heating systems is available, but highlevel demographic data on heating system coverage in the single/duplex building sector of British Columbia was compared to the clusters[53].

This case study repeated the same methodological steps, which are summarized in Fig. 14.

The SVR-fit to the data leads to slightly higher performance ($R^2_{mean} = 0.67$) than in case study 1 ($R^2_{mean} = 0.61$) which can probably be traced back to the lack of apartment buildings in this case study. While before we found a jump in the histogram around $R^2 \approx 0$, here the histogram has a continuous decay towards low R^2 scores.

Clustering energy signatures from electrically heated buildings is a harder problem than separating electrically heated from non-electrically heated buildings and consequently the average Silhouette score is lower (see Figs. 13, 14 *(ii)*). Three is the maximum achievable number of clusters with an average Silhouette score $S_{mean} > 0.5$. Afterwards, S_{mean} drops to a level of 0.4 for up to 6 clusters. Here, the distribution of Silhouette scores is less skewed as in the first case study with approximately 25% of all buildings having a score of less than 0.25. Similar to before, filtering is applied to eliminate these cases.

Cluster shapes Grouping the energy signatures into three clusters enables us to separate buildings with a strong linearly increasing heating demand (cluster 1) from those buildings whose demand levels-off or even decreases at low temperatures (cluster 2). Furthermore, one cluster which groups buildings not well represented by the other two clusters is found (cluster 3).

Cluster 1 is by far the largest of the three clusters (436 of 480 buildings). The two tails of the cluster, i.e. for OAT <3 °C and OAT >15 °C, show that a large band of varying energy signature profiles is included in that cluster. At cold temperatures the slope of the profiles ranges from strongly negative to positive, and at warm temperatures from zero to positive. To better group buildings from this large cluster, we increase the number of clusters until the next slight drop of S_{mean} can be observed (6 clusters, see 14 *(iiii)*).

Having 6 clusters leads to a mean Silhouette score of $S_{mean} = 0.4$, which is lower than the recommended 0.5 (see Fig. 14*(ii)*), and on first glance, leads to some clusters of rather similar shape (Fig. 14*(iiii)*). However, each cluster features some physically meaningful differences compared to the other clusters. In cluster 1, all buildings exhibit some kind of cooling behaviour. Cluster 2, 4 are rather similar with no cooling and a linear increase in heating demand. However, in cluster 4 we see the tendency of heating demand increasing at a higher rate at low temperatures. Cluster 3 differs strongly from the other three clusters, as the heating demand levels-off below 5 °C. Although cluster 5 features a similar shape, the overall temperature caused variance is a lot smaller. Cluster 6 groups buildings with very low temperature dependence.

Interpretation. We find that the two tails of the energy signatures allow to separate the buildings from each other. The tails level-off (convex shape), increase with a uniform slope, or increase at higher orders.

Table 4
Comparison of determined clusters to demographic data from British Columbia [53].

	Clusters with matching shape	Share of buildings [%] (# buildings)		
		Clustering		Demographic data [53]
		all	*filtered*	
Heat pumps (air-source)	Cluster 4	16.4 (77)	12.1 (54)	10.4 (44'900)
Resistance heaters	Clusters 1, 2	60.4 (283)	63.5 (233)	64.4 (276'600)
Electric heater with auxiliary system	Clusters 3, 5	23.2 (109)	24.4 (65)	25.2 (107'700)
	Sum	100	100	100

We interpret the clusters in the following:

- With positive slopes at warm temperatures, the shape of cluster 1 (all: 47/ filtered: 20 buildings) may resemble the power consumption of buildings with an electric cooling system installed. The constant slope at low temperatures implies the use of a resistance heater as the efficiency, i.e. slope, is not temperature dependent.
- Cluster 2 (236/213 buildings) may capture buildings without cooling system and a linearly increasing heating demand which again supports the use of a resistance heater.
- Cluster 3 (87/46 buildings) and cluster 5 (22/19 buildings) have a strongly convex signature profile. This shows that some other non-electrical heating source must be available in the building leading to a stabilizing or even decrease of electricity consumption at low OAT.
- Cluster 4 (77/54 buildings) is very similar to cluster 2. However, it includes some buildings with cooling behaviour and the slope of the heating demand is continuously increasing at low OAT. This indicates a decrease in heating efficiency at lower temperatures. This resembles the demand profile of an air-source heat pump whose efficiency decreases at cold temperatures [36].
- Lastly, cluster 6 shows no clear trend in the shape and the included signatures vary strongly. It appears to group those buildings which do not fit any other cluster well. Due the low number of buildings (11) and variation in the energy signature shapes, it is not further discussed.

Finally, we compare the fraction of buildings in each cluster to demographic data for British Columbia (Table 4). The share of buildings in the clusters matches the demographic data well. Furthermore, filtering improves these results. Having a sample size of 369 buildings for a total population size of 429.200 electrically heated buildings in BC, we reach a confidence interval of 5.1% [54]. This comparison relies on the assumption that the set of buildings is representative of the residential building stock in BC, and on the physical interpretations above.

5. Discussion and outlook

This paper provides a method to gain rapid insight into qualitative characteristics (heating system type and building type) of a building stock. Energy signature profiles were extracted for each building and subsequently clustered to find dominant energy signature profiles.

Using two case studies we show that the method allows separating electrically heated from non-electrically heated buildings, identifying groups consisting of one building type (case study I), and differentiating buildings with different electric heating systems (heat pump, resistance heater, electric with auxiliary heater; case study II). The latter could not be validated with metadata available for each building, but we found strong agreement between our clusters and high level demographic data.

5.1. Outlook

The method successfully grouped buildings by certain building characteristics, and in particular we could show that aligning signature profiles with dynamic time warping is a promising way to find similar energy signatures in building stocks. In future research the following five issues could be addressed to improve the method and its robustness.

1. More metadata on system types (heating, cooling, ventilation) of buildings would be highly valuable. It would allows us to reconfirm our findings and could also show how energy signatures capture certain ventilation systems, which was not considered in this paper. Apart from that, additional data would enable supervised models to learn and predict characteristics based on energy signatures.
2. As we show, DTW is highly valuable to quantify the pairwise similarity of energy signatures. In both case studies it produced higher Silhouette scores than common Euclidean distance based clustering. However, as shown in the second case study, the shape of the energy signature at its tails, i.e. at extreme temperatures is specifically valuable to differentiate buildings. Hence, a similarity metric which better quantifies differences in energy signature tails could improve the building characterization.
3. In this study, we used k-means clustering and hierarchical clustering. As we found in the first case study the occurrences of HVAC systems in building stocks may be highly imbalanced requiring algorithms to find clusters of different size. First experiments with density based clustering did not show any improvements but further research on the most suitable clustering algorithms may improve the cluster quality.
4. Cluster quality was quantified by the average Silhouette coefficient. It measures how dense and distinct each cluster is, and proved that significant clusters were found. However, between 20% (*case study I*) and 25% (*case study II*) of the buildings had low sample Silhouette scores (<0.25). This indicates that they lie in between clusters and implies that the range of energy signature profiles in a stock is continuous rather than discrete, with certain energy signature shapes reoccurring more often than others. The Silhouette coefficient is helpful, to identify those buildings with little resemblance to larger groups of buildings. More research could help determine if the characteristics of those buildings can be inferred anyway as for example by using information from neighbouring clusters.
5. Case study I showed that some clusters were composed of buildings with different heating systems (e.g. cluster 2). This makes sence, since the energy signature is a composition of all loads inside a building with varying temperature dependency (including fan loads, hot water consumption, auxiliary plug heaters, other loads). This composition of the load was not analysed further. When high frequency smart meter data is available, a combination of our approach with load disaggregation could be analysed in future [19].

5.2. Application realm

This work was motivated by the need for automated building characterization to enable large scale, automated retrofit analysis. This may include its use for building model calibration, for building benchmarking (see Fig. 15) and for gaining rapid insight into building stocks for sustainable policy design. Lastly, researchers may leverage energy signature clustering to identify similar types of buildings to test

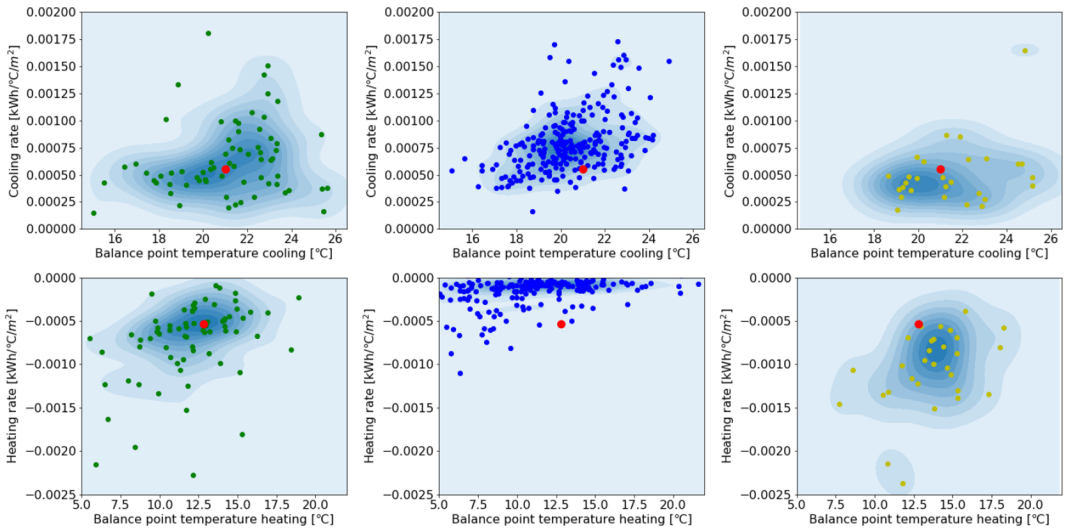


Fig. 15. Example of the use of energy signature shape clustering for building benchmarking. The thermal performance of one example building (red dot) is compared to the performance of buildings in clusters 2, 3 and 4 (left to right). The thermal performance is quantified by the temperature change point (x-axis) and the heating and cooling rate (y-axis).

new building analytics and control methods.

5.2.1. Model calibration

A comprehensive set of building characteristics allows us to use more suitable building models for parameter calibration processes [3,12]. Based on our results the clusters in *case study II* for example help us to understand that the calibration model must be designed such that it incorporates a non-electrical auxiliary heating system for extremely low temperatures. If not considered, the calibrated parameter values, for example the heating system efficiency, could be misleading. An example for that problem is given in [24], who found that either heat supply parameters (ventilation) or envelope parameters have to be known for accurate building performance estimates.

5.2.2. Building benchmarking

By clustering buildings into groups one can ensure that the performance of a building is benchmarked against buildings of similar type and heating system. We showcase that application in Fig. 15. In the plot, we derived the heating and cooling rate for each building of Clusters 2, 3 and 4 of *case study I* using a 5-point piecewise linear regression model [29]. We computed the average heating and cooling rate (red dot). When a building with similar performance would be benchmarked against the whole set of buildings it would have mediocre performance. This is also the case if benchmarked against buildings in cluster 2. However, if compared against buildings in cluster 3 it would have a relatively high heating rate (i.e. low efficiency) and if compared against buildings in cluster 4 the rate would be relatively low (i.e. high efficiency). This brief example shows that the addition of the cluster information could potentially enable to benchmark buildings in a more physically meaningful manner.

5.2.3. Demographic overview

The method provides a rapid overview of prevailing energy signatures in a stock, which can be used to infer the demographic distribution of heating system type or building type. For example, in *case study I* we immediately see the strong dominance of non-electrically heated buildings. Together with socio-economic insights from time-of-use analysis (see Fig. 1), customized policies could be developed based on smart meter data alone.

6. Conclusions

Knowledge on different building characteristics is required for the design of building energy retrofits [7]. In this study, we developed a novel, smart meter-based method to automatically retrieve thermal building characteristics, i.e. the heating system type and building type. It augments the set of methods to extract qualitative building characteristics using smart meter data only, which are essential to enable large scale, accurate building retrofit analysis. For example, it can improve building calibration, where the calibrated parameters are more accurate when the number of unknowns of a building is reduced [24].

While many previous studies extracted temporal patterns from smart meter data to obtain socio-economic characteristics (e.g. primary building use [14]), we developed a method for finding physical building characteristics. We used the concept of energy signatures which is an informative plot of the daily mean electricity usage and the daily mean outside air temperature. In our approach we extracted the shape of each buildings energy signature within a stock and subsequently, clustered them. That allows to sort buildings by heating system type and building type. This was validated in two case studies with the following quantitative results:

- In case study I, four clusters of different energy signature shapes were found. Comparing them to metadata showed that, each cluster is composed of more than 75% of buildings having the same heating system (heat pump, gas furnace). Furthermore, two of the clusters consist of more than 90% of the buildings with the same building type (apartments, single/duplex buildings). No cluster which strongly correlates to electric resistance heaters was found.
- In case study II, we focussed on the problem to differentiate buildings with electric heating systems only (heat pump, resistance heaters). The determined clusters allowed to split buildings equipped with heat pumps, from those with electric resistance heater, or non-electric auxiliary systems installed. Although no building-level validation data was available, we could show that the number of buildings in the clusters matches demographic data available with a maximum deviation of 1.7% in the share of buildings equipped with a certain heating system.
- Lastly, the approach allows to quantify the confidence of a building to be assigned to a certain cluster. We could show that the majority of all buildings has a sample silhouette score of larger than 0.25 (82% in case study I, 75% in case study II), a score lower than

indicates a low confidence in the cluster assignment.

For further research, we suggest to integrate the method into a large-scale building stock retrofit analysis. This allows to quantify, how much an automated building characterization can help to increase the accuracy of approaches like building calibration and building benchmarking.

Apart from that, we foresee this study as a start for more research leveraging energy signatures to retrieve building characteristics. More smart meter data sets with more metadata is required. This could potentially also guide to a supervised learning approach.

6.1. Code and data availability

The source code and analysis process for this work is available as a Python module and Jupyter notebooks. They are posted on GitLab and are hosted on the building analytics platform BESOS.[4,5] The data set from the first case study is available on [9]. The data from the second case study is unfortunately proprietary.

CRediT authorship contribution statement

Paul Westermann: Methodology, Software. **Chirag Deb:** Methodology, Conceptualization, Supervision. **Arno Schlueter:** Supervision, Resources, Validation, Writing - review & editing. **Ralph Evins:** Supervision, Resources, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The project benefited from data provided by [9] and BC Hydro. It was supported by grant funding from CANARIE via the BESOS project (CANARIE RS-327).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.apenergy.2020.114715.

References

[1] Edenhofer O. Climate change 2014: mitigation of climate change Vol. 3. Cambridge University Press; 2015.
[2] International Energy Agency, Tracking clean energy progress: 2017; 2017.
[3] Heo Y, Choudhary R, Augenbroe G. Calibration of building energy models for retrofit analysis under uncertainty. Energy Build 2012;47:550–60.
[4] Sun K, Hong T, Taylor-Lange SC, Piette MA. A pattern-based automated approach to building energy model calibration. Appl Energy 2016;165:214–24.
[5] Hong T, Yang L, Hill D, Feng W. Data and analytics to inform energy retrofit of high performance buildings. Appl Energy 2014;126:90–106.
[6] Chaudhary G, New J, Sanyal J, Im P, O'Neill Z, Garg V. Evaluation of "autotune calibration against manual calibration of building energy models. Appl Energy 2016;182:115–34.
[7] Miller C, Nagy Z, Schlueter A. A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. Renew Sustain Energy Rev 2018;81:1365–77.
[8] Nordström G. Use of energy-signature method to estimate energy performance in single-family buildings, Ph.D. thesis, Luleå tekniska universitet; 2014.
[9] Pecan Steet, Dataport: the world's largest energy data resource, Pecan Street Inc. <https://www.pecanstreet.org/>.
[10] Scully P. Smart meter market report, Tech. rep., IOT Analytics; 2019.
[11] Wang Y, Chen Q, Hong T, Kang C. Review of smart meter data analytics: applications, methodologies, and challenges. IEEE Trans Smart Grid 2018;10(3):3125–48.
[12] Nagy Z, Rossi D, Hersberger C, Irigoyen SD, Miller C, Schlueter A. Balancing envelope and heating system parameters for zero emissions retrofit using building sensor data. Appl Energy 2014;131:56–66.
[13] Miller C, Meggers F. Mining electrical meter data to predict principal building use, performance class, and operations strategy for hundreds of non-residential buildings. Energy Build 2017;156:360–73.
[14] Park JY, Yang X, Miller C, Arjunan P, Nagy Z. Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset. Appl Energy 2019;236:1280–95.
[15] Miller C. What's in the box?! towards explainable machine learning applied to non-residential building smart meter classification, Energy and Buildings.
[16] Tong X, Li R, Li F, Kang C. Cross-domain feature selection and coding for household energy behavior. Energy 2016;107:9–16.
[17] McLoughlin F, Duffy A, Conlon M. Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: an Irish case study. Energy Build 2012;48:240–8.
[18] Wang Y, Chen Q, Gan D, Yang J, Kirschen DS, Kang C. Deep learning-based socio-demographic information identification from smart meter data. IEEE Trans Smart Grid 2018;10(3):2593–602.
[19] Deb C, Frei M, Hofer J, Schlueter A. Automated load disaggregation for residences with electrical resistance heating. Energy Build 2019;182:61–74.
[20] Chambers JD. Developing a rapid, scalable method of thermal characterisation for uk dwellings using smart meter data Ph.D. thesis UCL (University College London); 2017.
[21] Coakley D, Raftery P, Keane M. A review of methods to match building energy simulation models to measured data. Renew Sustain Energy Rev 2014;37:123–41.
[22] Bacher P, Madsen H. Identifying suitable models for the heat dynamics of buildings. Energy Build 2011;43(7):1511–22.
[23] Bacher P, Andersen P. IEA Common Exercise 4: ARX, ARMAX and grey-box models for thermal performance characterization of the test box. Technical University of Denmark; 2014.
[24] Nagpal S, Mueller C, Aijazi A, Reinhart CF. A methodology for auto-calibrating urban building energy models using surrogate modeling techniques. J Build Perform Simul 2019;12(1):1–16.
[25] Borgeson SD. Targeted efficiency: Using customer meter data to improve efficiency program outcomes, Ph.D. thesis, UC Berkeley; 2013.
[26] Rabl A. Parameter estimation in buildings: methods for dynamic analysis of measured energy use. J Solar Energy Eng 1988;110(1):52–66.
[27] Fels MF. Prism: an introduction. Energy Build 1986;9(1–2):5–18.
[28] Perez KX, Cetin K, Baldea M, Edgar TF. Development and analysis of residential change-point models from smart meter data. Energy Build 2017;139:351–9.
[29] Kissock JK, Haberl JS, Claridge DE. Inverse modeling toolkit: numerical algorithms. ASHRAE Trans 2003;109:425.
[30] Danov S, Carbonell J, Cipriano J, Martí-Herrero J. Approaches to evaluate building energy performance from daily consumption data considering dynamic and solar gain effects. Energy Build 2013;57:110–8.
[31] Burkhart MC, Heo Y, Zavala VM. Measurement and verification of building systems under uncertain data: a gaussian process modeling approach. Energy Build 2014;75:189–98.
[32] Paulus MT, Claridge DE, Culp C. Algorithm for automating the selection of a temperature dependent change point model. Energy Build 2015;87:95–104.
[33] Baasch G, Wicikowski A, Faure G, Evins R. Comparing gray box methods to derive building properties from smart thermostat data. Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation. 2019. p. 223–32.
[34] Belussi L, Danza L. Method for the prediction of malfunctions of buildings through real energy consumption analysis: Holistic and multidisciplinary approach of energy signature. Energy Build 2012;55:715–20.
[35] Lundström L, Wallin F. Heat demand profiles of energy conservation measures in buildings and their impact on a district heating system. Appl Energy 2016;161:290–9.
[36] Love J, Smith AZ, Watson S, Oikonomou E, Summerfield A, Gleeson C, et al. The addition of heat pump electricity load profiles to gb electricity demand: Evidence from a heat pump field trial. Appl Energy 2017;204:332–42.
[37] Wei G, Liu M, Claridge DE. Signatures of heating and cooling energy consumption for typical ahus. In: Proceedings of the eleventh symposium on improving building systems in hot and humid climates, Forth Worth, TX; 1998.
[38] Liu G, Liu M. A rapid calibration procedure and case study for simplified simulation models of commonly used hvac systems. Build Environ 2011;46(2):409–20.
[39] C. Weatherstats, Weather data vancouver island (2016).
[40] Kleiminger W, Beckel C, Staake T, Santini S. Occupancy detection from electricity consumption data. Proceedings of the 5th ACM workshop on embedded systems for energy-efficient buildings. ACM; 2013. p. 1–8.
[41] Mahalanobis PC. On the generalized distance in statistics, National Institute of Science of India; 1936.
[42] Rousseeuw PJ, Leroy AM. Robust regression and outlier detection Vol. 1. Wiley Online Library; 1987.
[43] Kriegel H-P, Kröger P, Schubert E, Zimek A. Loop local outlier probabilities. Proceedings of the 18th ACM conference on Information and knowledge management. ACM; 2009. p. 1649–52.
[44] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12(Oct):2825–30.
[45] De Moivre A. The doctrine of chances. Annotated readings in the history of statistics. Springer; 2001. p. 32–6.

[4] https://gitlab.com/energyincities/energy-signature-analyser.
[5] https://besos.uvic.ca/.

[46] Yang J, Ning C, Deb C, Zhang F, Cheong D, Lee SE, et al. k-shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. Energy Build 2017;146:27–37.
[47] MacQueen J et al., Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, Oakland, CA, USA; 1967. p. 281–97.
[48] Cuturi M, Blondel M. Soft-dtw: a differentiable loss function for time-series. Proceedings of the 34th international conference on machine learning, vol. 70. JMLR. org; 2017. p. 894–903.
[49] Tavenard R. tslearn: A machine learning toolkit dedicated to time-series data, <https://github.com/rtavenar/tslearn>; 2017.
[50] Szekely GJ, Rizzo ML. Hierarchical clustering via joint between-within distances: extending ward's minimum variance method. J Classif 2005;22(2):151–83.
[51] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53–65.
[52] Air-Conditioning, Heating & Refrigeration Institute, AHRI Directory of Certified Product Performance; 2019. URL <https://www.ahridirectory.org>.
[53] Natural Resources Canada, National Energy Use Database, British Columbia, Residential Sector, Table 22 (2015–2016). <http://oee.nrcan.gc.ca/corporate/statistics/neud/dpa/menus/trends/comprehensive/trends_res_bct.cfm>.
[54] Kotrlik J, Higgins C. Organizational research: determining appropriate sample size in survey research appropriate sample size in survey research, Information technology, learning, and performance. Inform Technol, Learn, Perform J 2001;19(1):43.

Epilogue

This paper was the first in the literature to use energy signatures to segment buildings within a stock. Further, the information carried in the characteristic shape of a building's energy signature allowed us to infer building characteristics with an unsupervised learning algorithm. Here we used it to infer the heating system type and building type. This includes differentiation among heating systems with the same power supply, like heat pumps and electric furnace heaters.

The next step of this research would be to integrate the heating system information with other information like primary-use-type, building type and occupant behaviour [16][17]. Furthermore, it would be helpful to leverage other data sources like satellite data to receive better knowledge on the geometry of each building. Only afterwards the actual model calibration process can be done.

Instead, we shortcut that process and apply surrogate modelling to a synthetic data set. This ensures that a suitable base model is known. For this idealistic case, we benchmark surrogate based calibration with other popular bottom-up, i.e. physics-based, calibration approaches (including the balance point method, and resistance-capacitance model calibration) and with novel top-down deep learning methods. While other studies have benchmarked calibration performance by the predictive accuracy of the models, we are mostly concerned in retrieving accurate building design parameter estimates.

6.1 Benchmarking surrogate calibration

Advanced Techniques for Learning Quantitative Building Properties from Sensor Data: An Empirical Perspective on Competing Paradigms

Gaby Baasch[a], Paul Westermann[a], Ralph Evins[a,*]

[a]*Energy and Cities Group*
Department of Civil Engineering
University of Victoria, Canada

Abstract

Data-driven models are increasingly used to extract building energy performance characteristics from building sensor measurements. Building energy model calibration approaches are most frequently used. This may no longer be suitable when data becomes available from entire building stocks, as in calibration each building is considered individually. While it is possible to augment current approaches, e.g. with archetype energy model development, we argue that the ever-increasing amount of building time series data may enable a shift from building-by-building model calibration to supervised deep learning models, which excel at extracting temporal features from highly multivariate data streams from multiple buildings.

In this empirical study, we implement and benchmark the two disparate

Abbreviations: BES, building energy simulation; CNN, convolutional neural network; ES, energy signature; RC model, resistance-capacitor network model; RNN, recurrent neural network; HLC, heat loss coefficient;

*Corresponding Author

Email addresses: gbaasch@uvic.ca (Gaby Baasch), pwestermann@uvic.ca (Paul Westermann), revins@uvic.ca (Ralph Evins)

paradigms, i.e. building-by-building calibration and deep-learning-based building characteristics prediction. Seven different approaches are considered, including three lumped parameter model calibration methods, two building energy performance simulation model calibration methods and two deep learning methods. We test the methods on an open-source synthetic building data set consisting of 16,000 simulated buildings. It allows to describe practical efficacy and sources of inaccuracy for each of the seven approaches, providing novel and substantial insight including an analysis of the impact of climate, thermal mass, occupant behaviour and air-infiltration on the performance of each method.
Keywords: building characterization, data-driven retrofit, time series analysis, calibration, deep learning, surrogate model

---

1. Introduction

The building sector is on the brink of fundamental change. Digitization is transforming our understanding of a building from a passive, voiceless space into a constantly communicating, active service provider for healthy and sustainable living [1]. At the core of this transformation is sensor data, which provides a continuous stream of information on indoor comfort conditions and energy performance.

This time series sensor data also is a viable source of information for building diagnostics and analysis [2][3]. In particular, we can extract thermal characteristics which enables more effective energy retrofits [4], the derivation of accurate building stock models to predict future energy behaviour of neighborhoods, districts or cities [5], and also can be used for commer-

cial use like customer targeting (e.g. demand response targeting) [6].[1] The bandwidth of thermal characteristics we can extract is large and includes both categorical information, like the installed heating system type [8][9], and quantitative information like the whole building heat loss coefficient or the heating system efficiency [10].

In this paper we focus on the latter, which gives a estimate of the efficiency of a building. Traditional research in the field of building science, has retrieved these quantitative building properties using calibration of physics-based whole-building models, whose parameters are calibrated using the measurement data (*bottom-up approach*). The methods mostly differ in the complexity of the underlying model [11] ranging from the 1 or 2 parameters-based balance-point (or energy signature) model [12], to Resistance-Capacitor (RC) network models of various orders of complexity [13], and to complex building energy simulation (BES) model calibration approaches [10].

Outside the building domain many other disciplines have undergone a "big-data" paradigm shift. Although large amounts of sensor data are accumulating, many findings of various domains are yet to be transferred to the building sector. This is the key motivation for this paper and resulted in the following contributions.

[1]Please note, that not only thermal characteristics of a building can be extracted. For example, it has been shown that we can infer about the socio-economic situation of occupants [7]. This leads to privacy concerns and very few building sensor data sets are publicly available.

First, we compare traditional building calibration approaches with rapidly advancing, deep machine learning models [14] that have recently shown particularly high performance on time series modelling tasks [15]. Instead of a model with a predefined structure, like physics-based models, machine learning models are found algorithmically to optimally perform on a specified task (*top-down approach*) [16]. Here, we apply them to predict quantitative building characteristics, which is a supervised regression problem.

Industry adoption of characterization methods may be hindered by a lack of model transferability [17] caused by lacking robustness and reliability in the performance for practical use cases. We suggest to assess the methods using metrics that represent practical application cases such as building stock modelling and retrofit analysis. This is considered in this work where we quantify the robustness of the method to four confounding building factors (those mentioned above).

The rapid development of machine learning has largely been supported by the ecosystem of the research field. It includes public data sets, open code repositories and transparent benchmarking. In the domain of building characterization and building calibration many studies exist where individual methods are applied on specific buildings, whose data may even be undisclosed. This problem has been acknowledged, and large efforts are being undertaken both to share data and to benchmark thermal characterization approaches. For example, a repository hosting meter data of thousands of buildings was initiated [18] and companies have offered their meter data for

research purposes [19][20]; or in [21] multiple characterization methods were compared.

We contribute to these efforts and introduce an open-source, extensible simulation-based synthetic building sensor data set. In comparison to the public real world data sets, a synthetic data set provides us with full information on the building and ground truth data (labels) on any thermal characteristic the method developer is interested in is easily accessible. The data set can be continuously upgraded to be suitable for certain research objectives. In this paper for example, we use it to understand how climate, construction materials, air-infiltration and stochastic occupant behaviour affect the performance of seven different thermal characterization approaches.

All of the data and code used for this work is available in a GitLab repository.[2] As such, this work serves as a catalyst for future studies by providing: (1) a preliminary comparison of several popular methods in the literature and state-of-the-art machine learning approaches, (2) a reusable benchmarking ecosystem and (3) robust performance metrics that measure the practical efficacy of the methods.

1.1. Structure of the paper

In the following we first provide background knowledge on the data sources serving as input to building characterization methods and on the bottom-up and top-down building characterization approaches. Subsequently, we introduce the data set we generated and highlight its use to understand the impact

[2] `https://gitlab.com/energyincities/bp-benchmarker`

of four confounding factors on the model accuracy (climate, construction materials, air-infiltration and stochastic occupant behaviour). Then, we provide details on all seven characterization methods applied and benchmarked in this paper and discuss the results.

2. Background

2.1. Data sources

Here we infer thermal building properties using time series data produced by sensor measurements. While data on buildings have been collected for decades already, the number of sensors, their sampling rate and standardization of data acquisition is ground breaking. Worldwide, more than one billion smart metering devices will be installed by the end of 2020 [22], and large construction markets have or will have (e.g. Canada [23]) nationwide coverage. Types of measurement data include on site measurements such as indoor temperature data and heating power, as well as off-site measurement data from weather files such as temperature and solar radiation.
Different sensing devices exist and are usually packaged in smart thermostat systems or smart meters. Smart thermostats, for instance, may record the indoor temperature as well as other variables like the heating system usage [19]; smart meters most commonly only report hourly or subhourly energy consumption [6].

2.1.1. Whole-Building heat loss coefficient

The automated smart metering devices do not capture any explicit thermal characteristic information (e.g. U-values) of the buildings to validate our

methods or to train supervised machine learning models. Labeling the buildings usually requires extensive manual efforts (e.g. measuring the U-value of a building). That scarcity in building meter data sets is a pressing problem in the field [17] and automating labelling of buildings is being explored [24][25].

When using a synthetic data set, full information of the thermal properties of the buildings are given or derivable. Here, we compute the whole-building HLC (see Appendix 8) which quantifies the rate at which heat is lost through the building envelope via convective, conductive and radiative forces, as well as infiltration. This knowledge is instrumental for estimating the benefits of building retrofits [4] or assessing the quality of a building post-construction [26].

2.2. Modelling paradigms

The problem of data-driven thermal property estimation has been approached from several angles. Two distinct paradigms emerge [10]: (1) bottom up, where the unknown characteristics, represented by parameters of an engineering model, are found by parameter calibration; and (2) top down models, where the a purely data-driven model is trained on labelled data to predict characteristics of future buildings.

This section provides a high level overview of these approaches, including the major differences between the two and relevant barriers to application. More background on the specific methods implemented in this work can be found in Section 3.2.

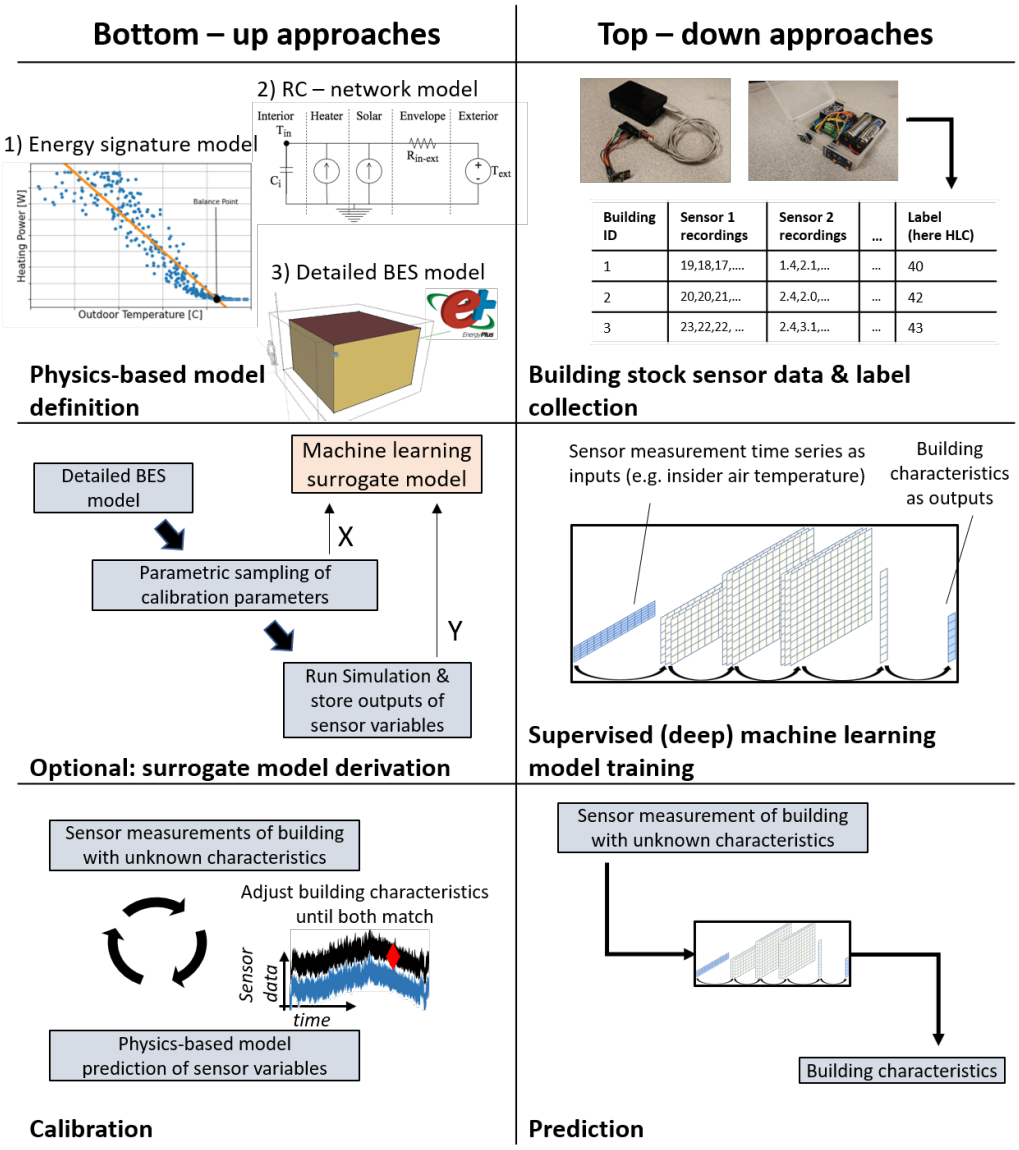


Figure 1: General overview of the bottom up and top down paradigms for thermal property estimation in buildings. The models in this image represent the those that are tested in this paper: (1) Energy Signatures, (2) RC-Networks, (3) Surrogate-Based Building Energy Simulation and (4) Supervised Deep Neural Networks.

2.2.1. Bottom up approach

Bottom-up engineering model calibration is the dominant approach in the thermal building property estimation literature [10] [21] [27]. It relies on an underlying physical model that uses physical laws to predict dynamic building behaviour (often indoor temperature time series). Parameters, representing the unknown thermal properties, are incrementally updated in an optimization loop that reduces the error between the real time series and the simulated time series (Figure 1). It is commonly used in the context of building control [28] [29] [30], but also can be used specifically to find building characteristics [31] [32] [33] [34] [35]. When used for control, the major goal is to achieve accurate building behaviour predictions. Hence, the predictive accuracy of the model after being calibrated is the key error metric. When used for property estimation it is crucial that the discovered building parameters match ground truth values. This is pursued in this study.

A variety of techniques for model calibration exist, they differ by complexity of the underlying physical model and technique to calibrate the parameters. For the purpose of this paper, energy signature calibration, RC network calibration and BES-based surrogate calibration (either using black-box optimization algorithms or using Markov-Chain Monte Carlo sampling) were considered and are described further in section 3.2. Each of these approaches have different data input requirements and implementation workflows, while the requirements for supervised learning are again unique. Figure 1 and Section 2.2.3 highlight these differences.

2.2.2. Top down approach

Top down approaches can be formulated as an unsupervised (ex. clustering) or supervised (ex. classification or regression) machine learning prediction problem. For numerical HLC quantification regression, i.e. supervised learning, is the natural choice. Unlike the physics-based calibration techniques described above, black-box methods require no prior knowledge of system dynamics, but create a model-agnostic mapping from time series inputs to the building quantity of interest by training on labelled data (Figure 1). Deep learning is a highly popular subfield of machine learning that creates such mappings through multiple layers of increasingly abstract representation throughout training [14].

Compared with calibration, applications of supervised learning to thermal property estimation are fairly limited. This is particularly true for deep learning, for which only a few studies exist [36] [37] [38]. Neural networks fell out of favour in this domain likely because the required ground-truth labels are rarely available. Recent work, however, showed that deep learning can be successfully applied for HLC estimation [39]. Its inclusion in this study therefore provides a novel prospective into the state-of-the-art for thermal property estimation. Further, the discussion section provides suggestions for overcoming label scarcity.

2.2.3. Data requirements and workflows

1. Energy signatures
    - No pre-training required.
    - Calibration error between measured and simulated time series.

2. RC networks
    - Require the selection of an underlying physical RC network model, which may be different for every building.
    - Calibration reduces error between measured and simulated time series.
3. Surrogate-based building energy simulation (BES)
    - Require a building energy simulation (BES) model that contains a very detailed description of a building [40].
    - BES has long simulation run times. To overcome this, machine learning models (i.e. a surrogate models[3]) are trained on a low number of simulation samples (with calibration parameter values as inputs, simulation outcomes as output) to emulate the BES model [44].
    - Calibration reduces error between measured and simulated time series.
4. Supervised deep neural networks
    - Require a large, representative training dataset with high-fidelity labels.

[3]In general, surrogate models are used in two kinds of paradigms. Either they are used to increase speed of an optimization algorithm minimizing the distance of simulation outputs and measured data [41], or they are used to increase the speed of sampling a probability density function (posterior) of the searched parameter value [42][43][31]. Both approaches are applied in this paper.

- Predicts based on characteristics and patterns of representation learned from the training data.

3. Methodology

In the following we introduce the synthetic data set, comprising sub-hourly metered energy data of 16,000 buildings, and seven methods to estimate the HLC.[4]

The synthetic data set is designed to assess the methods robustness towards confounding factors including stochastic occupant behaviour, air-infiltration, thermal mass of the buildings, and climate. We provide three error metrics to assess the overall performance of each method and their robustness when confronted with the four impact factors.

3.1. Synthetic dataset

We generated the meter data of 16,000 buildings by running parametric simulations using BESOS [45] and EnergyPlus [40].

Synthetic data was used for two fundamental reasons. First, labelled building meter data sets are rare [46]. Second, building simulation grants the control required for large-scale parametric studies that manipulate specific building characteristics relevant to understand the relative performance of characterization approaches. To perform robust, comparative modelling studies the use of simulation data is thus warranted - in fact, arguably unavoidable. Still,

[4]More information on the methods can be found in the code repo `https://gitlab.com/energyincities/bp-benchmarker`

the observed relative performance of the models on simulation data cannot necessarily be guaranteed on real data.

3.1.1. Data creation pipeline

The buildings data creation pipeline in this work is similar to that in [39]. The Building and Energy Simulation, Optimization and Surrogate-modelling (BESOS) platform [45] enables to run quasi-random latin-hypercube-sampling of building design parameters, as listed in Table 1 [47]. These parameter combinations are fed as input to the building simulation software EnergyPlus, version 9.2.0 [40]. Outputs from EnergyPlus include various information on the building performance such as time series values of all relevant thermal variables. A set of time series, and computed HLC values, were stored to form the final, labelled data set (see section 3.1.4).

3.1.2. Baseline building model

To generate the building data set, two baseline building models were first defined: one wooden building and one concrete building. Table 1 specifies the material composition of the buildings. Both buildings have a constant geometry, a simple $5m$x$5m$x$3m = 75\ m^3$ box with one zone, four $4m$x$1.5m = 6m^2$ windows and no unconditioned spaces (Figure 2). Additional modelling assumptions are listed below.

- The floors were designed as adiabatic. Ground heat loss effects are difficult to simulate and therefore, neglected for now [48].
- No mechanical systems were modelled, but EnergyPlus ideal air loads were used instead.

Surface	Material Layers	Thickness Ranges (m)
Wall	Stucco	[0.015, 0.030]
	Plywood or Concrete	[0.006, 0.03] or [0.2, 0.3]
	Insulation	[0.035, 0.3048]
	Gypsum	[0.00633, 0.0159]
Window	Glass	[0.001, 0.01]
	Air Gap	[0.006, 0.02]
	Glass	[0.001, 0.01]
Floor	Plywood or Concrete	0.0127 or 0.1016
Roof	Roof Membrane	[0.0012, 0.0095]
	Insulation	[0.1, 0.3]
	Metal Decking	[0.0007, 0.0015]

Table 1: Material composition of the buildings and the thickness ranges used for parametric generation of buildings meter data for our synthetic data set.

- Constant setpoint schedules were employed across all cases.
- Infiltration is modelled according to the DOE standard.[5] Complex airflow networks and ventilation were ignored.

3.1.3. Building design parameters

Figure 2 showcases the manipulated characteristics. For each of the wooden and concrete building baselines, the material thicknesses were varied to create 1000 buildings with distinct HLC values. To do so, thickness

[5] `https://bigladdersoftware.com/epx/docs/9-2/input-output-reference/group-airflow.html#zoneinfiltrationdesignflowrate`

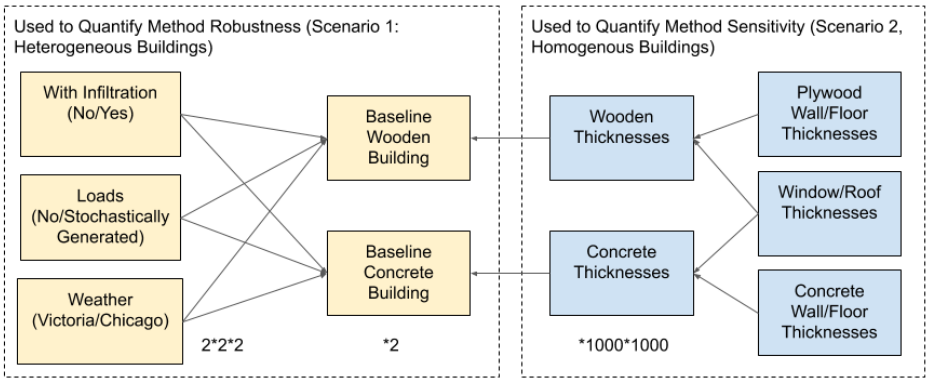


Figure 2: Manipulated building design parameters. In total 16,000 buildings were created. The building properties on the left were used to quantify the robustness of the methods to diver stock-level characteristics. The properties on the right were used to measure the sensitivity of the methods to changing HLC values.

ranges for each of the materials was defined according to engineering standards and randomly sampled for each new building. Each of these sets of buildings was then simulated with annual weather data from two different climates (Victoria, CA and Chicago, USA), with and without air-infiltration (flow per exterior surface area of 0 and 0.00085 m^3/sm^2), with and without equipment and occupancy loads[6], for a total of $1000 * 2 * 2 * 2 * 2 = 16,000$ simulated buildings. Considering the 2 material types, 2 infiltration rates, 2 climates and 2 load cases $2^4 = 16$ experimental conditions are considered.

3.1.4. Simulation outputs: temporal measurements and HLC

EnergyPlus outputs a myriad of time series variables that describe the detailed temporal behaviour of a building over the course of a simulation. Some of these variables can be measured with sensors in a real building, including external temperature (T_{ext}), internal temperature (T_{in}), heating

[6]The stochastic equipment and occupancy loads were generated with the richardsonpy library from `https://github.com/RWTH-EBC/richardsonpy`. A distinct stochastic schedule was generated for each building that included loads.

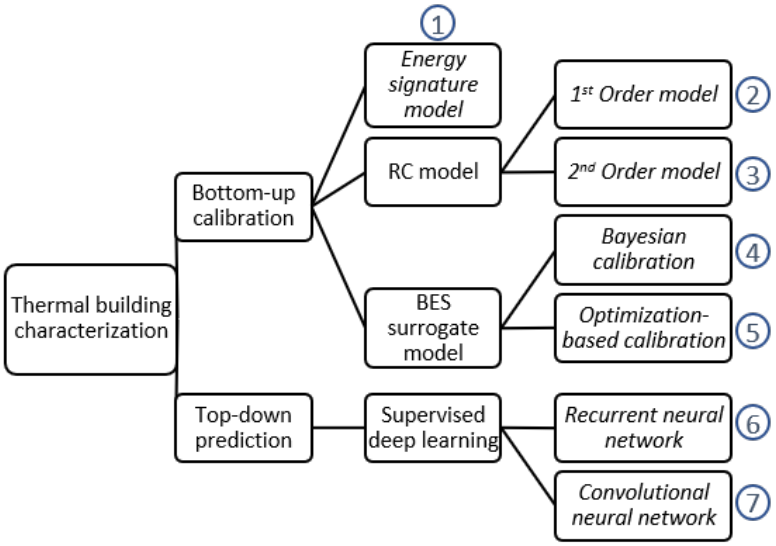


Figure 3: The investigated research paradigms and method implementations.

system power ($\dot{Q}_{hsys}$) and solar gains ($\dot{Q}_{sol}$). For this study, 5 minute time steps were output by EnergyPlus and used for all modelling approaches but the energy signature method, which aggregate values daily (see Section 3.2.1). Similarly, all models use one week's worth of data in January, aside from the energy signature method, which uses one years worth of data.

The selected time series variables, granularities and measurement periods were chosen according to previous studies [48] [13].

As described in Section 2.1.1 and Appendix 8, the whole-building HLC [W/°K] depends both on the heat loss from infiltration and the heat lost through the building envelope. These were calculated analytically from the outputs of EnergyPlus. The full calculation can be found in the Appendix.

3.2. Seven thermal building characterization approaches

Within the two paradigms, i.e. calibration of bottom-up models and top-down prediction, several approaches and implementations exist. Figure 3 summarizes the paradigms and corresponding implementations considered in this paper. They are not exhaustive, but provide a baseline for continued analysis and benchmarking. We propose that future researchers developing new methods use our implementations for comparison. In this section, a technical description of the computational approaches will be provided.

3.2.1. Energy Signature (ES) Calibration

The underlying physical model for ES calibration, which is a standard approach in building energy modelling [49] [12] [50] [33] [51] [52], is a basic reformulation (equation 1) of the whole-building energy balance (equation 9). The energy demand of a building is plotted against outdoor temperature. Typically, each point on the plot represents the mean heating load and outdoor temperature for a single day. Above a particular value for the outdoor temperature, known as the balance point, no energy is required to heat the building. The slope of the line of best fit below the balance point represents the HLC.

$$\dot{Q}_{h,d}(\overline{T}_{ext}) = HLC_{wb}(\overline{T}_{in} - \overline{T}_{ext}) + \dot{Q}_{baseline} \tag{1}$$

3.2.2. Resistance-capacitor (RC) network calibration

In this popular approach, a building is modelled using an RC network and an associated set of stochastic differential equations [53].[7] The RC model can be defined at differing orders of complexity, from simple networks with a single lumped capacitance to complex, multi-order systems [13].

Selecting the appropriate RC model for a given building is a non-trivial task [46] [27] [48].[8] In this paper, only the results for two, low-order RC network implementations are presented (see Appendix 10) because the parameter estimates worsened for higher model orders. This result will be discussed further in Section 4.

3.2.3. Surrogate-based BES calibration using optimization

Instead of lumped parameter models, more detailed physics-based models are sometimes favoured. For that purpose building energy simulation (BES) models [40] can be manually or automatically calibrated [10]. This calibration process relies on iteratively adjusting calibration parameters to match simulation time series outputs and measured data. This can be computationally expensive and machine learning based surrogate models are used instead [42][31][54]. A surrogate model approximates the BES model, by learning

[7]More detail is provided by Bacher and Madsen at [13] and [53], but, briefly, statistical maximum likelihood estimation is applied to estimate the unknown parameters in the model. Specifically, a Kalman filter is used to estimate the likelihood function, and an optimization algorithm is used to find the set of parameters that maximize the likelihood function.

[8]Building archetypes with pre-defined RC models might help to alleviate this problem. This is explored further in the discussion.

from a few simulation runs to estimate the effect of changes in parameter values (surrogate model inputs) to changes in simulation outcomes (surrogate model outputs) [55].

BES models produce non-linear, multi-modal outputs with possible discontinuities [56]. Therefore, black-box optimization approaches such as genetic algorithms (GAs) are often applied to determine suitable parameter choices [41]. Here, we use the NSGA-II optimization algorithm (population size = 200, offspring size = 100, iterations = 3000), minimizing the summed distance of simulated daily heating demand and measured daily heating demand [57]. The approach is similar to [41], but uses higher frequency data (hourly instead of monthly).

3.2.4. Surrogate-based BES calibration using Bayesian calibration

The BES model time series outputs $y = G(\mathbf{x}, \Theta)$ may be seen as a function of the known parameters of a building x, and the unknown characteristics Θ, i.e. the vector of calibration parameters [58]. When comparing simulation outcomes to measurement data z, Θ may be adjusted such that the error between y and z becomes small.[9]

Following Bayes' theorem, a posterior for the unknown parameters Θ, i.e. a probability density function approximation of the calibration parameters, can be inferred using the *(i)* sensor measurements z, *(ii)* simulated model outputs

[9]Following [58], the relationship of y and z can be modelled with $z(x) = y(x, \Theta) + \delta(x) + \epsilon(x)$, where z represents the measurements, y represents the BES outcomes, $\epsilon(x)$ represents errors in measurements (aleatoric uncertainty) and $\delta(x)$ corresponds to the error induced by the model bias (epistemic uncertainty). Often the model bias δ is not explicitly modelled for building calibration [31][59].

y, *(iii)* and a prior probability $p(\Theta)$ for the calibration parameters. The prior integrates existing knowledge of the modeller (e.g. range of common wall thickness values).[10]

$$p(\Theta|z) \propto p(z|\Theta)p(\Theta) \tag{2}$$

As we only know an proportional result of the actual probability density, we use Markov-Chain Monte Carlo (MCMC) sampling, here the Metropolis-Hastings (MH) algorithm, to approximate the posterior $p(\Theta|z)$. The algorithm generates a sequence of guesses for Θ, where the MH algorithm determines which guesses to keep and which to discard. The pool of accepted guesses approximates the true posterior $p(\Theta|z)$, often visualized as a histogram. In this study we derive only one histogram for one calibration parameter, the HLC.

That MCMC sampling process usually requires thousands of simulation runs which motivates the use of surrogate models. Most commonly a Gaussian Process surrogate model is used with a explicit formulation of the likelihood $p(z|\Theta)$ [31]. When using non-GP surrogate models, other approaches exist to compute a likelihood function exist. We use the approach suggested by [61], where the likelihood is given by $p(z|\Theta) = exp(-\frac{\sum|z-y|^2}{2\sigma^2})$ which can be used with any surrogate model type.[11] Further, we specified a uniform

[10] It should be carefully chosen as it has significant effects on the outcome of the posterior estimate [60].

[11] This assumes identically distributed errors in the BES approximations with zero mean and constant variance σ^2, see [61].

distribution for each parameter bound by the maximum and minimum heat loss coefficient observed in the data.

3.2.5. Gated recurrent neural networks

Recurrent Neural Networks (RNNs) account for temporal input structure and are a common choice of neural network architecture when working with time series data. Vanilla RNNs suffer from something known as the vanishing gradient problem which prevents them from learning long-term temporal dependencies in data. Several work arounds for this problem are available, including Gated Recurrent Units (GRUs) [62] and Long-Short-Term-Memory Units (LSTMs) [63]. GRU has been shown to outperform LSTM in terms of runtime and accuracy [64] so it was chosen for this paper.

3.2.6. Residual neural networks

Much of the success of deep learning can attributed to the Convolution Neural Networks (CNNs). Intuitively, these networks operate by detecting local correlations in input data and later merging semantically similar features to produce a final output. 1-Dimensional CNNs have shown to be useful for a variety of time series applications, include speech-to-text, music generation and time-series classification [14] [15]. Residual Neural Networks (ResNets), introduced by He et al., are a CNN variant that allow for the training of very deep neural networks by introducing "skip-layers" which propagate lower representations forward through the network. They have exhibited state-of-the-art performance on various tasks, including time series prediction [65] [15].

3.3. Performance assessment

In this section, we provide metrics to assess the performance of each of the seven approaches to extract building characteristic from a heterogeneous stock of buildings. The experimental design of the data set allows for quantification of the impact of thermal mass (wood vs. concrete), infiltration, occupant behaviour, and climate on each of the methods (see Figure 2. As a result, it lets us infer about robustness of each method with details on causes of low performance.

The metrics we use to quantify the performance of each method are introduced in the following:

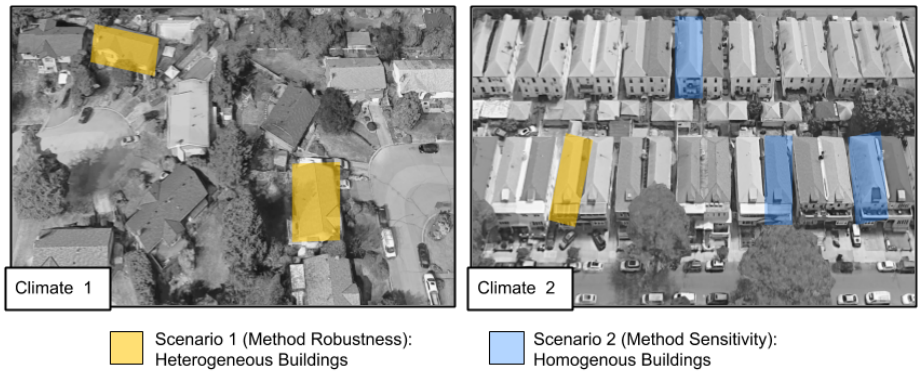


Figure 4: The quantitative metrics used in this paper are designed to measure the robustness and the sensitivity of the methods to changing building properties. Scenario 1 is concerned with the robustness of a method to heterogeneous buildings across the building stock. Scenario 2 considers measures the sensitivity of a method to differing material values (ie. HLC).

3.3.1. Performance metrics

The goal of each approach is to estimate the numerical HLC value for each of the 16,000 buildings. These estimates might be used for application cases such as non-intrusive, data-driven retrofit analysis and heterogeneous building stock modelling. For practical application cases it is not sufficient to measure only the goodness-of-fit of the model; the robustness and the sensitivity must also be quantified. The former quantifies model performance across heterogeneous building properties, while the latter measures sensitivity to differing HLC values for homogeneous buildings. These are formalized in the following.

The relationship between the measured and predicted HLC values provides descriptive statistics for these performance metrics. In general, a computational model is perfect if the predicted and actual values align on the diagonal when plotted against each other. In this case, the line-of-best-fit between these values will have a coefficient-of-determination (R^2-score) of 1 and a slope of 1. The mean absolute error (MAE) between the predicted and actual values for buildings will be 0. Note that even a model with a perfect R^2-score and slope can have an MAE of any magnitude if the line-of-best-fit is shifted up or down.

$$\mathrm{R}^2(HLC, \widehat{HLC}) = 1 - \frac{\sum_{i=1}^{n}(HLC_i - \widehat{HLC}_i)^2}{\sum_{i=1}^{n}(HLC_i - \overline{HLC})^2} \tag{3}$$

$$\mathrm{MAE}(HLC, \widehat{HLC}) = \frac{1}{n}\sum_{i=1}^{n}|HLC_i - \widehat{HLC}_i| \tag{4}$$

where $\widehat{HLC}_i$ is the predicted value for building i, HLC_i is the actual value for building i and $\overline{HLC}$ is the mean HLC.

To perform robustness and sensitivity analysis, separate linear regressions were performed on the modelling results for each of the 16 cases in Figure 2.[12] To understand why, consider the following two scenarios.

1. In the most common scenario we analyse a stock of buildings with heterogeneous characteristics (Figure 4: Scenario 1). A method can be considered robust and unbiased if it does not systematically over or under predict HLC for certain characteristics. We assess this systematic bias using the impact of each of the four experimental impact factors (construction material/thermal mass, infiltration levels, occupant behaviour, weather conditions). This impact can be quantified by computing and comparing the error distributions of buildings with similar characteristics within that heterogeneous building stock. For example, if a set of buildings with matching HLC values exist both in Chicago and in Victoria, a robust method must produce the same error distribution in both cases.
2. In some cases, heterogeneity in the building stock is small, i.e. similar[13] buildings are under assessment (Figure 4: Scenario 2). This may occur for a specific neighbourhood or district. In this scenario we can ease the requirements; a method must primarily be able to rank the HLC values of buildings correctly. This can be measured with the slope of the line-of-best-fit between the predict and actual values within a

[12] To avoid overfitting, machine learning methods require a separate data set for training and validation. The regression results presented in this work are on a pre-defined validation set that was not used for training.

[13] Here, similarity specifically refers to the building characteristics defined in Figure 2.

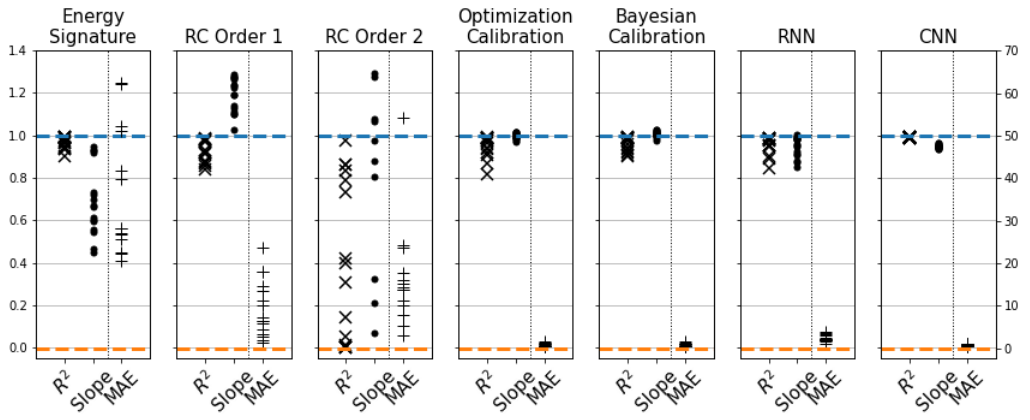


Figure 5: The summary statistics acquired by comparing the predicted and actual thermal property values for all the computational methods and building property case studies. The three columns in each subplot represent the R^2 (goodness-of-fit), the slope (sensitivity) and the MAE, respectively. Each of these three columns contains 16 data points; one for each experimental case. The values for R^2 and slope fall between 0 and 1.4, while the MAEs fall between 0 and 70. The dashed blue line represents R^2 and slope for a perfect model, while the dashed orange line represents the MAE for a perfect model. The wider the spread of the MAEs, the less robust the model.

particular case.

4. Results

In this section we assess the performance of the seven methods by quantifying the performance metrics on a building characterization task introduced above. They are summarized graphically in Figure 5 and numerically Figures 6, 7 and 9. The metrics provide us with a comparative overview on the performance of each of the seven approaches. We find a much larger spread in the performance of the studied gray-box methods in comparison to the surrogate-based BES model calibration and top-down deep neural network

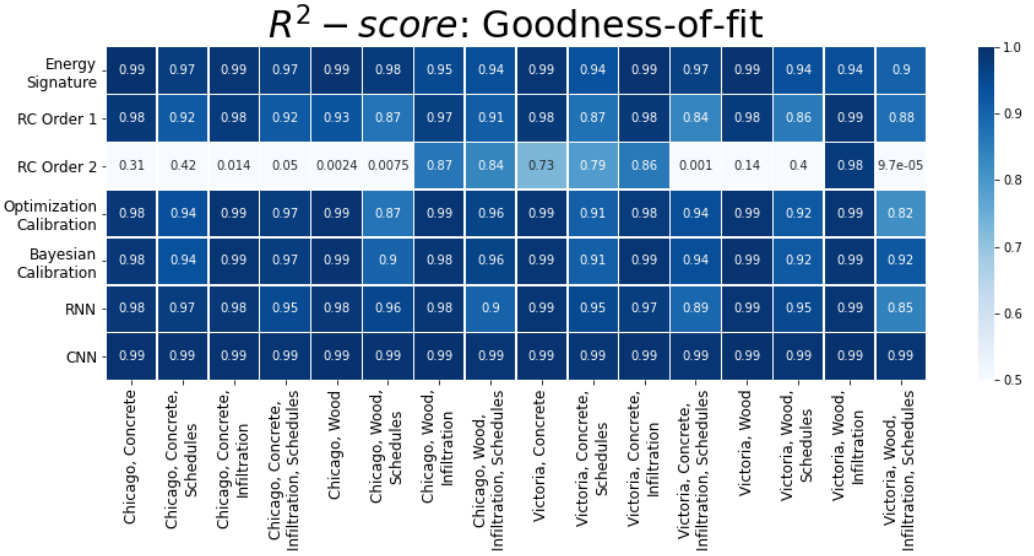


Figure 6: Numerical results for goodness-of-fit, as measured by the $R^2 - score$. A score of 1 indicates the best fit.

models. This is discussed in the following where the R^2 performance of the models within each experimental condition is discussed first.

4.1. Goodness-of-fit

In this work, the goodness-of-fit is measured by the $R^2 - score$ which was computed separately for all of the 16 experiments. The threshold for goodness-of-fit under which a method is no longer reliable is in reality somewhat arbitrary. In Figure 6 it can be seen that all of the methods aside from RC order 2 achieve a score over 0.8 for every experimental condition which indicates a strong goodness-of-fit within each case.

Surprisingly, the 2nd order RC model performs worse than the 1st order model. This can be explained by the calibration approach minimizing the error between the predicted and actual time series. Higher order models,

which are calibrated with more time series, have more parameters making them more variable but also easier to overfit to the input data. This explains the high variability in errors in the HLC estimates on our test data.

Figure 6 shows that the CNN consistently performs the best in terms of R^2-score. RC order 2 performs the worst by far; for many cases its R^2-score is close to 0. Still, in some cases (for example the wooden building with infiltration in Victoria) the method performs well. This shows that a single case study might yield the method to be reliable, even if this is not the case in general. Literature tends to run case studies that validate methods on only a single building without varying properties or climatic conditions; the result here provides strong evidence that this is not sufficient.

Evaluating the rows of the heatmap, it can be seen that the other methods have the lowest R^2-score for the cases with stochastic schedules. It follows that, of the experimental conditions that were tested in this study, the addition of stochastic loads has the largest effect on the method's goodness-of-fit.

4.2. Accuracy and robustness

We test if a method produces accurate and robust $\widehat{HLC}$ for a heterogeneous stock of buildings by checking how much the error is impacted by extraneous building properties. The distribution of the errors of $\widehat{HLC}$ must be similar for all experimental conditions for a method to be considered robust (recall Section 3.3).

Figure 7 provides a numerical summary of the MAEs and the boxplots in Figure 8 provide a visual summary of the distributions. The first thing to notice is the differences in magnitude of the errors between the methods. The Energy Signature and RC Models have much higher errors in general

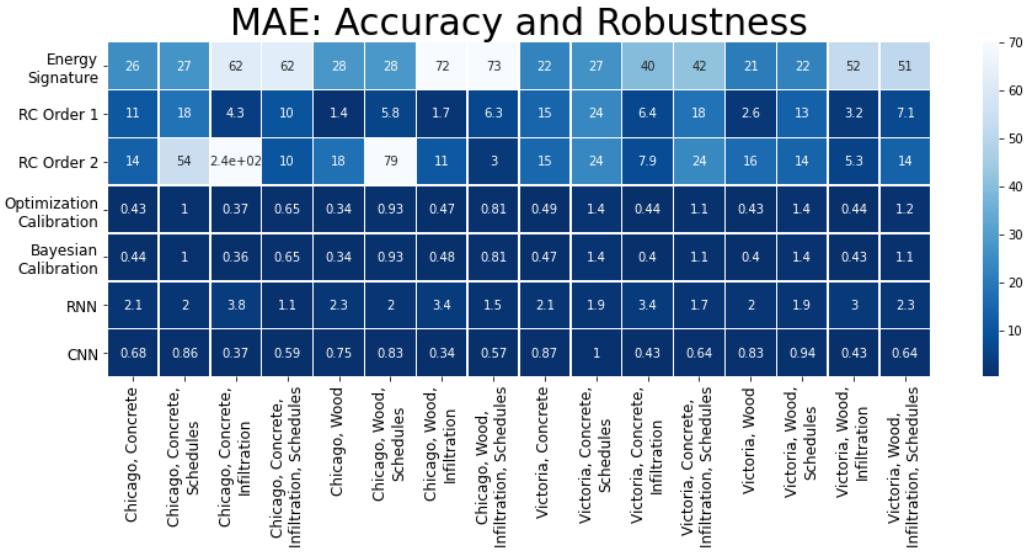


Figure 7: Numerical results for the MAE. This provides a metric for accuracy and begins to indicate the method robustness. And MAE of 0 indicates perfect accuracy. The larger the difference between MAEs for a method, the less robust the method.

than the surrogate-based BES calibration and deep learning approaches. The worst performing model in terms of absolute error is RC order 2, followed by the Energy Signature method and then RC order 1. For the surrogate-based BES calibration approaches and for the RNN, the errors are generally under 7, while the CNN has errors that are always below 3.

4.3. Shifted error distributions

A statistically significant difference ($p<0.05$) in the error distributions for all of the evaluated methods was found. While this is noteworthy, it is not the only consideration, as the distributions may have a statistically significant difference that is meaningless in practice.

In the remainder of this section we will highlight for each method whether the

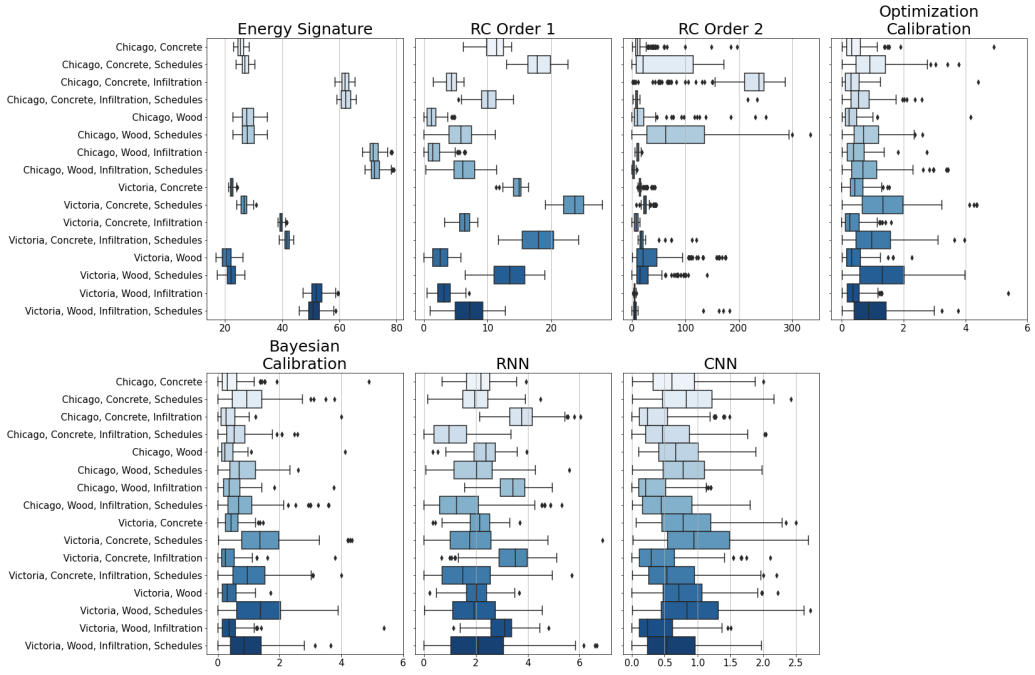


Figure 8: Boxplots are a standard approach for summarizing data distributions. In this figure, the error distributions for each of the 16 experimental cases are plotted. The difference between these distributions indicates a method's robustness to changing properties. Note that for the BES calibration approaches, a few outliers were found. For readability, these were not included.

shift in the error distributions is practically relevant. Analyzing the results in this way will also highlight which of the evaluated confounding factors have the most significant effect on the modelling results. It will be shown that the confounding factors do not affect the methods in the same way. For example, the largest difference in error distribution for the Energy Signature method is caused by adding infiltration, while this is not the case for RC order 1.

For clarity, only the most important features of the data are discussed. The

reader is encouraged to analyze the results further.

- **Energy Signature**: For this method, there is a clear distinction between the cases with and without infiltration. All else held equal, the buildings with infiltration result in a much higher MAE than the buildings without infiltration. This indicates that this method is not able to model infiltration properly and thus is systematically biased. To model HLC for a heterogeneous building stock a method must be able to account for infiltration. Thus, the Energy Signature method is not suitable for this application.

- **RC order 1**: From Figure 8, it is clear that there is a fairly large difference in the error distributions between the cases for this method. Unlike the Energy Signature approach, this method tends to find lower errors for the buildings with infiltration and the capacitance of the envelope (i.e. concrete vs wood) has an affect on the predicted HLC. The RC order 1 method also yields larger errors for the buildings with schedules than for buildings without schedules, and the spread in the errors tends to be wider. The largest MAE found was 24 (Victoria, concrete, schedules) and the lowest was 1.4 (Chicago, wood). Over all, this model is not robust to the extraneous factors tested in this work.

- **RC order 2**: Of the examined methods, RC order 2 exhibited the least robust. This is likely due to over-parameterization (discussed previously).

- **Optimization Calibration**: This method is most susceptible to the presence of stochastic schedules. In general, the errors for the cases

with schedules are larger and more variant. The largest MAE is 1.4 and the lowest is 0.34. In a practical scenario this is likely insignificant, but it is up to the user to decide.

- **Bayesian Calibration**: The results for Bayesian Calibration are very similar to those for Optimization Calibration.

- **RNN**: Unlike the other methods, the RNN exhibits lower MAEs for buildings with schedules than for building without schedules. Overall, it tends to perform most poorly in the infiltration cases without schedules. Compared with the surrogate-based BES calibration approaches, the RNN method finds a larger differences the error distributions, but these differences are still much smaller than for the gray-box calibration approaches. Again, it is up to the practitioner to decide whether the differences in the error distributions is significant in practice.

- **CNN**: The CNN has the lowest errors across all cases. The MAE is less than or equal to 1 in every case and there are less outliers with high errors. The greatest differences in error distributions are caused by infiltration and stochastic schedules, but these differences are likely insignificant in practice.

4.3.1. Sensitivity to material properties

The slope of the line-of-best-fit indicates the sensitivity of the method to differing material parameters (see Figure 1). Of the tested methods, the surrogated-based BES calibration approaches achieve slopes that are closest to 1, with Bayesian calibration slightly outperforming optimization calibration. Again, RC order 2 performs worst by far, but even so this method

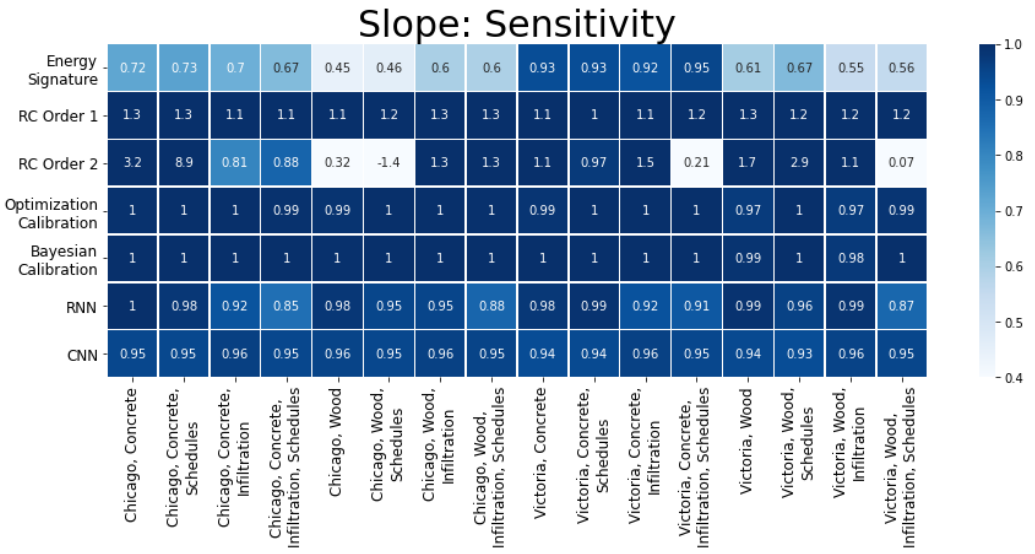


Figure 9: Numerical results for method sensitivity to differing material values. A value of 1 indicates perfect sensitivity.

achieved a slope of 1 ± 0.1 for 6/16 cases. RC order 1 tends to find slopes above 1, while the Energy Signature approach finds slopes that are less than 1, so the former is the most sensitive to changes in HLC, while the latter is the least sensitive. For the Energy Signature approach, the concrete buildings in Victoria have slopes closest to 1. Of the deep learning approaches, the CNN outperforms the RNN.

4.4. Summary

Based on the analysis above, all of the models except for RC order 2 exhibit a strong goodness-of-fit and reasonable sensitivity to differing material properties. Even so, none of the gray-box calibration approaches are robust to heterogeneous building properties. This is a first indicator that supervised deep learning approaches may become a key element in data-driven retrofits,

stock modelling and demand-response management, but the presented study is constrained by the experimental set-up and the results cannot be generalized without appropriate consideration. The limitations of this work and the implications on the results will now be discussed.

4.5. Assumptions and limitations

The provided comparison of methods is valid under strong modelling assumptions.

- The building design parameters, that are not calibrated, are assumed to be perfectly known. That means that the BES calibration model is the same one that was used for generation of the synthetic dataset. Similarly, the training and test data for the deep learning model was generated by the same BES model.

- RC-model design was kept at its minimum. Potentially, a better suited model structure exist. This however, acknowledges that automated RC-model design is currently lacking in literature.

- The CNNs were trained across all climates, materials, infiltration cases and cases and are thus less over-specified than the surrogate methods, which were trained individually for each experimental case. Further, the surrogate calibration only used heating rate as input, while all of the other methods used four time series variables. Regardless, the major conclusions of this study remain the same.

Based on these assumptions the given results are biased in favour of learning based methods. Future work is required to generalize these methods, e.g.

the expected accuracy-loss due to model misspecification in case of BES calibration needs to be studied and quantified. Nonetheless, the results highlight the sensitivity of well-established methods towards building material choice, air infiltration, stochastic occupant behaviour, and climate, and motivates further research into deep learning based approaches.

5. Discussion

The goal of this paper was to transfer machine learning research to foster data-driven building characterization. We contributed with a holistic test of novel machine learning methods to predict building characteristics using sensor data as input and compared them to traditional model calibration approaches. Therefore, we generated a synthetic data set which offers an experimental environment to test the methods' robustness towards four factors that possibly confound characterization accuracy.
The results show the risk of using data-driven methods for building characterization as commonly occurring factors, like stochastic occupant behaviour, significantly impact the performance of traditional methods. Novel deep learning methods reach higher overall performance and robustness, but remain far from application due to practical constraints like the lack of sufficient labelled training data.

In the following we discuss the experimental results. Further, we discuss the advantages and disadvantages of using synthetic data set to benchmark building characterization methods, which we complement with pointing out the key differences to real world data. Lastly, we propose promising directions for future research, where we set the focus on young machine learning based

paradigms like bottom-up surrogate-based BES calibration and top-down deep learning.

5.1. Comparison of methods

5.1.1. Overall accuracy and ranking accuracy

We compared all methods focusing on their robustness towards changes in extraneous impact factors. The robustness is measured by the change in error in the heat-loss coefficient estimate and by the change in accuracy of sorting the buildings by their HLC (ranking). For a method to be suitable for data-driven retrofit and building stock characterization it must perform well in both cases.

For ranking similar buildings with regard to their HLC, it was shown that the energy signature method and the second order RC model perform the worst, while the CNNs and the optimization-based BES calibration performed the best.

With regard to overall robustness in the model performance, CNNs are the most robust towards differing building properties, followed by the RNN and the surrogate calibration approaches. None of ES, first order RC model or second order RC model were robust to changing building properties. Given the stated assumption in section 5.2, we can conclude that ES, first order RC model or second order RC model cannot characterize building HLC across differing building properties with statistical certainty, rendering them unsuitable for HLC characterization in practice. This is a significant result of this study and should be confirmed with real world data. OBC, BC, RNN and CNN have strong predictive capabilities and are robust towards changes in

building properties, but they have practical barriers to usage.

5.1.2. Barriers to application

The given experiment provided comparative results on various building characterization methods. This disregards that their requirements and workflows differ strongly as introduced in Section 2.2.3.

As shown, the studied lumped parameters calibration approaches, i.e. the energy signature approach and the RC-modelling approach, do not require any model training. In comparison to BES calibration no surrogate model derivation is required, in comparison to the top-down approaches no sensor data from multiple buildings including building labels are required. Instead gray-box models can be calibrated for any building which has the right sensor data available.

All calibration approaches, i.e. both lumped parameter models and BES calibration, are highly dependent on the model to be calibrated. That building energy model can either be designed for each building individually, or archetype models can be derived if a large number of buildings is to be calibrated. In fact, segmenting a building stock into groups of similar buildings (archetype classification) and deriving a suitable building energy model (architecture characterization) are decisive steps in common calibration processes [42][43][66]. In the given case study, we used the same building energy model for calibration as we used for training. Hence, we fully omit the essential step of developing and characterizing an archetype suitable for the considered building stock, thus, biasing the results in favour of surrogate calibration.

That step of model derivation is avoided when using a supervised deep learning model. It learns features from building stock sensor data, that indicate a certain building characteristic. This process, however, is currently rarely possible as labelled building data sets are often not available for specific cities or districts. That label scarcity is a common problem in machine learning research and we encourage to leverage existing research from that domain.

Another limitation of supervised learning methods is that they do not necessarily generalize. For example, when a model is trained on data from residential buildings in Victoria, it may not characterize buildings in Chicago accurately. The best way to handle this issue is to continue to collect and publicize high-fidelity building data such that models are trained on more heterogeneous data sets.

Current literature often highlights that calibration approaches are attractive as they provide us with a modifiable physics-based model that can right away model the impact of retrofits on the overall building performance. This is not the case for top-down deep learning approaches. However, we propose that this argument should not prevent the field from exploring the use of deep machine learning models, as it might provide more accurate characterization results.

5.1.3. Methods to alleviate label-scarcity of building time series data sets

Supervised learning using labelled datasets has enabled great achievements in machine learning. For example, accuracy rates of 95% for classifying

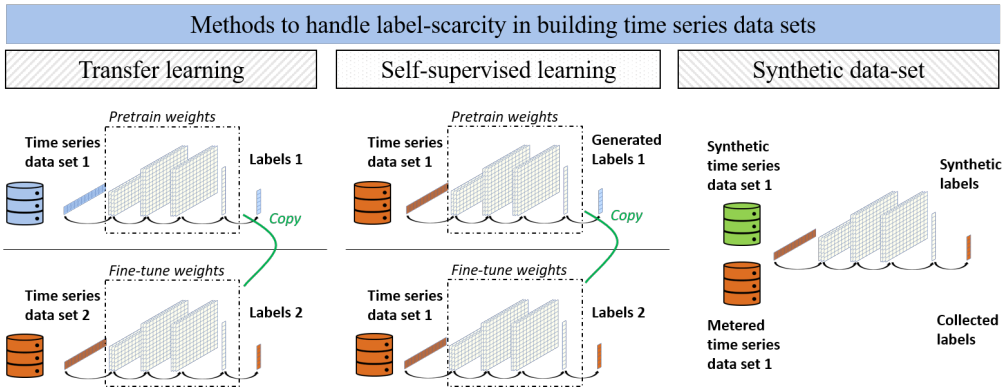


Figure 10: Caption

images contained in the Imagenet dataset are now possible . However, ImageNet consists of more than 14 million labelled images [67], which are often not available in supervision-starved areas like medical data or building time series data sets. Researchers have been tackling that problem by developing methods which receive similarly high accuracy with less and less data.

- **Transfer learning: From other labels to labels of real data:** Use existing building time series data sets with labels different to the one seeked to pretrain the model.

- **Self-supervised learning: From unlabelled real data to labelled real data** Another option is to pretrain neural networks on large unlabeled datasets and then fine-tune them on a smaller labeled dataset [68]. These methods are commonly called self-supervised learning methods, which aim at converting an unsupervised learning problem into a supervised one by creating automatically-derivable labels. So far the concept of self-supervised learning has barely been applied to time series problems, but it seems promising for the building domain.

- **Pretraining with labelled synthetic data sets.** This aims at pretraining a model using a dataset as provided in this study and subsequently refine it on a smaller real-world data set.

5.2. Synthetic data and real-world data

We used a synthetic data set to conduct controlled experiments on the robustness of building calibration methods. It allows to estimate the performance loss (increase in error) due to the four considered impact factors, but the actual errors will be higher in the case of real buildings. This has multiple reasons including that

- the heterogeneity of the synthetic building stock is small. We only considered one geometry with only one zone; we did not take surrounding buildings into account; only two climates were used and micro-climates were ignored; and the floors were assumed to be adiabatic.

- the quality of sensor measurements (outside air temperature, inside air temperature, heating system power and solar gains time series data in daily or 5-min intervals) is ideal. Here, we used the outputs from the BES tool as measurement data. No additional noise was added. In the real case, noisy data, missing recordings and dysfunctional sensors will necessarily lower the performance of all of the methods. Moreover, often climate data is often not available for a specific site but rather taken from a near-by weather station. This error is not considered.
  A great example on the quality of real world data is given by a recent Kaggle competition [69]. One of the reasons for winning the competition was large effort (including manual work) put into data cleaning.

Nonetheless, when comparing the synthetic and real world data it should be kept in mind that labelled building data sets, whose metadata contain labels on building characteristics like the HLC, are currently non-existent. By pointing out the two major differences of our synthetic data set and real world data, we aim at sparking future work. Many of the listed points, e.g. imposing noise on the synthetic data, can be addressed in future research and possibly let the characteristics of synthetic and real world data converge.

Furthermore, we suggest to augmenting the scope of the synthetic data set, such that the impact of more factors on characterization accuracy can be studied. They are listed in the following.

5.2.1. Future work: augmenting the synthetic data - based experiment

- **Values other than HLC:** Other building properties besides the heat loss coefficient are often wanted by building energy modellers. They may be continuous or discrete, e.g. the primary heating system of a building [8]. Here, we limited the study to the HLC, as the energy signature method allows to estimate it. The lumped parameter models are constrained in the type of characteristic to provide. BES calibration and supervised learning models are more versatile. In future work we would like to offer more labelled characteristics as part of the synthetic data set.

- **Other confounding factors:** We looked at the impact of climate, construction materials, air-infiltration, and stochastic occupant behaviour. We will extend that list, for example with the option to assess

the impact of versatile building geometries, of micro-climates, of the surroundings of a building, of ventilation strategies (incl. opening of windows) and others.

6. Conclusion

In this paper we benchmarked multiple methods to estimate quantitative building characteristics, here the heat loss coefficient, on a novel, extensible synthetic building meter data set.

The data set was used to conduct experiments assessing the impact of climate, building construction material, air-infiltration, and stochastic occupant behaviour on the performance of the methods. We could show both the lack of robustness of calibration-based methods towards these impact factors, and the practical shortcomings of more robust deep learning approaches. The latter is particularly caused by the lack of labelled building meter data sets.

We propose the experimental setup, i.e., a controlled environment of a synthetic, simulated data set, to further study the promising field of deep learning for automated building characterization. It is highly automated, less prone to errors due to mistakes of modellers, and can integrate large amounts of data for thousands of buildings for characteristics estimation of a specific building.

7. Acknowledgements

The project was supported by grant funding from CANARIE via the BESOS project (CANARIE RS-327) and the NSERC graduate scholarship.

References

[1] H. Schaffers, N. Komninos, M. Pallot, B. Trousse, M. Nilsson, A. Oliveira, Smart cities and the future internet: Towards cooperation frameworks for open innovation, in: The future internet assembly, Springer, Berlin, Heidelberg, 2011, pp. 431–446.

[2] C. Fan, F. Xiao, C. Yan, A framework for knowledge discovery in massive building automation data and its application in building diagnostics, Automation in Construction 50 (2015) 81–90.

[3] C. Miller, Z. Nagy, A. Schlueter, A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings, Renewable and Sustainable Energy Reviews 81 (2018) 1365–1377.

[4] Z. Nagy, D. Rossi, C. Hersberger, S. D. Irigoyen, C. Miller, A. Schlueter, Balancing envelope and heating system parameters for zero emissions retrofit using building sensor data, Applied Energy 131 (2014) 56–66. doi:10.1016/j.apenergy.2014.06.024.
URL http://www.sciencedirect.com/science/article/pii/S0306261914006060

[5] C. F. Reinhart, C. C. Davila, Urban building energy modeling–a review of a nascent field, Building and Environment 97 (2016) 196–202.

[6] Y. Wang, Q. Chen, T. Hong, C. Kang, Review of smart meter data analytics: Applications, methodologies, and challenges, IEEE Transactions on Smart Grid 10 (3) (2018) 3125–3148.

[7] Y. Wang, Q. Chen, D. Gan, J. Yang, D. S. Kirschen, C. Kang, Deep learning-based socio-demographic information identification from smart meter data, IEEE Transactions on Smart Grid 10 (3) (2018) 2593–2602.

[8] P. Westermann, C. Deb, A. Schlueter, R. Evins, Unsupervised learning of energy signatures to identify the heating system and building type using smart meter data, Applied Energy 264 (2020) 114715.

[9] S. D. Borgeson, Targeted efficiency: Using customer meter data to improve efficiency program outcomes, Ph.D. thesis, UC Berkeley (2013).

[10] D. Coakley, P. Raftery, M. Keane, A review of methods to match building energy simulation models to measured data, Renewable and Sustainable Energy Reviews 37 (2014) 123–141. doi:10.1016/j.rser.2014.05.007.
URL http://www.sciencedirect.com/science/article/pii/S1364032114003232

[11] J. D. Chambers, Developing a rapid, scalable method of thermal characterisation for uk dwellings using smart meter data, Ph.D. thesis, UCL (University College London) (2017).

[12] C. Ghiaus, Experimental estimation of building energy performance by robust regression, Energy and Buildings 38 (6) (2006) 582–587. doi:10.1016/j.enbuild.2005.08.014.
URL http://www.sciencedirect.com/science/article/pii/S0378778805001799

[13] P. Bacher, H. Madsen, Identifying suitable models for the heat dynamics of buildings, Energy and Buildings 43 (7) (2011) 1511–1522. doi:10.1016/j.enbuild.2011.02.005.
URL http://www.sciencedirect.com/science/article/pii/S0378778811000491

[14] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, number: 7553 Publisher: Nature Publishing Group. doi:10.1038/nature14539.
URL https://www.nature.com/articles/nature14539

[15] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.-A. Muller, Deep learning for time series classification: a review, Data Mining and Knowledge Discovery 33 (4) (2019) 917–963, arXiv: 1809.04356. doi:10.1007/s10618-019-00619-1.
URL http://arxiv.org/abs/1809.04356

[16] L. Breiman, et al., Statistical modeling: The two cultures (with comments and a rejoinder by the author), Statistical science 16 (3) (2001) 199–231.

[17] T. Hong, Z. Wang, X. Luo, W. Zhang, State-of-the-art on research and

applications of machine learning in the building life cycle, Energy and Buildings 212 (2020) 109831.

[18] C. Miller, F. Meggers, The building data genome project: An open, public data set from non-residential building electrical meters, Energy Procedia 122 (2017) 439–444.

[19] Ecobee, Donate your data, Ecobee Inc.
URL https://www.ecobee.com/donate-your-data/

[20] P. Street, Dataport: the world's largest energy data resource, Pecan Street Inc (2015).

[21] M. Senave, S. Roels, G. Reynders, S. Verbeke, D. Saelens, Assessment of data analysis methods to identify the heat loss coefficient from on-board monitoring data, Energy and Buildings 209 (2020) 109706. doi:10.1016/j.enbuild.2019.109706.
URL http://www.sciencedirect.com/science/article/pii/S0378778819319747

[22] P. Scully, Smart meter market report, Tech. rep., IOT Analytics (2019).

[23] J. Hiscock, Smart grid in canada 2014, Tech. rep., report# 2015-018 RP-ANU 411-SGPLAN, Natural Resources Canada, March 2015, 32 ... (2014).

[24] M. Frei, C. Deb, R. Stadler, Z. Nagy, A. Schlueter, Wireless sensor network for estimating building performance, Automation in Construction 111 (2020) 103043.

[25] A. Marston, E. Turner, A. Zakhor, O. Baumann, P. Haves, Testing rapmod: Can a portable scanner collect existing building data and create an energy model faster and more accurately than a human, eScholarship, University of California, 2015.

[26] D. Majcen, L. C. M. Itard, H. Visscher, Theoretical vs. actual energy consumption of labelled dwellings in the Netherlands: Discrepancies and policy implications, Energy Policy 54 (2013) 125–136. doi:10.1016/j.enpol.2012.11.008.
URL http://www.sciencedirect.com/science/article/pii/S0301421512009731

[27] A.-H. Deconinck, S. Roels, Comparison of characterisation methods determining the thermal resistance of building components from onsite measurements, Energy and Buildings 130 (2016) 309–320. doi:10.1016/j.enbuild.2016.08.061.
URL http://www.sciencedirect.com/science/article/pii/S0378778816307587

[28] M. Maasoumy, M. Razmara, M. Shahbakhti, A. S. Vincentelli, Handling model uncertainty in model predictive control for energy efficient buildings, Energy and Buildings 77 (2014) 377–392. doi:10.1016/j.enbuild.2014.03.057.
URL http://www.sciencedirect.com/science/article/pii/S0378778814002771

[29] K. Arendt, M. Jradi, H. R. Shaker, C. T. Veje, COMPARATIVE ANALYSIS OF WHITE-, GRAY- AND BLACK-BOX MODELS FOR

THERMAL SIMULATION OF INDOOR ENVIRONMENT: TEACHING BUILDING CASE STUDY (2018) 8.

[30] S. Prívara, J. Cigler, Z. Váňa, F. Oldewurtel, C. Sagerschnig, E. Žáčeková, Building modeling as a crucial part for building predictive control, Energy and Buildings 56 (2013) 8–22. doi:10.1016/j.enbuild.2012.10.024.
URL http://www.sciencedirect.com/science/article/pii/S0378778812005336

[31] Y. Heo, R. Choudhary, G. A. Augenbroe, Calibration of building energy models for retrofit analysis under uncertainty, Energy and Buildings 47 (2012) 550–560. doi:10.1016/j.enbuild.2011.12.029.
URL http://www.sciencedirect.com/science/article/pii/S037877881100644X

[32] P. Biddulph, V. Gori, C. A. Elwell, C. Scott, C. Rye, R. Lowe, T. Oreszczyn, Inferring the thermal resistance and effective thermal mass of a wall using frequent temperature and heat flux measurements, Energy and Buildings 78 (2014) 10–16. doi:10.1016/j.enbuild.2014.04.004.
URL http://www.sciencedirect.com/science/article/pii/S0378778814003041

[33] G. Nordström, H. Johnsson, S. Lidelöw, Using the Energy Signature Method to Estimate the Effective U-Value of Buildings, in: A. Hakansson, M. Höjer, R. J. Howlett, L. C. Jain (Eds.), Sustainability in Energy and Buildings, Smart Innovation, Systems and Tech-

nologies, Springer, Berlin, Heidelberg, 2013, pp. 35–44. doi:10.1007/978-3-642-36645-1_4.

[34] N. Pathak, J. Foulds, N. Roy, N. Banerjee, R. Robucci, A Bayesian Data Analytics Approach to Buildings' Thermal Parameter Estimation, in: Proceedings of the Tenth ACM International Conference on Future Energy Systems, e-Energy '19, Association for Computing Machinery, Phoenix, AZ, USA, 2019, pp. 89–99. doi:10.1145/3307772.3328316.
URL https://doi.org/10.1145/3307772.3328316

[35] V. Gori, P. Biddulph, C. A. Elwell, A Bayesian Dynamic Method to Estimate the Thermophysical Properties of Building Elements in All Seasons, Orientations and with Reduced Error, Energies 11 (4) (2018) 802. doi:10.3390/en11040802.
URL https://www.mdpi.com/1996-1073/11/4/802

[36] M. Lundin, S. Andersson, R. Östin, Development and validation of a method aimed at estimating building performance parameters, Energy and Buildings 36 (9) (2004) 905–914. doi:10.1016/j.enbuild.2004.02.005.
URL http://www.sciencedirect.com/science/article/pii/S0378778804001008

[37] S. S. Sablani, A. Kacimov, J. Perret, A. S. Mujumdar, A. Campo, Non-iterative estimation of heat transfer coefficients using artificial neural network models, International Journal of Heat and Mass Transfer 48 (3) (2005) 665–679. doi:10.1016/j.ijheatmasstransfer.2004.09.005.

URL http://www.sciencedirect.com/science/article/pii/S0017931004004065

[38] R. Singh, R. S. Bhoopal, S. Kumar, Prediction of effective thermal conductivity of moist porous materials using artificial neural network approach, Building and Environment 46 (12) (2011) 2603–2608. doi:10.1016/j.buildenv.2011.06.019.
URL http://www.sciencedirect.com/science/article/pii/S0360132311001934

[39] G. M. Baasch, R. Evins, Targeting Buildings for Energy Retrofit Using Recurrent Neural Networks with Multivariate Time Series, 2019.

[40] D. Crawley, L. Lawrie, F. Winkelmann, W. Buhl, Y. Huang, C. Pedersen, R. Strand, R. Liesen, D. Fisher, M. Witte, J. Glazer, EnergyPlus: Creating a New-Generation Building Energy Simulation Program, Energy and Buildings 33 (2001) 319–331. doi:10.1016/S0378-7788(00)00114-6.

[41] S. Nagpal, C. Mueller, A. Aijazi, C. Reinhart, A methodology for auto-calibrating urban building energy models using surrogate modeling techniques | Request PDF, Journal of Building Performance Simulation (Apr. 2018). doi:10.1080/19401493.2018.1457722.
URL https://www.researchgate.net/publication/324257575_A_methodology_for_auto-calibrating_urban_building_energy_models_using_surrogate_modeling_techniques

[42] J. Sokol, C. Cerezo Davila, C. F. Reinhart, Validation of a

Bayesian-based method for defining residential archetypes in urban building energy models, Energy and Buildings 134 (2017) 11–24. doi:10.1016/j.enbuild.2016.10.050.
URL http://www.sciencedirect.com/science/article/pii/S037877881631372X

[43] M. H. Kristensen, R. E. Hedegaard, S. Petersen, Hierarchical calibration of archetypes for urban building energy modeling, Energy and Buildings 175 (2018) 219–234. doi:10.1016/j.enbuild.2018.07.030.
URL http://www.sciencedirect.com/science/article/pii/S0378778818312532

[44] M. Gilli, E. Schumann, Calibrating Option Pricing Models with Heuristics, in: A. Brabazon, M. O'Neill, D. Maringer (Eds.), Natural Computing in Computational Finance: Volume 4, Studies in Computational Intelligence, Springer, Berlin, Heidelberg, 2012, pp. 9–37. doi:10.1007/978-3-642-23336-4_2.
URL https://doi.org/10.1007/978-3-642-23336-4_2

[45] G. Faure, T. Christiaanse, R. Evins, G. M. Baasch, BESOS: a Collaborative Building and Energy Simulation Platform, in: Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, BuildSys '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 350–351. doi:10.1145/3360322.3360995.
URL http://doi.org/10.1145/3360322.3360995

[46] A.-H. Deconinck, S. Roels, Is stochastic grey-box modelling suited

for physical properties estimation of building components from on-site measurements?, Journal of Building Physics 40 (5) (2017) 444–471. doi:10.1177/1744259116688384.
URL https://doi.org/10.1177/1744259116688384

[47] S. S. Garud, I. A. Karimi, M. Kraft, Design of computer experiments: A review, Computers & Chemical Engineering 106 (2017) 71–95.

[48] G. Reynders, J. Diriken, D. Saelens, Quality of grey-box models and identified parameters as function of the accuracy of input and observation signals, Energy and Buildings 82 (2014) 263–274. doi:10.1016/j.enbuild.2014.07.025.
URL http://www.sciencedirect.com/science/article/pii/S0378778814005623

[49] S. Hammarsten, A critical appraisal of energy-signature models, Applied Energy 26 (2) (1987) 97–110. doi:10.1016/0306-2619(87)90012-2.
URL http://www.sciencedirect.com/science/article/pii/0306261987900122

[50] P. Gianniou, C. Reinhart, D. Hsu, A. Heller, C. Rode, Estimation of temperature setpoints and heat transfer coefficients among residential buildings in Denmark based on smart meter data, Building and Environment 139 (2018) 125–133. doi:10.1016/j.buildenv.2018.05.016.
URL http://www.sciencedirect.com/science/article/pii/S0360132318302762

[51] S. Danov, J. Carbonell, J. Cipriano, J. Martí-Herrero, Approaches to

evaluate building energy performance from daily consumption data considering dynamic and solar gain effects, Energy and Buildings 57 (2013) 110–118. doi:10.1016/j.enbuild.2012.10.050.
URL http://www.sciencedirect.com/science/article/pii/S0378778812005841

[52] M. Brøgger, P. Bacher, K. B. Wittchen, A hybrid modelling method for improving estimates of the average energy-saving potential of a building stock, Energy and Buildings 199 (2019) 287–296. doi:10.1016/j.enbuild.2019.06.054.
URL http://www.sciencedirect.com/science/article/pii/S0378778819300398

[53] H. Madsen, J. Holst, Estimation of continuous-time models for the heat dynamics of a building, Energy and Buildings 22 (1) (1995) 67–79. doi:10.1016/0378-7788(94)00904-X.
URL http://www.sciencedirect.com/science/article/pii/037877889400904X

[54] M. Manfren, N. Aste, R. Moshksar, Calibration and uncertainty analysis for computer models–a meta-model based approach for integrated building energy simulation, Applied energy 103 (2013) 627–641.

[55] P. Westermann, R. Evins, Surrogate modelling for sustainable building design–a review, Energy and Buildings 198 (2019) 170–186.

[56] M. Wetter, E. Polak, A convergent optimization method using pattern search algorithms with adaptive precision simulation, Building Services

Engineering Research and Technology 25 (4) (2004) 327–338, publisher: SAGE Publications Ltd STM. doi:10.1191/0143624404bt097oa.
URL https://doi.org/10.1191/0143624404bt097oa

[57] K. Deb, S. Agrawal, A. Pratap, T. Meyarivan, A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii, in: International conference on parallel problem solving from nature, Springer, 2000, pp. 849–858.

[58] M. C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (3) (2001) 425–464, _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9868.00294. doi:10.1111/1467-9868.00294.
URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00294

[59] M. Farah, P. Birrell, S. Conti, D. D. Angelis, Bayesian Emulation and Calibration of a Dynamic Epidemic Model for A/H1N1 Influenza, Journal of the American Statistical Association 109 (508) (2014) 1398–1411, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2014.934453. doi:10.1080/01621459.2014.934453.
URL https://doi.org/10.1080/01621459.2014.934453

[60] A. M. Rysanek, J. A. Fonseca, A. Schlueter, Bayesian calibration of a building energy model by stochastic optimisation of root-mean square error, Working Paper, ETH Zurich, accepted: 2019-06-26T13:44:18Z

(Jun. 2019). doi:10.3929/ethz-b-000349836.
URL https://www.research-collection.ethz.ch/handle/20.500.11850/349836

[61] P. R. Miles, R. C. Smith, Parameter estimation using the python package pymcmcstat (2019).

[62] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Gated Feedback Recurrent Neural Networks, arXiv:1502.02367 [cs, stat]ArXiv: 1502.02367 (Feb. 2015).
URL http://arxiv.org/abs/1502.02367

[63] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Comput. 9 (8) (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.
URL http://dx.doi.org/10.1162/neco.1997.9.8.1735

[64] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, arXiv:1412.3555 [cs]ArXiv: 1412.3555 (Dec. 2014).
URL http://arxiv.org/abs/1412.3555

[65] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, arXiv:1512.03385 [cs]ArXiv: 1512.03385 (Dec. 2015).
URL http://arxiv.org/abs/1512.03385

[66] F. Johari, G. Peronato, P. Sadeghian, X. Zhao, J. Widén, Urban building energy modeling: State of the art and future prospects, Renewable and Sustainable Energy Reviews 128 (2020) 109902.

[67] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[68] C. Doersch, A. Gupta, A. A. Efros, Unsupervised visual representation learning by context prediction, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1422–1430.

[69] C. Miller, P. Arjunan, A. Kathirgamanathan, C. Fu, J. Roth, J. Y. Park, C. Balbach, K. Gowri, Z. Nagy, A. Fontanini, et al., The ashrae great energy predictor iii competition: Overview and results, arXiv preprint arXiv:2007.06933 (2020).

8. Appendix A

$$C\frac{dT_{in}}{dt}(t) = \dot{Q}_{int}(t) + \dot{Q}_{hsys}(t) + \dot{Q}_{sol}(t) + \dot{Q}_{env}(t) + \dot{Q}_{inf}(t) \tag{5}$$

Equations 6 and 7 express $\dot{Q}_{env}$ and $\dot{Q}_{inf}$ in terms of the difference between external and internal temperature.

$$\dot{Q}_{env}(t) = \frac{1}{R}(T_{ext}(t) - T_{in}(t)) \tag{6}$$

where R is the thermal resistance of the building envelope [K/W], T_{ext} is the external temperature and T_{in} is the internal temperature.

$$\dot{Q}_{inf}(t) = \dot{m} * c_{p,air}(T_{ext}(t) - T_{in}(t)) \tag{7}$$

where $\dot{m}$ is the air mass flow rate [UNITS] and $c_{p,air}$ is the air specific heat capacity [UNITS]. Equation 5 can thus be rewritten as:

$$C\frac{dT_{in}}{dt}(t) = \dot{Q}_{int}(t) + \dot{Q}_{hsys}(t) + \dot{Q}_{sol}(t) + HLC_{wb}(T_{ext} - T_{in}) \tag{8}$$

$$HLC_{wb} = HLC_{inf} + HLC_{env} \tag{9}$$

where $HLC_{inf} = \dot{m} * c_{p,air}$ and $HLC_{env} = \frac{1}{R}$. HLC_{wb} is the whole-building heat loss coefficient. By rearranging the thermal energy balance in this way we can see that it depends on both the infiltration rate and the thermal resistivity of the building envelope.

9. Appendix B

HLC_{inf} is the product of the air mass flow rate, $\dot{m}$, and the air specific heat capacity, $c_{p,air}$. The air mass flow rate was calculated directly by EnergyPlus and recorded as a time series output.[14] The mean yearly value of this output variable was multiplied by the specific heat capacity for air to calculate HLC_{inf}.

Note that for all the cases in which infiltration was 0, HLC_{inf} was also 0. The calculation for HLC_{env} is considerably more complicated. It can be calculated using an analogy to RC circuit model, where the thermal resistances of the building envelope are analogous to resistors in a circuit. The building

[14]The EnergyPlus output variable is called: Zone Infiltration Current Density Volume Flow Rate

envelope can be represented by three resistors in series: (1) the interior surface resistance, R_{int}, (2) the resistance of the material layers, R_{mat}, and (3) the exterior surface resistance, R_{ext}:

$$HLC_{env} = (R_{int} + R_{mat} + R_{ext})^{-1} \tag{10}$$

R_{int}, R_{mat} and R_{ext} represent the respective resistances of all the building surfaces in parallel. For instance, R_{int} represents the parallel resistances for each individual indoor surface. These values can therefore be found by taking the sum of the reciprocals of the resistances of each surface, as seen in equation 11. The resistances of each surface are reported by EnergyPlus, but the models outputs do not account for area. Therefore, the reciprocals of the resistances are multiplied by their associated surface areas as follows:

$$\frac{1}{R_i} = \sum_{s \in S} A_s * \frac{1}{R_{i:s}} \tag{11}$$

where $i \in \{\text{int, mat, ext}\}$, S is the set of all surfaces, $R_{i:s}$ is the resistance of the surface and A_s is the area of the surface.

The values for $R_{i:s}$ are calculate differently for the three resistance types. $R_{mat:s}$ is output directly by EnergyPlus. The calculation for R_{int} and R_{ext} requires the evaluation of time-resolved heat transfer coefficients (HTCs), measured in W/m^2 °K.HTCs are proportionality constants that dictate the given amount of heat exchange by convective and radiative forces at a building surface. Each HTC can be viewed as the inverse of a resistance. Each HTC at a given surface acts in parallel, so $h_{total} = h_1 + h_2 + ... + h_n$. In EnergyPlus, the equations for heat exchange due to convection and radiation

depend directly on HTCs so

$$R_{y:s} = 1/mean(h_{y:s:conv} + h_{y:s:rad}) \tag{12}$$

where $y \in \{\text{int, ext}\}$, $h_{y:s:conv}$ is the surface convective HTC and $h_{y:s:rad}$ is the surface radiative HTC.

The remaining calculation considerations for each of the three R values are summarized below:

1. $R_{int:s}$: The internal surface radiation HTCs calculated by EnergyPlus are modelled by the software internally and are not easily accessible to the user (see the EnergyPlus documentation[15] for more information). Therefore, for the purpose of this study, only convection is included in the calculation for $R_{int:s}$. The exclusion of the radiative HTCs may result in a slightly larger absolute errors in HLC estimation, but it should not affect the comparisons between methods or the parametric analysis within the methods.
2. $R_{mat:s}$: The material R value is the sum of the resistance of the material layers that compose the surface. It is calculated directly by EnergyPlus and is reported as the surface U-value, or $1/R_{mat:s}$, in [W/m^{2}K].
3. $R_{ext:s}$: As described by the EnergyPlus documentation,[16] at the external surfaces convection and radiation to the ground, air and sky are modelled by EnergyPlus and the associated HTCs are directly available

[15] https://bigladdersoftware.com/epx/docs/9-2/engineering-reference/inside-heat-balance.html

[16] https://bigladdersoftware.com/epx/docs/9-2/engineering-reference/outside-surface-heat-balance.html#outside-surface-heat-balance

to the user as time-resolved output variables. These output variables are used to find $R_{ext:s}$.

10. Appendix C

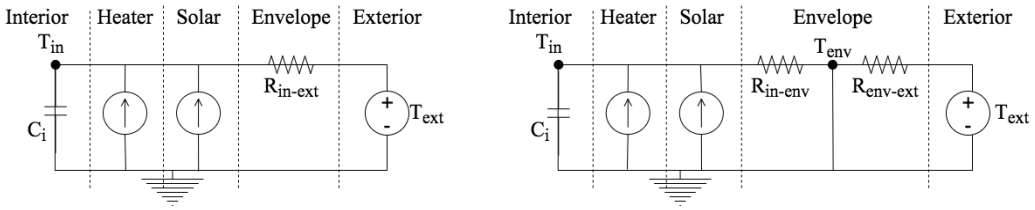


Figure 11: RC Network models, as presented by Bacher and Madsen [13]. The left model is a 1st order representation of the building envelope with a single lumped capacitance, while the right model has two lumped capacitances.

1^{st} Order Model:

$$dT_i = \left(\frac{1}{C_i R_{ia}}(T_a - T_i) + \frac{1}{C_i}(A_w \phi_s) + \frac{1}{C_i}(\phi_h)\right) dt + \sigma_1 dw_1 \tag{13}$$

2^{nd} Order Model:

$$dT_i = \left(\frac{1}{C_i R_{ie}}(T_e - T_i) + \frac{1}{C_i}(A_w \phi_s) + \frac{1}{C_i}(\phi_h)\right) dt + \sigma_1 dw_1 \tag{14}$$

$$dT_e = \left(\frac{1}{C_e R_{ie}}(T_i - T_e) + \frac{1}{C_e R_{ea}}(Ta - Te)\right) dt + \sigma_e dw_e \tag{15}$$

Epilogue

The given paper contributes with a collection seven different approaches to extract quantitative building properties. Furthermore, we applied them to a synthetic dataset where we have full knowledge on the buildings for which we try to find a calibrated model.

The two surrogate-based calibration methods showed high performance in comparison to the other five approaches. However, these findings currently do not generalize to real world data and future work is required. In particular, we had full knowledge on a suitable parametric BPS model to train our surrogate model on. In reality this model has to be found prior to calibration (see Chapter 6) [25][15] which introduces large uncertainty.

Chapter 7

Thesis conclusion

This thesis was inspired by the vision of fast, interactive machine learning based surrogate models to support architects and engineers in finding sustainable building designs. The core idea of surrogate models is to be trained on physics-based simulation results and subsequently approximate building energy performance estimates almost instantaneously. This creates an interactive environment for end users to explore the energy performance of a large space of design alternatives.

The goal was to lay the technical foundations for a large scale application of surrogate models in our domain. As such we relied on rapidly accumulating knowledge in the machine learning world and transferred the most promising elements to our domain. This involved the use of Bayesian deep learning models and deep temporal convolutional neural networks as surrogate models.

As a result, we provide tools which allow us to train uncertainty-aware surrogate models, which can be applied over a large range of climates. They can be embedded into web-platforms to be widely accessible. Furthermore, by using calibration techniques they can link simulation models with the physical world. This allows the use

of surrogate model for the design of new buildings and for the retrofit of the already built environment.

Our collection of work may be considered as a starting point for future work. We propose two major research paths, i.e. to empirically study the interaction between surrogate models and building designers, and to further increase the scope of surrogate models such that they can be used for more design problems without retraining:

Surrogate models in practice: Our studies were motivated by the first success with interactive early design tools [1][24][13][20]. Similarly, our group, the Energy in Cities group at the University of Victoria, is soon hosting surrogate models on a web-platform. In dedicated survey, we will be able to collect empirical data on the interactive design sessions, which may help to assess the use of our tools, i.e. uncertainty-aware surrogate models and location-independent surrogate models, and beyond that point us towards further technical needs.

Generalizing surrogates: Combined with our work, it is possible to derive surrogate models for simulations of buildings with varying geometry [7] and located at varying locations. Other unpublished work showed that also the impact of the built environment (e.g. wind channelling effects, or shading effects) can be modelled with fast machine learning models. A key element of future research will be to integrate all these approaches into one model. Also, we foresee that in future surrogate models will be able to generalize over various occupancy load profiles (by using the same approach we used for climate-independent surrogates) and detailed mechanical systems. First studies have been initiated. A high degree of generalization may also enable a higher automation of surrogate-based calibration.

Real world data-based calibration: To date, we have only applied surrogate-based calibration on a synthetic data set. Other authors already managed to use them for a larger set of buildings [25], however, they only assessed the predictive performance of the models instead of the accuracy of the building parameter estimates. This is essential for retrofit assessment and proposes to widen our benchmarking study to real world data sets.

Bibliography

[1] Nathan C. Brown. Design performance and designer preference in an interactive, data-driven conceptual building design scenario. *Design Studies*, page , 2020.

[2] Daniel Coakley, Paul Raftery, and Marcus Keane. A review of methods to match building energy simulation models to measured data. *Renewable and sustainable energy reviews*, 37:123–141, 2014.

[3] Drury B Crawley, Linda K Lawrie, Frederick C Winkelmann, Walter F Buhl, Y Joe Huang, Curtis O Pedersen, Richard K Strand, Richard J Liesen, Daniel E Fisher, Michael J Witte, et al. Energyplus: creating a new-generation building energy simulation program. *Energy and buildings*, 33(4):319–331, 2001.

[4] Pieter De Wilde. The gap between predicted and measured energy performance of buildings: A framework for investigation. *Automation in construction*, 41:40–49, 2014.

[5] Ralph Evins. A review of computational optimisation methods applied to sustainable building design. *Renewable and Sustainable Energy Reviews*, 22:230–245, 2013.

[6] Sushant S Garud, Iftekhar A Karimi, and Markus Kraft. Design of computer experiments: A review. *Computers & Chemical Engineering*, 106:71–95, 2017.

[7] Philipp Geyer and Sundaravelpandian Singaravel. Component-based building performance prediction using systems engineering and machine learning. *Applied Energy*, 228:1439–1453, 2017.

[8] Elisabeth Gratia and André De Herde. A simple design tool for the thermal study of an office building. *Energy and buildings*, 34(3):279–289, 2002.

[9] Rob Guglielmetti, Dan Macumber, and Nicholas Long. Openstudio: an open source integrated analysis platform. In *Proceedings of the 12th conference of international building performance simulation association*, 2011.

[10] Jan LM Hensen and Roberto Lamberts. Building performance simulation–challenges and opportunities. *Building Performance Simulation for Design and Operation*, pages 1–10, 2019.

[11] Jennifer Hiscock. Smart grid in canada 2014. Technical report, report# 2015-018 RP-ANU 411-SGPLAN, Natural Resources Canada, March 2015, 32 âĂę, 2014.

[12] Internation Energy Agency. Tracking buildings. Technical report, Internation Energy Agency, 2020.

[13] Thomas Jusselme. *Data-driven method for low-carbon building design at early stages*. PhD thesis, EPF Lausanne, 2020.

[14] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[15] Martin Heine Kristensen, Rasmus Elbæk Hedegaard, and Steffen Petersen. Hierarchical calibration of archetypes for urban building energy modeling. *Energy and Buildings*, 175:219–234, 2018.

[16] Clayton Miller and Forrest Meggers. Mining electrical meter data to predict principal building use, performance class, and operations strategy for hundreds of non-residential buildings. *Energy and Buildings*, 156:360–373, 2017.

[17] Clayton Miller, Zoltán Nagy, and Arno Schlueter. A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. *Renewable and Sustainable Energy Reviews*, 81:1365–1377, 2018.

[18] Open Technologies. The building pathfinder. Online.

[19] Torben Østergård, Rasmus L Jensen, and Steffen E Maagaard. Building simulations supporting decision making in early design–a review. *Renewable and Sustainable Energy Reviews*, 61:187–201, 2016.

[20] Torben Ostergard, Rasmus L. Jensen, and Steffen E. Maagaard. Early building design: Informed decision-making by exploring multidimensional design space using sensitivity analysis. *Energy and Buildings*, 142:8–22, 2017.

[21] Torben Ostergard, Rasmus Lund Jensen, and Steffen Enersen Maagaard. A comparison of six metamodeling techniques applied to building performance simulations. *Applied Energy*, 211:89–103, 2018.

[22] June Young Park, Xiya Yang, Clayton Miller, Pandarasamy Arjunan, and Zoltan Nagy. Apples or oranges? identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset. *Applied Energy*, 236:1280–1295, 2019.

[23] Steffen Petersen. *Simulation-based support for integrated design of new low-energy office buildings*. DTU Civil Engineering, Technical University of Denmark, 2011.

[24] Fabian Ritter, Philipp Geyer, and AndrÃľ Borrmann. Simulation-based decision-making in early design stages. In *32nd CIB W78 conference, Eindhoven, The Netherlands*, pages 27–29, 2015.

[25] Julia Sokol, Carlos Cerezo Davila, and Christoph F Reinhart. Validation of a bayesian-based method for defining residential archetypes in urban building energy models. *Energy and Buildings*, 134:11–24, 2017.

[26] Liesje Van Gelder, Payel Das, Hans Janssen, and Staf Roels. Comparative study of metamodelling techniques in building energy simulation: Guidelines for practitioners. *Simulation Modelling Practice and Theory*, 49:245–257, 2014.

[27] Michael Wetter and Jonathan Wright. A comparison of deterministic and probabilistic optimization algorithms for nonsmooth simulation-based optimization. *Building and Environment*, 39(8):989–999, 2004.

Appendix

While surrogate modelling is the core topic of this PhD thesis, my studies involved a diverse set of research on applying machine learning methods to analyse retrofit performance and to improve building operation using predictive models. A lot of that work involved the supervision of Bachelor's and Master's students leading to a set of conference publications which are given below.

Building performance prediction

Paper 1

The first paper tackles the use of rule-based algorithms to provide building occupants indoor condition predictions including explanations. It can serve as a basis for an application for virtual assistant device to interact with building occupants.

BS 2019 2-4 SEPT ROME 16th IBPSA INTERNATIONAL CONFERENCE AND EXHIBITION

INTERNATIONAL BUILDING PERFORMANCE SIMULATION ASSOCIATION

Insight Into Predictive models: On The Joint Use Of Clustering And Classification By Association (CBA) On Building Time Series

Paul Westermann, Joel Grieco, Johanna Braun, Eamon Murphy, Ralph Evins[1]
[1]Energy Systems and Sustainable Cities group,
Department of Civil Engineering, University of Victoria, Canada

Abstract

Data-driven, black box machine learning models have received a lot of attention in the field of building control. They have been used successfully to predict building behaviour given information like weather forecasts and real time sensor information. In these models, the occupant behaviour is considered to act exogenously on the building.
We consider the users as active elements of the building operation control loop. To make educated control decisions they have to be informed about how the building will behave. Therefore, we propose a prediction model which explains to occupants the day-ahead building behaviour using a clustering and classification by association model. We benchmark this approach to a neural network regression model and only observed a small loss of accuracy.
Knowing the upcoming building behaviour, occupants can adjust their behaviour (e.g. putting on clothes) or the building systems settings (e.g. set points) accordingly. The proposed method is a promising way to decode complex regression models into readable rules, which in future may be useful in conjunction with for example voice-based virtual assistants.

Introduction

Buildings are a major energy consumer accounting for 36% of final energy and 55% of final electricity consumption worldwide (IEA, 2017). 80 to 90% of that energy is attributed to building operation (Ramesh et al., 2010). Therefore, optimizing building operation through effective energy management is a strong element of current research on sustainable buildings (Shaikh et al., 2014).
Building occupants have a major impact and partially explain why high performing building technologies (e.g. efficient HVAC systems) do not guarantee low energy use (Andersen et al., 2009). In a simulation-based study on office buildings, a difference in energy use of up to 50% is found if the worker is proactive in energy savings or not (Lin and Hong, 2013). Behavioural differences are found in their adaptive actions (e.g. opening/closing of windows, adjusting set-points) or non-adaptive actions (operation of office equipment, movement through space, etc.) (Hong et al., 2017). This shows that engaging occupants in the energy efficient control of the building will be crucial to achieving energy use targets.
Researchers have developed tools which incorporate occupancy data as input into *supervisory* building control algorithms. Supervisory control logic is implemented at a higher level than the individual controllers of the building systems. Two approaches are prevailing in research: rule-based, and model-predictive control. While rule-based control uses rules defined by HVAC specialists, MPC conducts an operational optimisation over a specified prediction horizon. In both approaches temperature set-points for the whole building are adjusted, or HVAC systems activated taking occupant actions (adapative or non-adaptive) into account. The occupant behaviour is either hard-coded in schedules or detected based on data (Lu et al., 2010). Detection of occupancy patterns (e.g. sleeping, or absent) is a key element of smart thermostat technologies which already exist.[1]
A characteristic of rule-based and model-predictive control is that they *monitor* human behaviour instead of involving occupants as sensing and active element in the control loop (*direct* human-in-the-loop control, HIL). Recent publications envision an interplay of occupants and automated controls where comfort conditions are traded-off with minimizing energy use (D'Oca et al., 2018). This negotiation of comfort conditions demands not only machines to learn occupancy patterns, but also occupants to understand the computer controlling the building.
This study contributes by providing a forecasting method which features a human-readable set of information to explain the expected building behaviour given the computer-based controls already existing in the building. We use a combination of clustering and associate rule mining. Cluster analysis enables to find typical 24-hour temperature profiles and

[1]See for example: https://nest.com/thermostats/nest-learning-thermostat/overview/

https://doi.org/10.26868/25222708.2019.211236

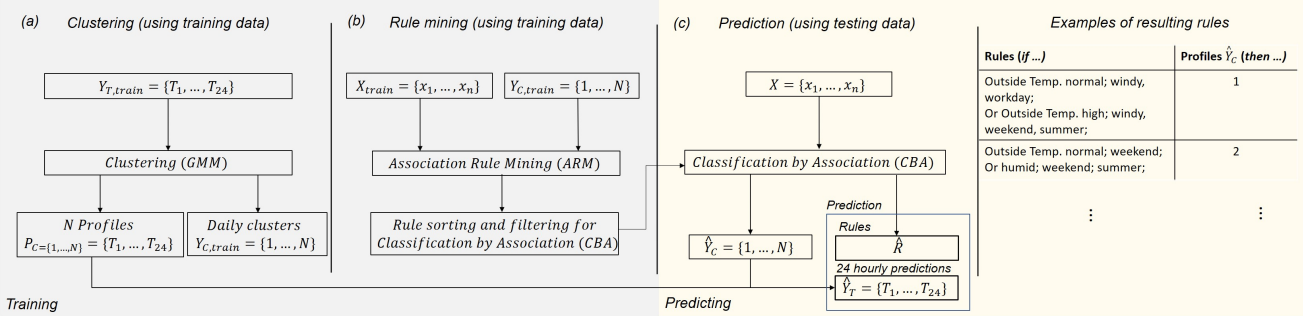


Figure 1: Overview of the proposed approach.

associate rule mining allows to assign a set of rules connecting each of the profiles to weather conditions and occupancy. Based on those rules we select a 24-h profile for the upcoming day with the classification by association (CBA) algorithm. As a result, the occupant has access to numerical building behaviour predictions which are explained by human-readable association rules in the form of "*the predicted profile is x because y*" .

The combination of clustering and association rule mining has been leveraged on building time series data before. Mirebrahim et al. (2017) and Xiao and Fan (2014) used it to receive insight on the control of heating, ventilation and air conditioning (HVAC) systems. Both cases exemplify the strength of the approach for analytical purposes, however it has never been used for forecasting of building time series.

We showcase the use of the method in a study where we derive 24-h indoor temperature forecasts for the upcoming day. Indoor temperature was chosen as it inherently captures the trade-off between occupant comfort and energy demand.

The use of a set of temperature profiles and of categorical features (e.g. binned outdoor air temperature) instead of continuous ones for rule-based prediction limits model complexity. We benchmarked our approach against a 24h prediction of a deep multiple output feed-forward neural network.

In this paper we familiarize the reader with the applied method and provide details on the clustering algorithm used (Gaussian Mixture Modelling), associate rule mining and the classification by association algorithm. Then, the performance and limits of the approach are shown in a case study on indoor temperature prediction in an office building.

Methodology

The proposed approach combines clustering (Fig. 1, a) and rule-mining (Fig. 1, b) to give insightful time series predictions which provide numerical forecasts as well as explanatory rules causing that forecast (Fig. 1, c). The method can be applied to any time series data which is formatted as daily sets of 24 hourly values. In the case study below we focussed on indoor temperature forecasting only, hence the model outputs ($\hat{Y}_T$) are labelled T.

The methodology consists of two steps to train the model:

1. Derive N typical daily profiles using a Gaussian Mixture Model (GMM). The number of profiles has to be chosen by the modeller and is treated as a hyperparameter to be optimized in a grid search (see Table 2).
   Clustering converts hourly output values $Y_{T,train}$ to daily ones $Y_{C,train}$ which contain the derived cluster numbers for each day of the training data.

2. The prediction model, a CBA model, uses association rules for the cluster number $Y_{C,train}$ given features X_{train}. In our case study, the n number of features include daily mean weather forecast data, date-time information (incl. holidays), occupancy data and the cluster of the previous day.
   To derive the CBA model, we first generate association rules between X_{train} and $Y_{C,train}$ using the Apriori algorithm (Agrawal et al., 1994). Then the number of rules is reduced to a small set which only includes those rules with the highest confidence. The high confidence rules form the CBA model.

After this model training process is terminated, the CBA model can be used to predict hourly indoor temperatures for the upcoming day given a new set of unseen features X (Fig. 1, c). It uses the profiles (cluster centroids) and rules determined on the training data. Note that, in the following prediction performance is quantified solely by comparing predicted hourly values, $\hat{Y}_{T,test}$, to observed hourly values, $Y_{T,test}$. We fully neglect whether clusters are predicted correctly.

In the sections below we provide more details on the two steps to derive the prediction model.

Clustering (Gaussian Mixture Model)

The GMM is suitable for clustering problems. It has been applied to time series data before (Eirola and Lendasse, 2013) and specifically on building re-

lated time series data (Melzi et al., 2017) (Mirebrahim et al., 2017). It is a classification algorithm which describes a cluster by its mean and covariance. Both are composed of a mixture of Gaussian distributions. This allows it to identify inhomogeneous, multimodal clusters as required for time series profile clustering of temperature data. In comparison to the k-means or hierarchical clustering, GMM is a soft clustering algorithm, i.e. individual samples influence the centroids of all clusters and not only the one they belong to (a comparison of both approaches is found in Park et al., 2019). Soft clustering may be suitable for the given problem as indoor temperatures are inherently continuous and cannot be sorted into discrete, separable bands. Comparing and picking the best performing clustering algorithm is not within the scope of this study but would be valuable future work.
The output of the GMM is a probability density function $P_k(x)$ for each of the clusters $k \in K$ given a set of features X. The density functions consist of a linear combination of multiple Gaussian distributions $N(x; \mu_{kr}, \Sigma)$ (Hastie et al., 2009).

$$P_k(X) = \sum_r \pi_{kr} N(X; \mu_{kr}, \Sigma) \quad (1)$$

Here all clusters share the same covariance matrix Σ. The optimum value of all parameters, i.e. the mean of each Gaussian distribution, the mixing proportion $\pi_{k,r}$ for each of the R Gaussian distributions and covariance matrix Σ are chosen by maximising the log-likelihood

$$\sum_k^K \sum_{g_i=k} log \left[\sum_{r=1}^{R_k} \pi_{kr} N(x_i; \mu_{kr}, \Sigma) \prod_k \right] \quad (2)$$

of all clusters $k \in K$ simultaneously, where $\prod_k$ represents the clusters prior probability. The cluster with the highest probability given a set of parameters x is the one proposed by the GMM. Fitting the GMM is done using the expectation-maximisation (EM) algorithm (Dempster et al., 1977).

Before the GMM is fitted to the data, the number of clusters is picked manually. The common way is to use information criteria like BIC or AIC which enable to qualitatively compare accuracy of models with different number of clusters. In our case, we optimized the number of clusters to maximize predictive accuracy of the whole approach in Figure 1.

Model derivation

Association rule mining

Like GMM, association rule mining (ARM) is an unsupervised learning technique that identifies interesting relationships between features and targets (Jirı and Kliegr, 2012). It was initially applied to market basket analysis for the identification of simple rules to understand consumer behaviour.
First, the discretized features X and targets Y_C are stored in a transactional database. The transactional database is scanned for association rules using one of the existing ARM algorithms (here: Apriori algorithm, Agrawal et al., 1994). The quality of a rule is quantified by calculating support and confidence of each rule described in the following equations.

$$supp(A) = |t \in T; A \subseteq t| \,/\, |T| \quad (3)$$

$$conf(A \Rightarrow B) = supp(A \cup B)/supp(A) \quad (4)$$

Let A be a feature set, A ⇒ B an association rule and T a set of transactions of a given database. Support captures how likely it is that A and B occur jointly ($P(A, B)$) while the confidence provides a value for how likely the occurrence of B is if A is given. A minimum value for support is used to place a limit on the number of rules.
For classification purposes, the rule mining algorithm is adjusted to restrict the consequent B to only contain the target variable Y_C. Ma and Liu (1998) formulated the framework for creating association rules in this manner, naming them class association rules (CARs). A predictive classification model is created by a subset of CARs which are picked using Classification by Association (CBA).

Classification by Association (CBA)

CBA is a supervised machine learning algorithm which stands out due to its simplicity. It takes CARs as inputs, sorts them and outputs a subset of useful rules that can classify sets of features. As outlined by Ma and Liu (1998), to derive the CBA model, CARs are sorted by the confidence, then support, and then the order the rules are generated in. Each entry of the training data is covered by at least one rule.
The CARs derivation, sorting and deleting of rules is conducted based on training data and therefore may be regarded as model training. Afterwards, the remaining rules can be applied to unlabelled data picking the first rule within the list of sorted rules that is satisfied by a given set of new features.
The rules picked by CBA are a useful output in themselves, because they provide a human readable list of the most predictive features for target selection. Ma and Liu (1998) describe this as the discovery of understandable rules. The CBA framework can provide more understandable and more predictive rules than association rule mining alone. In addition to the prediction of targets on unseen data the outputted rule set can assist in achieving the human readable functionality desired in many applications.
In this study we used the pyFIM and PyARC libraries for ARM and CBA implementation (Borgelt, 2012)(Jirı and Kliegr, 2012).

Case Study

The methodology is applied to predict indoor temperatures in a small room ($\approx 10m^2$, one worker, one window) of an office building in British Columbia.

Table 1: Overview on the dataset split into target and features.

Type	Sensor name	properties
Target Y_T	Indoor air temperature [°C]	hourly mean (15 min. data)
Features X:	outdoor air temperature forecast* [°C]	daily mean on-site measured data
	wind chill* [°C]	daily mean, on-site measured data
	heat index* [°C]	daily mean, on-site measured data
	relative humidity* [%]	daily mean, on-site measured data
	occupancy [%]	daily mean, on-site measured data
	lagged profile number []	profile number from previous day (predicted by GMM)
	date - time []	day of the week, month, season

*discretized by equal frequency binning.

The indoor climate of the room is controlled by a trickle vent and slab heating or cooling. The trickle vent preheats or cools fresh air using a coil. Both systems are connected to a central heat pump.

Data and Feature selection

In the proposed approach the selection of input features is crucial as they form the rules shown to occupants to understand temperature predictions. For now, we limit the set to only a small selection of features, constrained by data availability and quality.
The considered data set spans three years (2014-2017). It consists of measured values on the building systems and the internal and external climate conditions. The data is not public but information on the building are publicly available.[2] The data from multiple sensors was cleaned and aligned to a frequency of one hour (Y_T) or one day (X).[3] All continuous features are discretized into bins with equal numbers of samples. Besides the listed features, we also had access to temperature set point (occupant input) data of the room which was constant over the whole period and therefore ignored.
Among the features in Table 1, we selected a subset based on an exhaustive grid search (see next section). In future applications, more occupant inputs (adaptive actions, see Section 1), building system data and sensor data of adjacent rooms might be important to ensure useful explanations for forecasts.

Model derivation

The model was trained on two years of data (Nov. 2014 to Nov. 2016) and tested on the following year. As the CBA algorithm ranks rules based on support and confidence values derived on the training data, it is crucial that the training data consists of the same number of samples from each season. Otherwise, the support (Eq. 3) for rules of an underrepresented season will be relatively low in comparison to rules of other seasons. Similarly, the confidence of rules (Eq. 4) would be skewed.
In Figure 2, the resulting seven temperature profiles generated with Gaussian Mixture modelling are shown. All results in this section were derived using the optimized number of clusters, bin size and set of features (see Table 2). The profiles may be sorted from hot to cold and by differences in shape. Two profiles are rather flat with low average temperature. The other five profiles fluctuate strongly between day and night and the temperature is warmer on average.
Next we apply associate rule mining and extract the classification by association rules (CARs). We receive distinct explanations for each cluster (see Fig. 3). The rules in Fig. 3 show the three rules with the highest support value for each cluster. Some clusters have less than three rules in which case all of the associated rules are shown.
The most days (highest support) in the training data

[2] `https://www.reliablecontrols.com/corporate/facility/`

[3] Instances of sensor outages were found at various points in the data set. Days with one or more missing indoor temperature values are ignored leading to a loss of 12.9% of samples. In future, measurement gaps could be filled with rolling mean values.

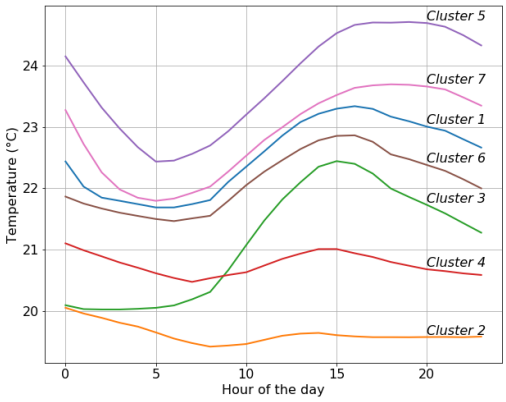


Figure 2: Temperature profiles for each of the seven clusters.

Consequent (then...)	Antecedent (if...)	confidence	support
Cluster 1	day-of-week=Tuesday,Previous_day=0.0	0.556	0.009
	season=Fall,day-of-week=Tuesday,Wind_Chill=high	0.556	0.009
Cluster 2	day-of-week=Sunday,month=12	1.000	0.013
	day-of-week=Sunday,Previous_day=Cluster 4,Heat_Index=low	0.600	0.011
	oat_mean=very low,Occupancy=False,month=12	0.600	0.011
Cluster 3	Occupancy=True,day-of-week=Monday,season=Winter	0.400	0.015
	Occupancy=True,day-of-week=Monday,quarter=4,Previous_day=Cluster 2	0.750	0.011
Cluster 4	Previous_day=Cluster 6,Occupancy=False,Heat_Index=low	0.750	0.022
	Heat_Index=low,day-of-week=Saturday	0.857	0.022
	day-of-week=Sunday,Occupancy=False,quarter=1	0.588	0.018
Cluster 5	Wind_Chill=very high,Previous_day=Cluster 5,Occupancy=False,season=Summer	1.000	0.015
	Previous_day=Cluster 5,OA_RH=low	0.667	0.011
	day-of-week=Monday,month=6	0.556	0.009
Cluster 6	quarter=1,Occupancy=True	0.785	0.133
	Wind_Chill=medium,oat_mean=medium	0.737	0.102
	Occupancy=True,season=Spring,Wind_Chill=medium	0.800	0.051
Cluster 7	quarter=3	0.664	0.165
	Heat_Index=very high,quarter=3	0.795	0.113
	Heat_Index=very high,Occupancy=True,season=Summer	0.753	0.100

Figure 3: Top 3 rules for classification of each cluster (sorted by support).

are members of *Cluster 6* and *7*. *Cluster 6* represents occupied days during winter (quarter 1) and shoulder season (spring) with medium outside air temperature and wind chill, and *Cluster 7* is the typical profile for occupied days in summer (quarter 3, Season = Summer). *Cluster 4* and *2* show the profiles for unoccupied days. During unoccupied days in winter the temperature typically drops to below $20°C$. *Cluster 3* has a very distinct shape. It captures the reheating process after unoccupied days in winter which typically occurs on Mondays. *Cluster 5* describes overheating inside the room. The rules show that this happens on days when wind chill is high meaning high ambient temperature and low wind speeds. The strong impact of wind speed is due to the fact that the room features trickle vents which rely on natural ventilation for cooling. Lastly, *Cluster 1* has very low support values. This is surprising as it lies between the two most common clusters. The reason may be that the control routine of the heating and cooling system leads to indoor profiles very close to *Cluster 7* OR *Cluster 6* and nothing in between.
Finally, we apply the derived model to unseen data and compare the results to the observed indoor temperature profiles. Model derivation and testing was conducted iteratively in an exhaustive grid search with the number of clusters, the number of bins for variable discretization and the selection of features as hyperparameters. To speed up the process the features were grouped into four sets (Table 1). The optimal parameter settings are shown in Table 2. Especially, the use of clusters of the previous day increased the accuracy significantly. They were derived using the mixture model trained on the training data.

Table 2: Results of grid search.

Hyper-parameter	Range	Final choice
No. of clusters	[1,15]	7
Bin size	[2,10]	5
Feature subsets	[Date time],[Lagged Clusters],[Weather], [Workday], [Occupancy]	[Date time],[Lagged Clusters], [Weather], [Occupancy]

Model validation

Testing the method on unseen data gives a Mean Absolute Error (MAE) of $0.558°C$ and 62.5% of the variation in the indoor temperature is explained ($R^2 = 0.625$). Figure 4 shows the characteristics of cluster based prediction with a cap at high temperatures and floor at low temperatures. Furthermore, due to the discrete classification of profiles the predictions exhibit a gap between $19.7°C$ and $20.3°C$. To better understand the performance and the causes of inaccuracies, we decomposed the inaccuracies and benchmarked our algorithm to two different applications of neural networks.
In a first step, the loss of variance caused by using

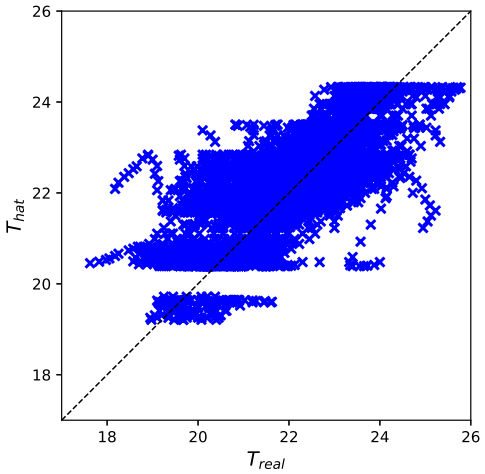


Figure 4: Observations vs. predictions on test data.

daily temperature profile clusters instead of predicting each hourly temperature value individually is shown in Figure 5. Using seven clusters, which was determined to be optimal by the grid search, a maximum R^2 of 0.79 is theoretically achievable if all clusters are predicted correctly. Hence, there is a 21% loss in theoretically explainable variance by the process of converting continuous hourly target values to seven discrete daily clusters.
Another simplification of the prediction process is the use of association rules instead of a complex statistical regression model. To quantify the loss of accuracy induced by rule based prediction, we conducted the cluster prediction with a parameterized black-box classifier. Here, we use a feed-forward neural network classifier whose parameters were again optimized in a grid search. It outperformed the CBA algorithm only by a little ($R^2 = 0.665$).
After having decomposed the loss of accuracy, we benchmarked the algorithm against a state-of-the-art deep neural network regressor which predicts 24 temperature values individually for each day. The regressor is fed with the same set of inputs as before while having 24 temperature outputs. The network is composed of three layers with 200 neurons each and was pruned by increasing the regulation term α (Hastie et al., 2009) step-by-step until optimum performance was achieved.[4] The accuracy is much higher (R^2=0.815) than the proposed cluster- and rule-based approach but with loss of explainability. Also it shows that given the current set of features the neural network fails to explain 19.5% of the variance. Probably, more features on occupants and other unknowns may be helpful to further increase accuracy.

[4]The process of *pruning* refers to gradually increasing the regularization term until variance and bias of the neural network are balanced.

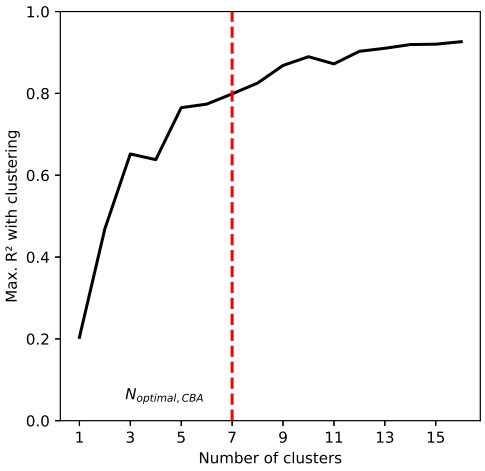


Figure 5: Maximum achievable R^2 score for a given number of clusters.

Model application

The functionality of the proposed algorithm is shown in Fig. 6. One week of indoor temperature predictions for each season is shown. The weeks were selected randomly among the weeks without missing data. The rules which caused the CBA algorithm to predict one of the seven profiles for the upcoming day are shown below each of the predicted 24h temperature profiles (split by black lines). For example, on 8th April 2017 the model predicts *Cluster 2* because the previous day was *Cluster 6*, the heat index is medium, and the building was unoccupied.

Generally, we find that the cluster- and rule-based prediction is capable of capturing the indoor temperature behaviour well. Weekends are identified and depending on weather conditions different profiles are picked during the week (see winter week). However, we also see that *Cluster 6* in winter and *Cluster 7* in summer are classified on most days. Rarely, a significant misclassification of a day can be observed as for example found on 22nd October 2017.

The dominance of two clusters is explainable due to the impact of the heating and cooling system, and due to the fact that the temperature set point was never changed by the worker in the training and testing data. As a consequence, our classifier mainly distinguishes between the seasons and between days where the HVAC system is switched on and those when it is switched off.

Misclassification may be caused by ambiguous information provided by the features. On 22nd October 2017, the classifier predicts the building to be heated but instead the heating system was switched off as it is Sunday. On that day the occupancy sensor recorded some activity in the room. This triggers the CBA algorithm to predict the wrong cluster, because in this specific case occupancy-based rules have higher confidence than rules which consider that it is a Sunday and the room should be unheated. A similar misclassification is observed on the 19th March which was also a Sunday.

Discussion

The case study showed that the proposed method is convenient to apply. Once a pipeline of clustering and rule-mining is established, it generates forecasts alongside of comprehensible sets of rules. In Fig. 6 a maximum of four variables per rule were generated which seems suitable for rapid forecast analysis.

The data available for the case study lacks information on occupant action. The rules like *it will be hot (Cluster 5) because wind chill is high and the building is occupied* (see Fig. 3), do not recommend any occupant action.[5] In further applications, the data should be complemented with behavioural features. For example, if an occupant knows that *it will be hot because wind chill is high, the building is occupied and windows are closed*, he or she will open the window to increase comfort.

Model parameters and model performance considerations

The number of clusters is the only model parameter of the GMM which was optimized. Its covariance matrix, another parameter of the GMM, was set to be full, i.e. each cluster has a different, full covariance matrix. A brief study showed that this is better than all other choices of covariance matrix type (all clusters sharing the same matrix or the matrices may only have diagonal elements).

The rule mining process has four modelling parameters, i.e. minimum support, minimum confidence, the bin size of the variable discretization and the involved features. We included the latter two into the hyper-parameter optimization process. Minimum confidence was removed (set to zero) and minimum support set to five days. This ensures that any derived rule is found at least five times in the data.

The accuracy of the model is significantly lower than 24h predictions of a deep neural network as shown in Table 3. However, one could argue that a loss of $0.19°C$ in MAE may be acceptable if the method helps occupants to improve energy efficiency of the building by adjusting their behaviour. This trade-off in loss of accuracy and improved occupant behaviour has yet to be studied in a field test.

The prediction accuracy of the model can be improved by deriving better rules to predict more

[5] High wind chill index refers to high ambient temperatures and low wind speeds.

Table 3: Model validation and benchmarking for 24h predictions.

Error Type	GMM + CBA	GMM + ANN_C	ANN_{Reg}
MAE $[°C]$	0.558	0.548	0.37
R^2	0.625	0.665	0.815

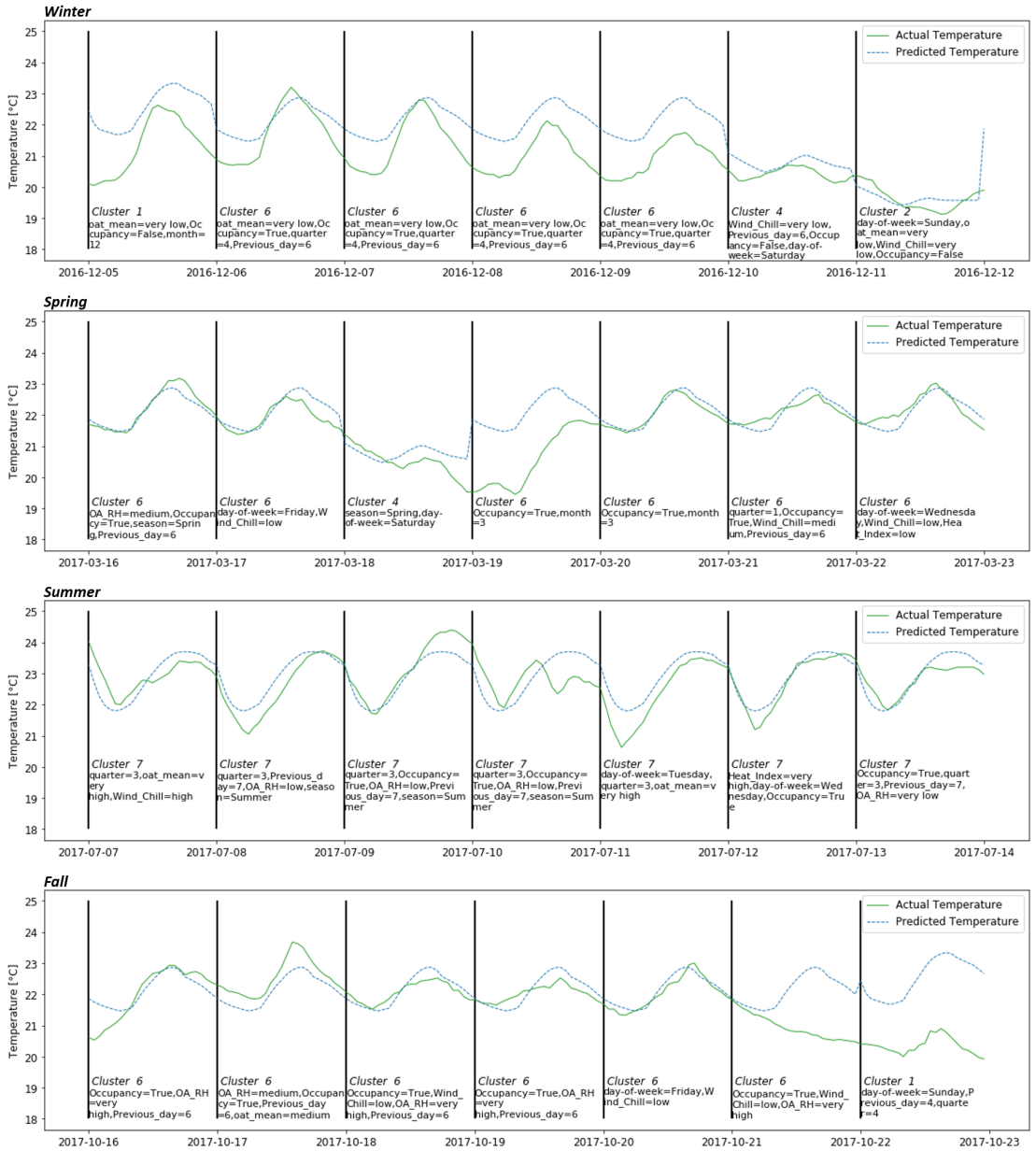


Figure 6: Predictions and associated rules for one week of each season in the test data.

clusters accurately (see Figure 5). For example, by using ten clusters the maximum achievable R^2 score would increase from $R^2_7 = 0.79$ to $R^2_{10} = 0.88$. With the current set of features and the resulting rules, we determined seven clusters to optimal. Our rule set is not explanatory enough to accurately predict more clusters. If more or better features are found, more clusters could be accurately predicted.

The benchmarking analysis showed that our method with the current way of feature engineering does not fully leverage the information hidden in the data. A neural network achieved much higher accuracy given the same set of information. More work on feature engineering could be done, but also it may be concluded that an increase of explainability leads to a loss in accuracy.

Conclusions and Future Work

This study introduced and benchmarked a novel approach to provide hourly forecasts on building behaviour for the upcoming day. It combines the analytical power of unsupervised machine learning (clustering, associate-rule mining) with the prediction ability of supervised machine learning methods given by the CBA algorithm. As a result each forecast is complemented with rule-based explanations why a certain forecast was given. This would enable occupants to adapt and adjust their actions. In future, we imagine the method could help to involve occupants in the building control loop which may lead to an increase in building energy efficiency.

After having benchmarked the accuracy of the

method against *black-box* models, the next step is to conduct a second case study where rules are provided to actual occupants of a building. This could be done by implementing the forecasting method on an intelligent personal assistant device to communicate the explanations and recommendations associated with temperature or energy consumption forecasts. This will clarify if influencing occupant actions can increase overall building efficiency.

Acknowledgements

We thank Reliable Controls Corporation for providing the data and NSERC for funding the research work.

References

Agrawal, R., R. Srikant, et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, Volume 1215, pp. 487–499.

Andersen, R. V., J. Toftum, K. K. Andersen, and B. W. Olesen (2009). Survey of occupant behaviour and control of indoor environment in danish dwellings. *Energy and Buildings 41*(1), 11–16.

Borgelt, C. (2012). Frequent item set mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2*(6), 437–456.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

D'Oca, S., T. Hong, and J. Langevin (2018). The human dimensions of energy use in buildings: A review. *Renewable and Sustainable Energy Reviews 81*, 731–742.

Eirola, E. and A. Lendasse (2013). Gaussian mixture models for time series modelling, forecasting, and interpolation. In A. Tucker, F. Hppner, A. Siebes, and S. Swift (Eds.), *Advances in Intelligent Data Analysis XII*, Lecture Notes in Computer Science, pp. 162–173. Springer Berlin Heidelberg.

Hastie, T., R. Tibshirani, and J. Friedman (2009). The elements of statistical learning: data mining, inference, and prediction.

Hong, T., D. Yan, S. D'Oca, and C.-f. Chen (2017). Ten questions concerning occupant behavior in buildings: The big picture. *Building and Environment 114*, 518–530.

IEA (2017). Energy technology perspectives. Technical report, International Energy Agency.

Jirı, F. and T. Kliegr (2012). Classification based on associations (cba)-a performance analysis.

Lin, H.-W. and T. Hong (2013). On variations of space-heating energy use in office buildings. *Applied Energy 111*, 515–528.

Lu, J., T. Sookoor, V. Srinivasan, G. Gao, B. Holben, J. Stankovic, E. Field, and K. Whitehouse (2010). The smart thermostat: using occupancy sensors to save energy in homes. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, pp. 211–224. ACM.

Ma, B. L. W. H. Y. and B. Liu (1998). Integrating classification and association rule mining. In *Proceedings of the fourth international conference on knowledge discovery and data mining*.

Melzi, F. N., A. Same, M. H. Zayani, and L. Oukhellou (2017). A dedicated mixture model for clustering smart meter data: Identification and analysis of electricity consumption behaviors. *10*(10), 1446.

Mirebrahim, S. H., M. Shokoohi-Yekta, U. Kurup, T. Welfonder, and M. Shah (2017). A clustering-based rule-mining approach for monitoring long-term energy use and understanding system behavior. In *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments*, pp. 5. ACM.

Park, J. Y., X. Yang, C. Miller, P. Arjunan, and Z. Nagy (2019). Apples or oranges? identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset. *Applied Energy 236*, 1280–1295.

Ramesh, T., R. Prakash, and K. Shukla (2010). Life cycle energy analysis of buildings: An overview. *Energy and buildings 42*(10), 1592–1600.

Shaikh, P. H., N. B. M. Nor, P. Nallagownden, I. Elamvazuthi, and T. Ibrahim (2014). A review on optimized control systems for building energy and comfort management of smart sustainable buildings. *Renewable and Sustainable Energy Reviews 34*, 409–429.

Xiao, F. and C. Fan (2014). Data mining in building automation system for improving building operational performance. *Energy and buildings 75*, 109–118.

Paper 2

The second paper, was written in collaboration with SES Consulting Inc.[1] One of their clients was facing large peak pricing costs due to their cooling devices. A predictive model was implemented to identify days with high cooling loads and the potential cost abatement was analysed.

[1] http://sesconsulting.com/

eSim 2018
«BUILDING SIMULATION TO SUPPORT BUILDING SUSTAINABILITY»

IBPSA Canada
The Canadian Chapter of the International Building Performance Simulation Association
Le chapitre canadien de International Building Performance Simulation Association

Machine Learning Recommendations for Control of Complex Building Systems Using Weather Forecasts

Paul Westermann[1], Nigel David[2], Ralph Evins[1]
[1]Energy Systems and Sustainable Cities group, University of Victoria
[2]SES Consulting Inc., Victoria

Abstract: *We present a machine learning model used to provide recommendations on chiller operation based on the prediction of cooling demand using a weather forecast. A long short term memory (LSTM) formulation was used, and achieved favourable results compared to a standard approach. The model captured the data to a reasonable extent (R^2 = 0.70), but was unable to predict very high loads at unexpected times. The model is intended to be used as an aid to a human operator, not as a replacement, and it is likely that many of these unexpected events could be overridden by the operator. Overall, the predictive model reduced the number of occasions in which a chiller was operating unnecessarily by 80.5%, or 469 hours. This demonstrates the power of data-driven predictive control to assist in the efficient operation of complex building systems, saving money, energy and operator time.*

Keywords: *Energy management system, human-in-the-loop control, Machine Learning*

Introduction

Non-residential, commercial and institutional buildings consume large amounts of energy. In Canada, they account for more than 10% of the end-use energy consumption (NRCan, 2017). Furthermore, in 2009 their energy intensity exceeded the ones of average Canadian households by almost 40% (NRCan, 2012).
Due to the sheer size of the buildings a share of 20% of energy in the total operating costs depicts a large amount of absolute energy payments. Hence, a strong driver to implement building retrofit measures exists. Regarding the complexity of heating and cooling systems of those large buildings, it is common that a specialised energy manager supervises the operation of all systems. Based on many factors like weather, occupant behaviour and other disturbances he decides which systems are switched on and how the temperature set-points are selected.

Recent advances in software development as well as increasing amounts of available data are promoting the development of methods to support or even automate this human-based energy management of complex building systems. One approach which received a lot of attention in building control is model predictive control (MPC) as shown by Oldewurtel et al. (2012). MPC uses a physical model to determine optimal system inputs for the upcoming hours or days. At each hour, the system inputs are optimised following certain objectives (e.g. lowering cost or energy consumption) subject to human comfort constraints. The approach takes inputs like weather conditions, electricity prices and occupancy patterns into account. However, wide application is yet to come, especially, because deriving a physical model of the building is work intensive. It requires to transfer all architectural and building system information of into a model. Furthermore, the modeller has to make assumptions about typical building operation patterns which may lead to modelling errors.

Hand in hand with the wide spread rise of machine learning (ML), building scientists have explored ways to exploit sensor data in buildings to train models capable of predicting future building behaviour and its energy consumption. Accurate forecasts can be used in a similar fashion as proposed by MPC, i.e. to optimise heating and cooling system inputs. Wei et al. (2018) explain the fundamentals and give a broad review on existing data driven models. A good example on the application of the most common ML model, artificial neural network (ANN), is given by Jetcheva et al. (2014), who generated an ensemble of multiple ANN models. Massana et al. (2015) address which kind of data is required to forecast building energy demand accurately. Especially, they highlight the importance of occupancy data.

Both physical model-based and ML-based optimisation of the building energy consumption struggle with the stochastics laying in building energy consumption due to multiple factors like weather forecast, or occupant behaviour uncertainty. One pathway in MPC research is the application of stochastic MPC Oldewurtel et al. (2010), where uncertainties and constraint violations are considered probabilistically to allow acceptable levels of risk in optimal control. In data driven research, methods which provide probabilistic forecasts, e.g. Gaussian

ISBN 978-2-921145-88-6

Processes, attract a lot of attention (Gray and Schmidt, 2016). Another option is the use of human domain knowledge as it is suggested for example in the field of crowd sourcing where human intelligence is used in cases when machine intelligence is less effective (Kamar et al., 2012). In case of buildings, one could exploit the knowledge of an on-site energy manager to remove some uncertainty the ML model fails to take into account. As an example, an energy manager of an office building may know of extraordinary events like meetings or conferences and thus, would know that heating demand will be higher than predicted by a data-driven model. Incidents like this are not straight forward to incorporate into a ML model.

We complement the existing literature with an application of ML based human-in-the-loop control. We use ML to generate point forecasts of the energy demand of a building and give recommendations to an energy manager on-site to switch individual systems on and off. The goal is to reduce the number of hours, when more systems run than required, which causes overall efficiency losses. We aim for a solution that avoids complex feature acquisition and extraction but only uses weather forecasts provided by an API as well as time information. Therefore, our approach offers a plug-and-play solution being highly generalizable as the model features do not need to be customized to changing available sensor data on different buildings.

Case study building

This study uses ML to optimise the operation of a complex cooling system of a large building in British Columbia, Canada. The system is equipped with three chillers to guarantee comfort within the building. Currently, the control of the cooling system is determined by an energy manager who decides on the use of the three different chillers. The chillers are all of different types with different efficiencies as shown in Table 1 and Figure 2. A simple operation routine for the three chillers is chosen by the building operator:[1]

$$\begin{aligned} 0t \leq D < 1100t \quad & \textit{Base operation}\text{: only chiller A} \\ 1100t \leq D < 3000t \quad & \textit{Co-operation}\text{: chiller A \& B} \\ 3000t \leq D \quad & \text{All chillers} \end{aligned} \tag{1}$$

As chiller C is supposed to act as a back-up chiller, which is switched on only in rare events like maintenance, outages of other chillers or heat waves, its operation is excluded in the following study. The given routine represents the most efficient operation strategy if the co-operation of both chillers runs optimally (see Figure 2). At the threshold of $1100t$ the co-operation becomes more efficient than the use of chiller A only. The optimisation of the simultaneous operation of both chillers is not part of this study, but we focus on the prediction of whether the *co-operation* of both chillers is necessary.

[1]The unit used is refrigeration ton t (also: RT). It originates from refrigeration with natural ice. $1t$ corresponds to the refrigeration supplied by melting one short ton of water over 24 hours(Avallone et al., 2006).

Name	System	Properties
Ch. A	Centrifugal chiller with variable speed drive (VSD) and condenser relief	*most efficient, good for changing loads*
Ch. B	Centrifugal chiller with with condenser relief	*less efficient worse for changing loads*
Ch. C	Back-up chiller (absorption chiller)	*least efficient, back-up or during heat waves*

Table 1: Overview of chiller systems of case study building.

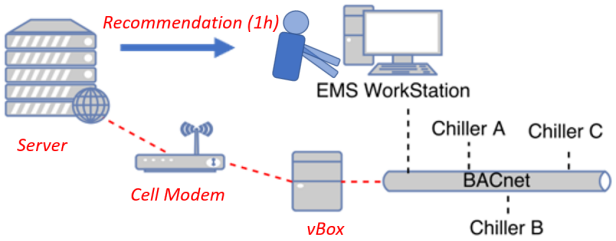


Figure 1: System architecture as introduced in Evins (2017).

Based on the considered data set (8760 hours), we observe that the actual operation did not always follow the presented routine which may have lead to significantly higher operating costs and energy use. It is found that both chillers run during 1998 hours however only during 1484 both chillers were actually needed, i.e. in 583 hours two chillers were running although only one chiller was required. Furthermore, ”cold” start of the second chiller occurred, leading to demand peaks which generates significant demand charges for the building operator (up to CAD$ 14k were reported).

The scope of the problem is therefore using demand forecasts to give recommendations to the energy manager to reduce

1. the number of hours of unnecessary *co-operation*,
2. peak demands.

The latter can be avoided by slow-starting the second chiller prior to the peak demand event (if the high peak demand is forecasted by the model) or by shifting the additional starting load to low demand hours (*load shifting*).

Human-in-the-loop control

Instead of developing a fully independent and automated control algorithm, this study proposes the use of ML based recommendations to support the building energy manager. Based on the recommendation, he decides to either switch

ISBN 978-2-921145-88-6

a chiller on or off. Once the system is switched on, it is controlled autonomously. As the energy manager fundamentally takes part in that control scheme we classify it as a human-in-the-loop approach.

The architecture of the intended control loop is shown in Figure 1. The ML model training and predictions are conducted on an external server connected to the internet. The server is also used to collect energy demand data and information on which chillers are running. The data is gathered via a mini computer running Volttron software, which is designed to access the BACnet of the building and retrieve data.[2] The transmission of the data happens over a cell modem.
Based on the information of which chiller is running the ML model recommends each hour of the day if chiller A or chiller B needs to be switched on or off. In case the building energy manager takes action, he accesses the energy management system (EMS) to control the individual chiller units. For practicability reasons this could also be done using wearables like a smart watch.

Having a human in a control loop, who is adjusting supervisory control parameters, is well understood (Stankovic, 2014). In those applications, the control loop is running autonomously with a human only intervening when it is necessary. An example for human-in-the-loop control in the building context may be found in Mirebrahim et al. (2017).

It has to be noted, that the provided solution is appropriate for large commercial and institutional buildings like hospitals, malls, offices and others. Due to the complexity of the heating and cooling system as well as the large uncertainty introduced by human activity, an interplay of machine intelligence and an energy manager seems promising in this case. In comparison, in buildings with low uncertainty in demand(e.g. data centres) pure machine learning methods have proven great success in lowering the energy demand. The company DeepMind reported a reduction of 40% in cooling demand using ML control.[3]

Energy demand forecasting

Methodologies in energy demand forecasting have a broad application including the optimisation of power systems or regarding buildings, either early building design optimisation or system control optimisation as found in this paper.
In this study we derived three models and compare them. In addition to two neural network models, a feed-forward neural network and a long short term memory (LSTM) model, we also included a common linear regression model. The models were programmed using Keras (Chollet et al., 2015) and Tensorflow (Abadi et al., 2015).
All models map weather data as well as time information (inputs) to cooling demand (outputs).[4] We received hourly chiller use data for 1 year (Dec. 2015 to Nov. 2016). For the same period weather data is available online via an API (www.vancouver.weatherstats.com) or was computed (sun position). Due to availability, we used weather measurements instead of forecast data. Thus, some uncertainty given by the inputs is ignored in this study and needs to be addressed in future. The heat map in Figure 3 shows the correlation of all used weather inputs to cooling demand.
The non-recursive structure of our models enables to choose the prediction horizon. For each input of hourly weather data the corresponding hourly cooling demand is given as an output of the model. Furthermore, no sensor data on the building state (e.g. temperature, internal gains, etc.) or on chiller A and B is used for predictions.

Traditional feed-forward neural networks are one of the most common ML algorithms. Each network consists of multiple layers with multiple cells (neurons). In comparison to LSTM networks they do not feature the capability to change an internal state depending on previous model outputs. Nonetheless, they have shown to be suitable for building energy demand predictions. Wei et al. (2018) give more explanations and a review on existing studies.

LSTM neural networks (Hochreiter and Schmidhuber, 1997) are part of the group of recurrent neural networks (RNN). RNN received a lot of attention in the time-series forecasting domain because they can store previous model outcomes. RNN are usually trained using back-propagation algorithm in real-time recurrent learning. Those algorithms are prone to vanishing or exploding gradients. The LSTM algorithm overcomes this problem and thus, it is popular in the ML world. A LSTM network consists of multiple cells whose outputs are looped back to the cell. Current model outputs can be stored in the cell via an internal state. This internal state is overwritten, erased or read at each model evaluation step.
LSTM cells can be stacked in a multi-layer architecture. Here, the final model architecture was determined by using a simple grid search. We explored different number of cells per layer (10 or 50), different number of layers (1,2) and different sizes of training batches (50 hours to 320 hours). The latter is specifically interesting as it determines on which size of individual batches the LSTM model is trained. The impact of the batch size had the most significant impact on the training performance. The final model consists of two layers with 50 cells each and was trained during 200 epochs on batches of 200 hours each.

[2] https://volttron.org/
[3] https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/
[4] For the specific building we also provide information if and to which extent chiller C is used to control for special events like maintenance.

ISBN 978-2-921145-88-6

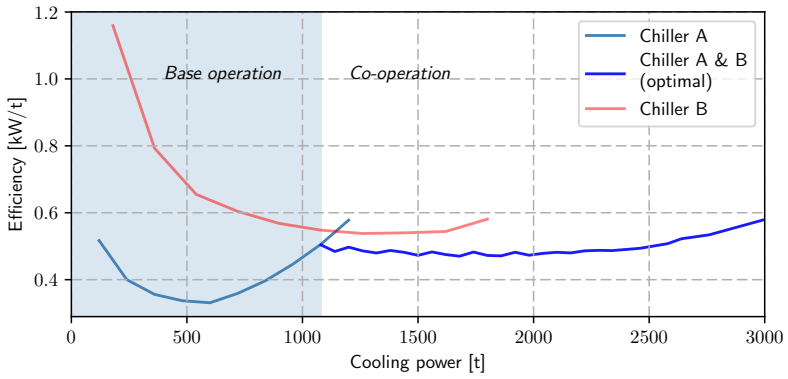


Figure 2: Overview of efficiency of Chiller A and B as a function of cooling power. Efficiency is quantified by the amount of electricity required to provide one refrigeration ton [kW/t].

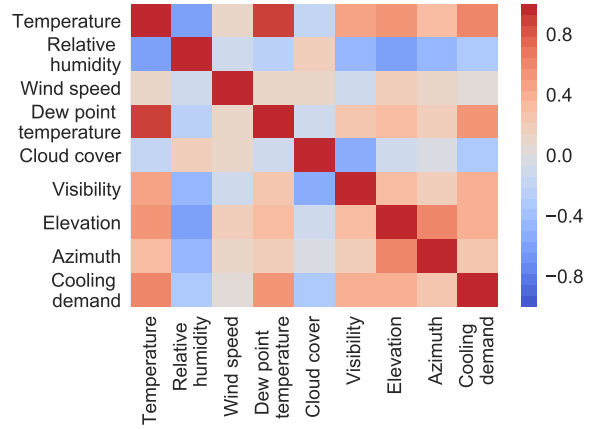


Figure 3: Heat map of correlation between weather inputs and cooling demand.

To generate outcomes with the model, the whole set of previous inputs has to be considered as the model predictions depend on the internal state of the LSTM, which depends on previous model outputs.

Results

The models were trained on a sequential block of data (70% of the data) and tested on the last block of the data set(30%). Last block testing is necessary for time-series data to receive an accurate estimation of the model performance (Bergmeir and Benitez, 2012). The training set consists of data from December to August and the test set from August to the end of November. The period of the test set covers late summer, autumn and the beginning of winter as such a large variety of cooling operation is included. Figure 4 shows that in the beginning of the test set consistently a high demand for cooling whereas later only occasional peaks occur.

	LR	**ANN**	**LSTM**
R^2_{train}	0.55	0.72	0.72
R^2_{test}	0.43	0.64	0.70

Table 2: Performance of applied models. The one of the LSTM surpassed the ones of the LR and ANN model. It is the most accurate one with the lowest degree of overfitting. The mean absolute error of the LSTM predictions on the test set is 24.2%.

The LSTM model surpassed the performance of all other models (see Table 2). The linear regression model had the worst performance which confirms that there are non-linearities in the relationship of cooling demand and weather data. The feed-forward network structure enables to capture any non-linearity and performed significantly better than a linear model. The best-performing (based on test data performance) neural network of the candidates in our grid search results in some overfitting as performance on the test data is significantly worse than on the training data. Similar overfitting behaviour was observed in the linear regression model. The LSTM model only showed weak overfitting. It seems that a network with memory for previous outputs is specifically well suited to our problem as no building state data (e.g. temperature) is given as an input to the model.

The performance of the LSTM model on the test data is shown in Figure 4. We find that in the beginning the model follows the real values well. However, at the end of the data two demand spikes occur which the model does not capture well. Looking into the data, it was observed that weather data does not indicate those peaks, i.e. temperature and solar gains do not show any abnormally high values. The model failure in those cases shows that eventually further input data on internal heat gains of the building would be required unless the building manager can compensate these inaccu-

ISBN 978-2-921145-88-6

racies. Overall, a mean absolute error of 24.2% was found for the LSTM model on the test data. If the demand peaks are not taken into account this value would be significantly better.

Another perspective on the model behaviour is given in Figure 5 where the real data is on the X-axis and predictions on the Y-axis. All perfect predictions lay on the 45° line. We see the strong correlation of predictions to the real data. However, two distinct areas are determined where the model struggles to give accurate predictions (see red dashed ellipses). The vertical ellipse shows the case when cooling demand was low and the chillers were close to idling operation, but the model predicts significant cooling. The horizontal ellipse indicates the opposite, i.e. the model predicts idling chiller operation but in reality, significant cooling was supplied. Again, this behaviour may be caused by a lack of information on occupant behaviour in the building. Either internal gains are higher than expected (horizontal ellipse) or lower than expected (vertical ellipse). Furthermore, a change in user comfort (i.e. changing temperature set points in multiple rooms) may cause modelling errors.

Recommendations for chiller operation

The predictions were processed to recommend if either chiller A or both chillers need to be switched on. Here, we programmed this process corresponding to the control routine defined above: any prediction of a cooling demand above 1100t means that the model recommends *co-operation* of both chillers. The resulting recommendations are shown in Figure 6 and emphasize that a ML approach to the problem is very suitable.

The three columns in the Figure represent which operation would be optimal according to the defined control routine (left), which operation was observed in the data (center) and which operation our prediction model suggests. We used boolean operators to filter the following cases:

1. The control routine suggests chiller A (*base operation*) is running AND only chiller A is observed/recommended to run (light blue).
2. The control routine suggests chiller A (*base operation*) is running AND both chillers are observed/recommended to run (green).
3. The control routine suggests both chillers (*co-operation*) are running AND both chillers are observed/recommended to run (white).
4. The control routine suggests both chillers (*co-operation*) are running AND only chiller A is observed/recommended to run (red).

The second case represents the erroneous switching on of two chillers during low demand hours. It was observed that in 92.9% only one chiller was running when the actual cooling demand was lower than 1100t (light blue area) and our prediction model recommended in 98.4% of the low demand hours to only use one chiller. In 7.1% of the cases it was found in the observed data that both chillers were running. The prediction model suggested only in 1.6% of the low demand hours to use two chillers (compare green area of center and right bar). With the prediction model 400 hours of unnecessary running of two chillers could have been avoided.

The third and fourth case addresses the prediction of the cooling operation during high demand hours (white and red area). In case the demand was between 1100t and 1200t (maximum capacity of chiller A), it was observed that only chiller A was running. For simplicity, we omit those samples (27) from this analysis, as the efficiency losses are small compared to the other cases. Therefore, these samples are not shown in Figure 6 and the white area of the left and the center bar are equally large. However, in the right bar a large red area is shown. This area shows that the prediction model recommends for 48.7% of the

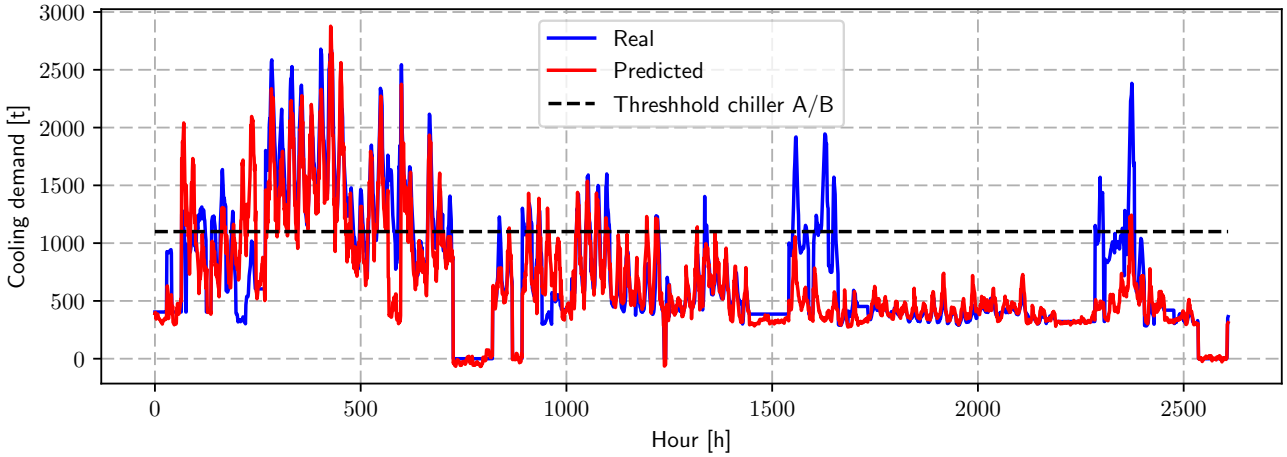


Figure 4: Sequential display of predicted and observed cooling demand of the test data set.

ISBN 978-2-921145-88-6

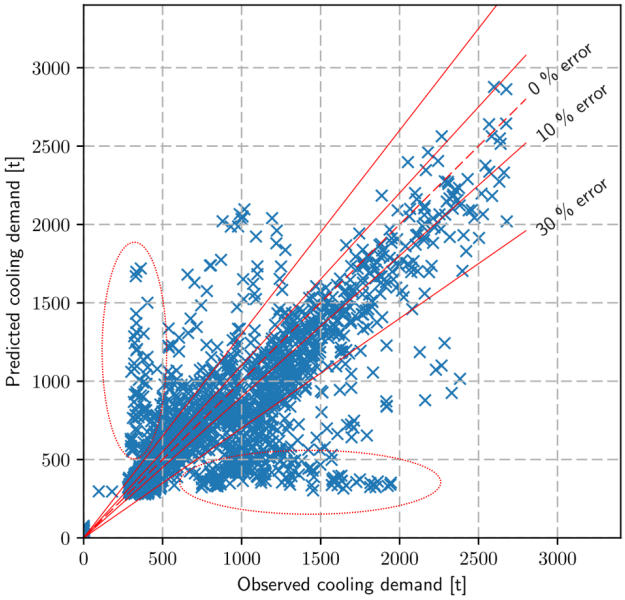


Figure 5: Predicted and observed cooling demand for test data set. The predictions which match the observed cooling demand lie on the 45° line. Ellipses indicate two distinct cases when the model fails to predict demand accurately. The horizontal ellipse highlights the area where the model predicts an idling chiller, but significant cooling is provided; the vertical ellipse corresponds to the opposite case.

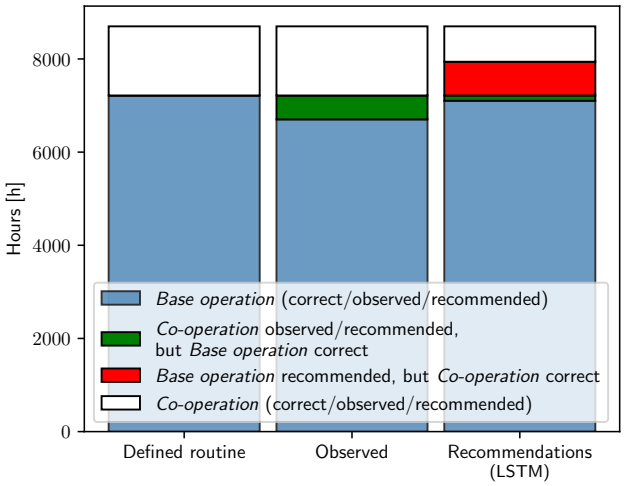


Figure 6: Performance of ML-based recommendations. Observed controls and recommendations are both compared to the optimum operation given by the previously defined control routine (see Eq. 1).
The unnecessary co-operation of two chillers (green section) is reduced if one follows the recommendations of the LSTM model. The red section shows that the LSTM recommendations sometimes fails to predict that two chillers are required. This emphasizes that even by using ML predictions, still unexpected cold starts of the second chiller may occur.

ISBN 978-2-921145-88-6

high demand hours that only one chiller is necessary. This possibly leads to cold starts of the second chiller and electricity demand spikes. In 51.3% of the time, the model predicts high demand well and prepares the energy manager for the upcoming need for the second chiller. This enables to slowly ramp up Chiller B or shift the incurred starting load to low demand hours. Nonetheless, without the prediction model no recommendation for the future are given and hence, the correct prediction in 51.3% of the time represents a benefit for the building.

Conclusions and Future work

The ultimate goal in building control research is optimal automated control of buildings. Although there are promising paradigms in development there is still significant work to be done.

We offer a simple approach to improve HVAC control by exploiting a combination of building operator's domain knowledge and the power of current ML algorithms. In this human-in-the-loop control scheme, a LSTM model is trained to predict energy demand of the building based only on weather forecast data. The demand predictions are used to give control recommendations to the building operator who can override recommendations if he receives additional information, which are not considered by the ML model as for example extraordinary occupant behaviour like conferences or meetings.

In this study, we applied that scheme to optimise chiller control. The cooling demand of a large commercial building was predicted with a LSTM model trained on one year of data. The model gives recommendations whether one or two chillers should be switched on. We compared the recommendations of the model to the observed decisions taken by the building operator: It was found that the model would have reduced the unnecessary use of the second chiller by 469 hours within one year (80.5% reduction). Furthermore, it correctly predicted the need for the second chiller in 51.3% (760 hours) of the cases it was required.

However, the model failed to predict the need for the second chiller in 48.7% of the cases. We investigated those errors and found that cooling was abnormal considering the associated weather inputs. This points out that further information on special happenings in the building are required which a building operator has. In those cases he should override model recommendations.

We aimed for a solution which can be generalized to a large variety of buildings. On the one side, the suggested method only uses weather forecast data and thus, no elaborate studying and selection of available sensor as model inputs has to be conducted. On the other, our approach requires good knowledge of the individual building system set-up. Furthermore, only buildings which are supervised by an energy manager are suitable to fully benefit.

The next step in research is to deploy hardware in the building (see Evins, 2017) to provide actual recommendations to the building operator. Recommendations of the model and control decisions of the operator need to be recorded. This will give understanding how well the collaboration of model and human works. It will be interesting to see if the accuracy of the model, which only relies on weather forecasts as inputs, is sufficient such that recommendations are accepted and in periods when the ML model is inaccurate, if the energy manager overrides the recommendations.

References

Abadi, M. et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org.

Avallone, E., I. Baumeister, and A. Sadegh (2006). *Marks' Standard Handbook for Mechanical Engineers. 10*. New York: McGraw-Hill.

Bergmeir, C. and J. M. Benitez (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences 191*, 192–213.

Chollet, F. et al. (2015). Keras. `https://github.com/keras-team/keras`.

Evins, R.; David, N. (2017). Using simple predictive models to improve control of complex building systems. In *ACM International Conference on Systems for Energy-Efficient Built Environments (BuildSys), Delft, The Netherlands.*

Gray, F. M. and M. Schmidt (2016). Thermal building modelling using gaussian processes. *Energy and Buildings 119*, 119–128.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural computation 9*(8), 1735–1780.

Jetcheva, J. G., M. Majidpour, and W.-P. Chen (2014). Neural network model ensembles for building-level electricity load forecasts. *Energy and Buildings 84*, 214–223.

Kamar, E., S. Hacker, and E. Horvitz (2012). Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pp. 467–474. International Foundation for Autonomous Agents and Multiagent Systems.

Massana, J., C. Pous, L. Burgas, J. Melendez, and J. Colomer (2015). Short-term load forecasting in a non-residential building contrasting models and attributes. *Energy and Buildings 92*, 322–330.

ISBN 978-2-921145-88-6

Mirebrahim, S. H., M. Shokoohi-Yekta, U. Kurup, T. Welfonder, and M. Shah (2017). A clustering-based rule-mining approach for monitoring long-term energy use and understanding system behavior. In *ACM International Conference on Systems for Energy-Efficient Built Environments (BuildSys), Delft, The Netherlands.*

NRCan (2012). Survey of commercial and institutional energy use – buildings 2009, *Detailed Statistical Report*. Technical report, Natural Resources Canada.

NRCan (2017). Energy fact book 2016-2017. Technical report, Natural Resources Canada.

Oldewurtel, F., A. Parisio, C. N. Jones, D. Gyalistras, M. Gwerder, V. Stauch, B. Lehmann, and M. Morari (2012). Use of model predictive control and weather forecasts for energy efficient building climate control. *Energy and Buildings 45*, 15–27.

Oldewurtel, F., A. Parisio, C. N. Jones, M. Morari, D. Gyalistras, M. Gwerder, V. Stauch, B. Lehmann, and K. Wirth (2010). Energy efficient building climate control using stochastic model predictive control and weather predictions. In *American control conference (ACC), 2010*, pp. 5100–5105. IEEE.

Stankovic, J. A. (2014). Research directions for the internet of things. *IEEE Internet of Things Journal 1*(1), 3–9.

Wei, Y., X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, and X. Zhao (2018). A review of data-driven approaches for prediction and classification of building energy consumption. *Renewable and Sustainable Energy Reviews 82*, 1027–1047.

ISBN 978-2-921145-88-6

Building stock retrofit Analysis

Paper 3

In this paper, we using data from the Victoria building stock which included pre- and post-retrofit building performance estimates for a large set of buildings. We applied multiple linear regression to determine the most cost-effective retrofit options to abate carbon emissions.

eSim 2018
«BUILDING SIMULATION TO SUPPORT BUILDING SUSTAINABILITY»
IBPSA Canada
The Canadian Chapter of the International Building Performance Simulation Association
Le chapitre canadien de International Building Performance Simulation Association

Using Multiple Linear Regression to Estimate Building Retrofit Energy Reductions

Wesley Bowley[1]*, Paul Westermann[1], Ralph Evins[1]

[1] Department of Civil Engineering, University of Victoria, Victoria, Canada
*corresponding author

Abstract: *This work applies multiple linear regression to a building energy retrofit database of the City of Victoria in order to determine the energy reductions associated with different retrofit measures. The results of the regression are then used to construct marginal abatement cost curves for retrofit options. A comparison between continuous and binary variables is performed to examine their effect on accuracy. It was found that the accuracy is comparable (R^2 for binary: 0.81, R^2 for continuous: 0.76). The regression results estimated that building envelope retrofits could reduce energy use by 40%, and heating system retrofits can reduce energy use by up to 30%. Switching to electric heat pumps could reduce emissions by an estimated 80%.*

Keywords: *retrofit, building stock, multiple linear regression*

INTRODUCTION

Retrofitting residential buildings has great potential to reduce carbon emissions through both improvements to the building envelope and by upgrading the heating systems. In British Columbia, the low carbon content of grid electricity makes converting to electrically-driven heating systems an excellent way to decarbonise the building stock. Retrofitting can also reduce energy bills for occupants. However, retrofitting measures incur significant up-front costs, which must be balanced against the possible benefits.

There are numerous ways to analyze the cost effectiveness of retrofit actions as well as how much each particular retrofit action reduces energy use. Physical modeling software can estimate the energy use of a building given many parameters and environmental conditions. However it is time consuming and impractical to model every building in a municipal building stock, and the required data is often not available.

One way around this is to collect data by surveying building characteristics as was done by Dall'O' et al. (2012). Another option is to use aggregate data from a national level and assume that this is representative of the local building stock as in Constantinos (2007), which may not be accurate.

Another option is to create building archetypes that are representative of the buildings in the stock, so that detailed simulation can be performed on a smaller number of archetypes rather than on all the buildings in the stock, while still being representative. Linear regression is sometimes combined with archetypal analysis as in Chidiac (2011), however this study only covered office buildings.

Martinez et al. (2018) use multivariate linear regression to assess the energy use reduction of retrofits that include and exclude building envelope upgrades. They found that upgrading building envelopes increase the energy savings. However the dataset is somewhat limited in size, in addition to no consideration to specific components of the retrofits (eg. Insulation, windows, etc.).

Walter and Sohn Walter (2016) use a multivariate linear regression model to predict energy use intensity with variables representing building parameters such as climate zone, heating system type, etc. The model quantifies the contributions of each characteristic to the overall energy use, then the energy saving from modifying or retrofitting that particular characteristic is inferred. The analysis is limited however in that it uses only pre retrofit data and isn't validated using pre and post retrofit energy use data.

This work aims to use multiple linear regression (MLR) to derive the statistical impact of each retrofit measure on the total percentage energy reduction. This has also been extended to carbon emissions and energy bills by making assumptions about the breakdown of energy use. Our method is similar to that used in Walter (2016) , however the key differences are that we performed the regression on the percentage energy reduction between the pre and post retrofit energy use as opposed to just on the pre retrofit energy use. The accuracy of the regression is discussed as well as potential ways to improve it.

The results of this analysis were then used to construct marginal abatement cost (MAC) curves, which quantify the cost and benefit of each possible retrofit measure. MAC curves provide a simple way of expressing this relationship. They are simplified representations of the underlying problem in that they rely on the assumption of linearity, i.e.

ISBN 978-2-921145-88-6

that separate measures can be recombined in any manner, and that the total impact will be the linear sum of their individual impacts.

The study is based on a dataset of several thousand building retrofit evaluations in the City of Victoria compiled by National Resources Canada (NRCan). This gives the retrofit actions that were recommended and performed across 50 categories alongside the pre- and post-retrofit energy use as estimated using software called HOT2000.

METHODOLOGY

There are several steps to the analysis. First the available NRCan data on the energy use reduction of building retrofits has been cleaned and processed. The cost data associated with each measure has also been collated. Next a multiple linear regression process has been used to approximate the contribution of each individual measure to the total reduction. These coefficients are used to generate MAC curves, which are analysed and then scaled to the whole building stock. Please note that due to space constraints, we are limited in the amount of data that can be shown. This includes many building parameters such as pre and post retrofit heating system efficiency and retrofit measure costing.

Database analysis

The database is created from pre and post retrofit energy audits where parameters are recorded such as wall, foundation and ceiling insulation, number of energy star windows, and information about the heating system type (various types of gas or oil furnace, ASHP, electric base boards, etc.) and fuel type (oil, natural gas, electricity, wood). These parameters were used to create HOT2000 models of the buildings and the pre and post retrofit energy use was estimated. It is the difference between these values, i.e. the change in energy use, which is used for our calculations.

Ideally it would be better to obtain energy use from direct measurements or from simulation using a more advanced tool such as EnergyPlus. However, pre and post retrofit measurements are rarely available, nor are the many parameters needed for more detailed simulation. This paper describes a methodology that can be used on other building energy databases that could perhaps have direct energy use measurements, or are for different cities.

Before the dataset could be used, it was organized and cleaned. Building entries that did not perform post retrofit energy audits were removed since they provided no way of assessing improvements due to retrofits. Building entries were grouped based on different parameters, and erroneous values were removed.

Multiple linear regression analysis

Multiple linear regression models are an extension of the standard linear regression approach that can be used to quantify the impact of multiple inputs on one output. They are a class of statistical model that generate aggregated statistical insights from many individual observations. In this study it is used to analyse retrofit measures on city level using data on building level.

Multiple linear regression generates very useful results: unlike other methods, the fitted coefficients relate directly to the variables of interest, in our case the different retrofit measures. The weakness of the method is that it assumes all relationships between the inputs and the output to be linear and independent, i.e. that there are no non-linear relationships and no interactions between variables so that the total impact will be the linear sum of the individual impacts. Since this is also an assumption of the MAC curves that the outputs will be used to construct, this is not particularly detrimental.

In this study, we use linear regression methods to quantify the impact of different building retrofit measures (e.g. wall insulation improvement, replacement of heating system, etc.) on the reduction in the annual energy consumption, carbon emissions and energy costs of a building. The model is fitted using 7000 data entries relating to retrofitted buildings within the City of Victoria. The impact of each retrofit measure is captured by the regression coefficients p_i of the fitted model as shown by the mathematical formulation of the regression model:

$$\Delta E = p_{air}X_{air} + p_{window}X_{window} + p_{ASHP}X_{ASHP} + \ldots = \sum p_i X_i \,, \quad (1)$$

where $X_i \in [0,1]$, $p_i \in \mathbb{R}$, *i = measure index.*

The output variable ΔE represents the percentage reduction in energy consumption per unit floor area. Each coefficient p_i is multiplied by a binary variable X_i which indicates whether the respective retrofitting measure i was performed (X_i=1) or not (X_i=0). The method provides the values of p_i, which here can be interpreted as the percentage by which the energy consumption is lowered if each of the different retrofit options is implemented independently. The larger p_i, the larger the impact of retrofitting measure i. The output variable ΔE is the difference between the pre- and post-retrofit annual energy use as estimated in the HOT2000 simulation on building level divided by the building area, in units of GJ/m^2/a.

As an example, we consider a simple case where there are three possible measures: windows can be retrofitted, an air

ISBN 978-2-921145-88-6

source heat pump can be installed, and wall insulation can be improved. Fitting the model to lots of different observations on buildings having conducted these measures will give the coefficients p_{window}, p_{ASHP} and p_{wall}, and the linear regression model estimates the reduction in energy consumption ΔE to be:

$$\Delta E = p_{window}X_{window} + p_{ASHP}X_{ASHP} + p_{wall}X_{wall} \quad (2)$$

For a specific building in which the windows and walls are upgraded but no heat pump is added, the percentage reduction in energy consumption is predicted to be:

$$\Delta E = p_{window} * 1 + p_{ASHP} * 0 + p_{wall} * 1 \quad (3)$$

i.e. the sum of the coefficients for the measures that were implemented. The full model is an extension of this to include all 17 measures, and hence has 17 coefficients.

Model fitting

The coefficients of the model are determined using ordinary least squares (OLS) methods. The model fitting and all related computations were programmed using the Python SKLearn Toolbox. To guarantee a statistically robust and accurate model, multiple steps were undertaken:

- The physics of the building heat balance show that the actual reduction due to building envelope and heating system retrofits are interlinked. For example, improving the insulation of a building with a low efficiency heating system is much more influential than of a building with a highly efficient heating system. To remove this link, the model was fitted to the percentage reduction in energy, emission or energy cost of a building. This modification eliminates the need to generate multiple models for each heating system type.
- The data set was scanned for outliers and 18 data points were removed.
- The coefficients resulting from the OLS fit were tested for statistical significance using the p-value score. All variables that are not statistically significant (i.e. whose p-value is larger than 0.005) are rejected from the model. The associated samples in which the associated measure is present are also removed, to reduce the variation in the remaining data.
- To verify the accuracy, the model was fitted to 90% of the data and its performance validated on the other 10% of the data. The samples for the validation set were chosen randomly.

MAC CURVES

Marginal abatement cost curves are used to compare the cost effectiveness of all retrofit measures in reducing carbon emissions. MAC curves integrate the previous findings on the impact of different retrofits on building energy consumption and the respective costs. The major advantage of MAC curves is the way they incorporate cost and emissions goals into one graph and display the most economical pathway of actions to reach a specific target.

First the energy consumption reductions must be converted in to carbon emissions reductions by multiplying the reduction by the carbon factor associated with that of the heating system and fuel type. The carbon factors for each fuel type was obtained from the BC Ministry of Environment (2016). Efficiencies of the heating systems were also accounted for.

MAC curves represent each retrofit measure according to the following metrics:

– *Annual $kgCO_2$ savings (per m^2 floor area), horizontal axis:* This number uses the coefficients of the multiple linear regression model as shown in the previous section. The percentage reduction value of each measure is multiplied by the total average pre-retrofit emissions in $kgCO2/m^2$.

– *Annual cost per $kgCO_2$ savings ($ per m^2 floor area), vertical axis:* The value above is divided by the cost of the measure. We compute the *equivalent annual cost* (EAC) to compare assets with different lifetimes, as determined for different building retrofit measures. EAC also considers the cost of capital by integrating current interest rates and inflation rates in Canada; a value of 1.16% was used Bank of Canada (2017).

MAC curves also have an advantage when paired with linear regression that they make the same assumptions regarding linearity and independence. This means that the assumptions of one method do not limit the ability or accuracy of the other method.

Energy consumption reductions are also converted into energy bill reductions by obtaining fuel cost data for Victoria, and then multiplying these factors by the energy reductions according to the fuel types (BC Hydro (2016), NRCAN (2015), FortisBC (2017)). All three metrics are examined in the results section.

*Table 1: Variables used in the multiple linear regression. R_{SI} insulation have units of m^2*K/W*

ISBN 978-2-921145-88-6

Variable	Description
thermostat	Addition of a thermostat
e2e	Upgrade of an electric heating system to a newer electric heating system.
E2G	Change from electric to gas fired heating system
E2O	Change from electric to oil fired heating system
G2E	Change from gas fired to electric heating system
G2G	Renewal of gas fired heating system
G2O	Change from gas to oil fired heating system
O2E	Change from oil fired to electric heating system
O2G	Change from oil to gas fired heating system
O2O	Renewal of oil fired heating system
GSHP	Change from any system to a ground source heat pump
e2ASHP	Change from electric furnace to air source heat pump
G2ASHP	Change from gas furnace to air source heat pump
O2ASHP	Change from oil furnace to air source heat pump
Upgrade	Renewal of air source heat pump
Air	Increasing air tightness of building, e.g. by fitting draft excluders
Window	Replacing windows
CRSI 0-4	Improving the ceiling insulation by an R_{SI} value between 0 and 4
CRSI 4+	Improving the ceiling insulation by an R_{SI} value of more than 4
FRSI 0-1	Improving the foundation insulation by an R_{SI} value between 0 and 1
FRSI 1-2	Improving the foundation insulation by an R_{SI} value of more than 1
WRSI 0-0.75	Improving the wall insulation by an R_{SI} value between 0 and 0.75
WRSI 0.75+	Improving the wall insulation by an R_{SI} value of more than 0.75

RESULTS AND DISCUSSION

In this section we first present the results of the model fitting, followed by an analysis of model accuracy, and finally the MAC curves derived from the model results.

Multiple linear regression results

The coefficients p_i of the multiple linear regression analysis give the average percentage reduction in energy use associated with each retrofit measure. The measure indexes i are given in Table 1. The results are shown in Figure 1; the numbers in brackets beside each retrofit option give the number of associated entries present in the data. The error bars display the standard error associated with each regression coefficient p_i. This is equivalent to the standard deviation of the model error, and therefore if the error is assumed to be normally distributed, then 68% of values will have an error less than or equal to the standard error.

Energy consumption

Energy consumption is lowered most effectively by installing more efficient heating systems, ideally an air source heat pump. The model suggests that a change from an electric furnace to an ASHP lowers the total energy consumption by 24%, a change from a gas furnace to an ASHP by 29% and a change from an oil boiler leads to a reduction of 37%. Installing new furnaces (especially gas or electric furnaces) leads to significant reductions in energy demand of between 10% and 17%. The reduction potential of ground source heat pumps is estimated to be 30%, but unfortunately since the dataset only features a very low number of samples (12), this value may not be accurate, and a detailed analysis of their impact is not possible.

Improving the building envelope also helps to lower energy consumption. Installing a highly effective wall insulation (R_{SI}-value > 0.75 m^2K/W) cuts energy consumption by 16%; major improvements in the floor insulation lower the energy consumption by around 10%. Improving the ceiling insulation, replacing the windows or increasing air tightness have a smaller impact. However, it should be highlighted that the building envelope retrofits can be combined, and accumulate such that they may have a similar impact to a heating system upgrade. If all possibly combinable building envelope improvements (Air tightness, window replacement, ceiling R_{SI}-Value > 4 m^2K/W, wall R_{SI}-value > 0.75 m^2K/W and foundation R_{SI}-Value > 1 m^2K/W.) are conducted a total energy consumption reduction of 41% is predicted.

The model results in negative coefficients (i.e. energy use is predicted to increase) for two of the retrofit measures: a change from electricity-driven heating to a gas powered system, and adding a thermostat. The former is explained by the reduced efficiency from 100% (electric) to rather less for gas, and also possibly the reduced cost of heating leading to increased use. The small increase in energy consumption due to installation of a thermostat may be caused by the use of the thermostat to increase comfort rather than to decrease energy use.

Some retrofit measure options do not occur in the dataset: no samples feature electric furnace upgrades, electric to oil conversions or gas to oil conversions (unsurprisingly since running costs for an oil boiler are higher than gas). As a consequence, they have coefficients of zero, and we omit them in this study.

Reduction in carbon emissions

The model suggests that electrifying the heating system is the strongest driver to reduce carbon emission. It is found that replacing gas and oil furnaces by air source heat pumps helps to cut emission by almost 80% and even replacing them by standard electric heating systems lowers emissions by more than 60%. Other heating system upgrades like changing from oil to gas or from electric heaters to a heat pump still have significant reductions of 31% and 20% respectively. The reduction in emissions by building envelope improvements are similar to those for the reduction in energy demand. It is important to note however

ISBN 978-2-921145-88-6

that the carbon factor if British Columbia's electricity grid is very low due to abundant hydro power, and these findings may not be the same for grids with a higher carbon factor.

Reduction in energy costs

The fundamental driver of energy costs are current fuel prices in Victoria as well as the effectiveness of the envelope and the efficiency of the heating system. Natural gas currently has the lowest cost and heating oil the highest cost per kWh; heat pumps have the highest efficiency of all heating systems. Based on this, the analysis of the results in the plot below are straight-forward. Changes from any system to a natural gas-fired system are estimated to reduce energy bills by at least 40% (electricity to gas) to 50% (oil to gas). The model suggests that installing a heat pump lowers bills by 24% (electric furnace to ASHP) to 38% (oil to ASHP). Two buildings which removed a gas system and installed an electric furnace instead suffered an increased energy bill of 61%. The reduction in bills by building envelope improvements are similar to the ones found for the reductions in energy demand.

Retrofit sequence effects

The order in which retrofits are applied to buildings can have an effect of the cost effectiveness of retrofits. The most obvious case is increasing envelope insulation and changing heating system type. If a building has poor insulation, it is going to require more heat through the year which will increase fuel and maintenance costs. If the heating system were to be upgraded, then the cost effectiveness will be high, since the use is high. If insulation were added first, it would decrease demand, and reduce the fuel costs, and lowering the cost effectiveness of a heating system upgrade.

The effect is more complex when emissions are considered. Switching from a fossil fuel heating system to an electric based one could be much more cost effective in terms of emissions than upgrading insulation or windows once electrifying the heating system has taken place. This is mainly due to the carbon factor of electricity being very low, so the reduction in emissions due to envelope upgrades after heating system electrification is almost negligible. Energy reductions obtained through envelope upgrades are still desirable however.

Model accuracy and prediction performance

The quality of the fitted model may be assessed by its ability to predict the energy reduction of the 10% of buildings that were not included in the fitting process (see Methodology section). The mean absolute error (MAE) and the standard deviation (SD) are given in Table 2. These indicate how much the model prediction of the annual reduction (in energy, emission or cost) deviates from the actual annual reduction. For example, for predicting the energy reduction we obtained a mean absolute error of 6.3% +/- 5.0%. Hence, in 68% of the cases (assuming normally distributed errors) the absolute prediction error is between 1.3% and 11.4%.

The prediction performance of the model was significantly improved over the course of this study, predominantly by converting the values to be estimated to percentage changes, adding further variables (e2ASHP, g2ASHP, o2ASHP) and eliminating outliers from the data.

The MAE and SD remain reasonably similar between the fitting data (90% of samples) and the testing data (10% of samples). This implies that the model is not 'over-fitted' to reproduce the fitting data as well as possible but then failing to accurately predict new testing data. The similarity implies that this is the limit of how well a linear model of this nature can represent the data available. Improving on this would either require more data (a greater number of samples), or better data (giving more details on the nature of the buildings or the actions performed). The latter is likely to give the best improvements, since the standard error values are reasonable.

Table 2: Model fitting results showing mean absolute error (MAE) and standard deviation (SD) for fitting and validation data for energy, emissions and cost models.

	Energy reduction		Emissions reduction		Cost reduction	
	Fitting Data	Val. Data	Fitting Data	Val. Data	Fitting Data	Val. Data
MAE [% reduction]	6.13	6.34	6.61	6.45	6.34	6.13
SD of error [% reduction]	4.98	5.01	6.33	5.42	6.02	5.51

Linear vs continuous variables

The regression analysis was performed using binary variables as opposed to continuous variables for several reasons. Firstly the retrofit measures that were recorded were a mix between continuous and binary with the majority being binary. For example, heating system upgrade was binary whereas insulation R value was continuous. The continuous values were separated into levels (e.g. wall R value increased by 0 to 2 m^2K/W, or 2 to 4 m^2K/W or by more than 4 m^2K/W); a binary variable was assigned to each level and the appropriate binary activated depending on the R value change that each entry performed.

ISBN 978-2-921145-88-6

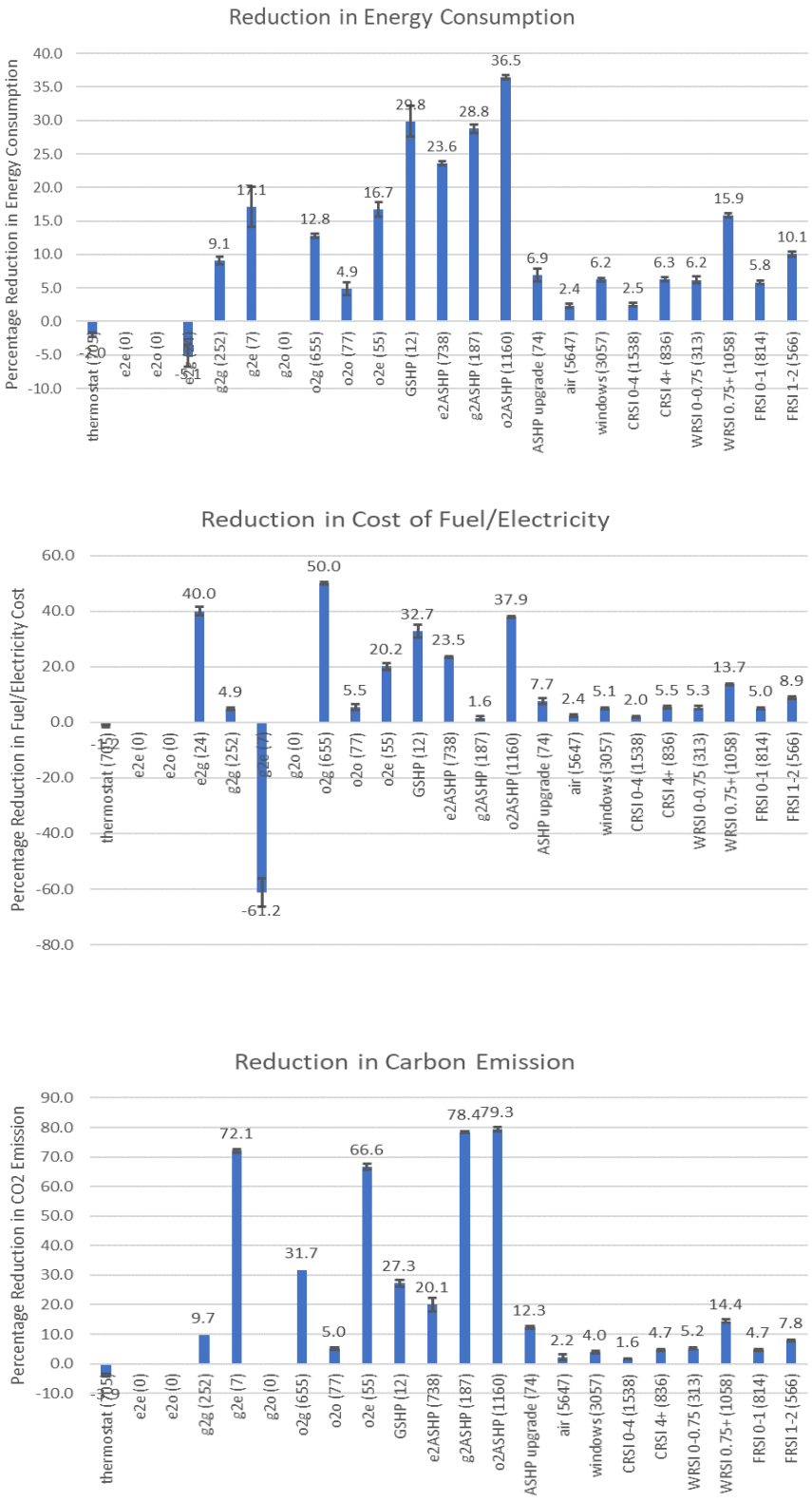


Figure 1: Results of the multiple linear regression for energy consumption, cost of fuel and carbon emissions. Each column shows the percentage reduction due to that variable. Variable descriptions are given in Table 1.

ISBN 978-2-921145-88-6

Secondly having the different levels of binaries for continuous retrofit measures also made it easier to determine if there were diminishing returns associated with different levels of that variable, whereas it could be more difficult to determine that with continuous variables due to the p_i coefficient needing to be constant over the whole range. Effectively the use of binaries is capturing high-level non-linearities in the system at the expense of low-level precision.

Thirdly the binary values may more accurately represent retrofit measures as they would be performed in reality. Wall R-value would not typically increase by 1.37 for example, but rather would be increased in discrete intervals determined by the way the construction materials are sold and installed. The discrete levels could represent separate consecutive applications of spray foam or layers of fiberglass batting. This could have practical advantages in applying this method and its results to creating municipal policy for retrofit incentives as it is simpler to communicate the requirements to residents or contractors. Interpreting the discrete variables is as simple as reading the number from the plot, whereas with a continuous variable it is necessary to account for the units of the factors before multiplying them by the result.

A comparison between using continuous variables to represent the continuous data and binary variables, as opposed to entirely discrete variables was performed on the retrofit data, to determine its effect on accuracy. Continuous variables were used for insulation R values for foundations, walls and ceilings, as well as furnace efficiency, while the rest of the variables were left as binaries since the data only indicated if they were performed or not.

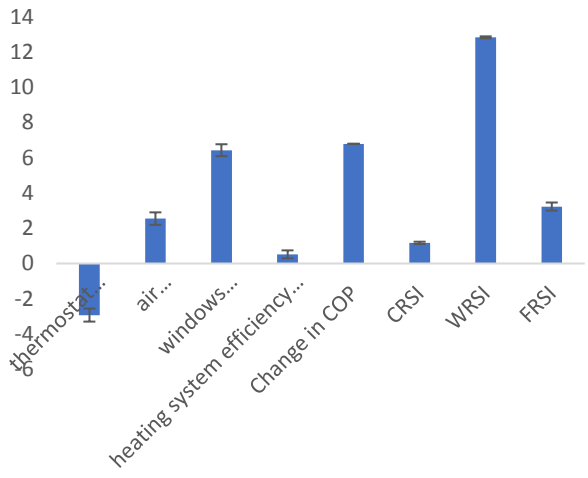


Figure 2: Regression coefficients of continuous variables.

Figure 2 gives the regression coefficients obtained for continuous variables. This gives a good example of the issue of units discussed above. The change in heating system efficiency appears to be small compared to the other variables, but this is due to the units being in percentage (usually between 70% and 100%) and the other variables having different units. This can be misleading to someone not familiar with linear regression.

A comparison of the binary and continuous fitting results showed that there was little change in the accuracy of the MLR, with the R^2 value decreasing slightly when continuous variables were used (0.81 for binary, 0.77 for continuous). One potential reason for the similar accuracy is that although we used binary variables, we had previously discretized continuous data into brackets that were each represented with a binary variable. If a single binary was used to represent an entire continuous range of data then this would likely give much poorer accuracy.

MAC curves

The results of the multiple linear regression have been combined with cost data and scaled by the city building stock to produce MAC curves, which we present in the following two sections.

Envelope and heating curves

First, we give separate results for building envelope retrofits and for HVAC retrofits. These are presented separately because the HVAC options are dependent on both the initial heating system type and on the preferences of the building owner (e.g. in prioritizing cost reductions over emissions savings).

Figure 3 shows that nearly all the retrofits that can be performed on the building envelope have negative annual cost over their lifetimes, meaning that they will pay back in energy bill savings over this period. Figure 4 shows the MAC curve for heating systems. It shows that switching oil furnaces to electric or ASHP are the most cost effective carbon reduction options. The negative cost indicates that owners would save money by switching from oil to any other heating system. Likewise, switching from gas to electricity provides large carbon reductions, however due to the low price of gas there is a positive cost over the lifetime.

Whole building stock results

The MAC curves were then used to assess the cost effectiveness of different heating system retrofits applied to the City of Victoria residential building stock. This was done by estimating the proportions of residential buildings that had gas, oil and electric heating systems according to utility connection data, BC. Ministry of Environment (2012).

ISBN 978-2-921145-88-6

The retrofit measures were then applied in these proportions to the total residential stock area. It is assumed that all building envelope items that have a negative cost will be implemented. Regarding the heating system retrofit, two different approaches are studied:

1: *Green* approach: Based on the results above the most emissions can be avoided if gas and oil furnaces are replaced by energy efficient air source heat pumps (expected emissions reductions of 78% and 79%). This scenario represents the CO_2 emissions that can be avoided if all carbon-intensive furnaces in Victoria are replaced by air source heat pumps.

2: *Cost-effective approach*: In this scenario those heating system retrofits are considered which offer the lowest abatement cost per kg CO_2 while providing significant CO_2 reductions. All gas furnaces and electric furnaces are replaced by air source heat pumps, while oil furnaces are changed to low cost gas fired heating systems. Note, that the only difference between the green and cost-effective approach is the change of oil furnaces to ASHPs instead of a change to gas furnaces.

The results for these scenarios are given in Figure 5. Total carbon emissions, equivalent annual costs and the initial investments are shown. Equivalent costs include the annualized initial investment using the current Canadian interest and inflation rates over 20 years, as well as savings from the lowering of energy bills.

The initial investment in heating system upgrades is expected to be 72M$ for the *cost-effective* approach (gas furnaces and air source heat pumps) and 90M$ for the *green* approach. The building envelope upgrades have an initial investment cost of 166M$. However, it has to be noted that the building envelope cost can be reduced if fewer measures (e.g. only wall insulation and air tightness upgrades, no ceiling or foundation insulation upgrades) are conducted. This is not possible for heating system upgrades as a full system must be purchased. This gives total initial costs of between 238 and 256M$. The estimated total annual emissions savings when the building envelope upgrades are combined with the *green* option for heating system upgrade is around 49,000 t CO_2. The equivalent annual costs are all negative which indicates a long-term cost saving by performing the retrofit scenarios through reduced energy bills.

The CO_2 abatement cost calculated in this study was compared to other MAC curves from nearby studies. The abatement costs range from $-14 to $-250 CAD$/tCO_2$ compared to our value of $-210 (Municipality of North Cowichan (2013), Canadian Association of Petroleum Producers (2015), City of Toronto (2017), McKinsey & Company (2007)). The negative values indicate that money is saved. It is worth noting that those studies are performed for different spatial scales and specific retrofit measures performed were not well defined.

LIMITATIONS AND FUTURE WORK.

A limitation of this work is the assumption of linearity in retrofit measures and their effects. Namely that the effect of two retrofit measures together do not necessarily equal the sum of effects if they were implemented individually. We recognize that assuming linearity is not entirely accurate representation of reality. However in the absence of detailed building dimensions for creating physical models, the only other option is to do more complex machine learning and non-linear modeling methods, which become more and more "black box" with complexity. We want to use a simple method that is as "white box" as possible so that it can be understood and adopted by municipalities as a tool for meeting their emissions targets.

Another limitation is that the database that was used calculates primary energy use based on the output of a HOT2000 simulation of a model with the recorded building parameters. A database that uses has directly measured energy use values pre and post retrofit would be ideal.

Future work could include moving to a non-linear model or machine learning algorithm to analyze the effects of retrofit measures, to get around the assumption of linearity that is made for this analysis. It would be interesting to then compare the results.

CONCLUSION

In this paper a novel methodology for estimating stock-level energy use reductions for building retrofits is applied to a dataset for residential buildings in the City of Victoria. The method uses multiple linear regression to estimate the amount of energy that each retrofit measure can save when applied to a building. The results of the MLR analysis are used to construct marginal abatement cost curves indicating the most cost effective and carbon saving measures. The MAC curves were then scaled by the residential building stock of Victoria to get an idea of the citywide potential for carbon reductions and the associated costs.

MLR is a relatively simple yet powerful tool that can be applied to datasets created from actual measurements from energy audits or simple simulations based on building surveys. The model was formulated using binary variables, with discrete intervals used to represent continuous data such as insulation R values. This resulted in a relatively quick set up and gives results that are simple to understand and use without post processing. A comparison was performed using the same dataset but with continuous variables where possible, and the results showed that there

ISBN 978-2-921145-88-6

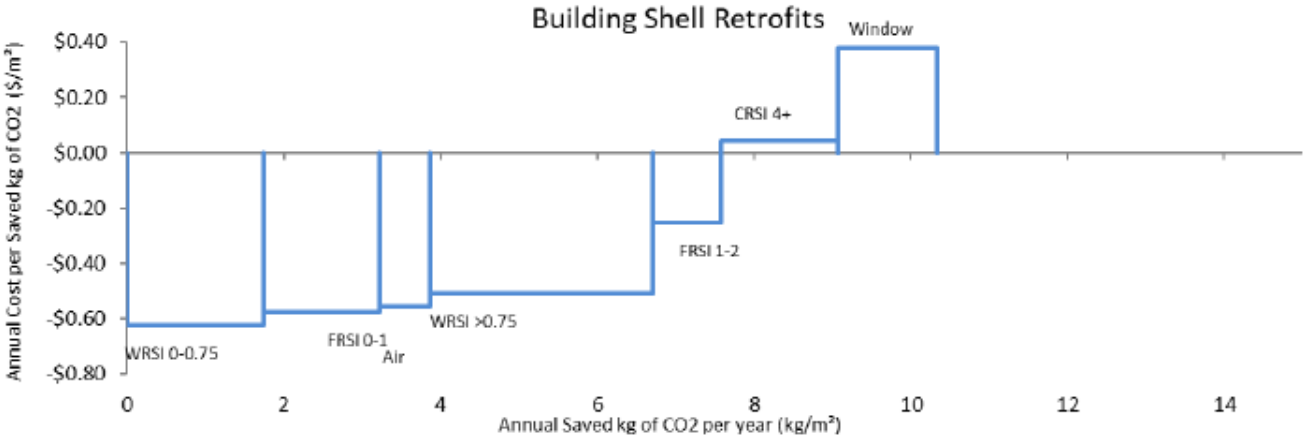


Figure 3: MAC curve for building envelope retrofits.

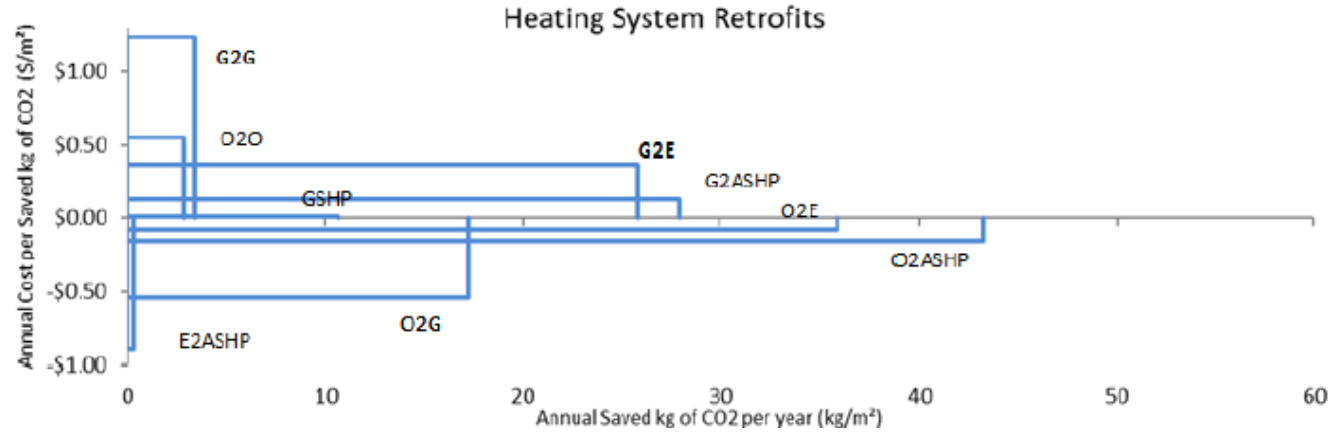


Figure 4: MAC curve for heating system retrofits. The different types overlap since only one can be performed at a time, so it is not a true MAC curve, but the comparison between options is still useful.

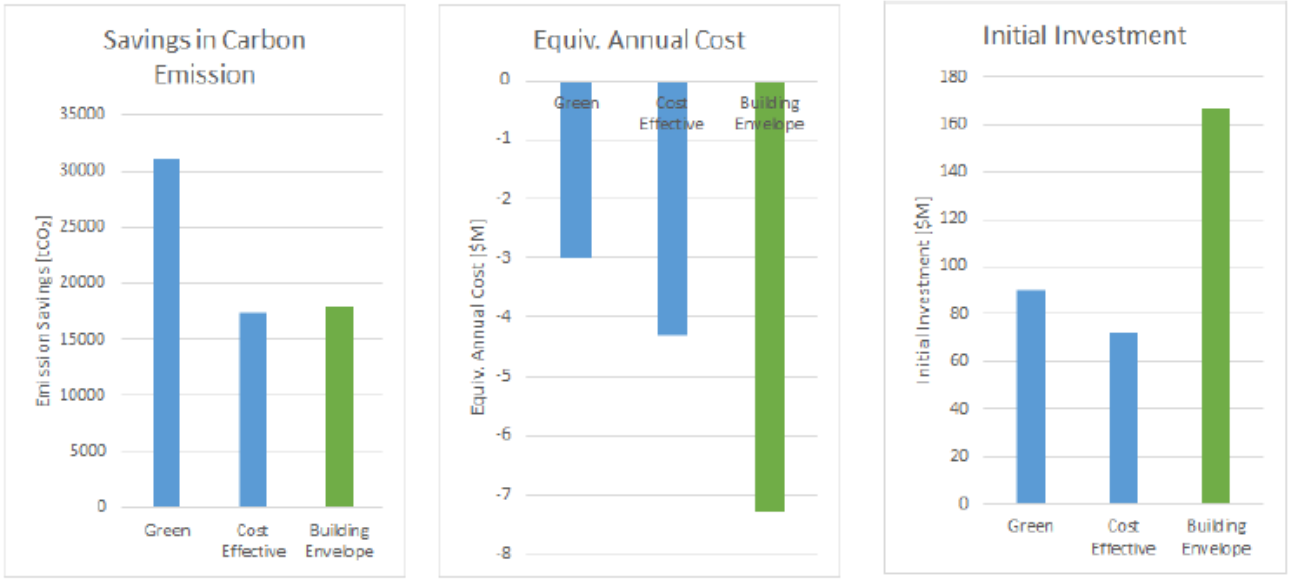


Figure 5: Equivalent annual cost, initial investment and carbon emissions savings of retrofit scenarios 1 and 2.

ISBN 978-2-921145-88-6

was little change in accuracy, and even a slight decrease for the analysis using continuous variables. The results of the MLR analysis are then used to create MAC curves, one for building envelope retrofits, and another for heating system upgrades. These are then scaled by the number of residential buildings in the City of Victoria to get an estimate of the magnitude of energy and emissions savings that could be achieved if these measures were applied. If all combinable building envelope retrofits are performed, energy use could be reduced by as much as 40%. Switching heating system types from oil and/or gas to electric, preferably with an ASHP, can give significant reductions in emissions. If all gas and oil heating systems were changed to ASHP then emissions could potentially be reduced by up to 80%. Part of this is due to the efficiency of ASHPs, but it is also due to the low carbon intensity of grid electricity in BC. Even if oil and gas were converted to electric resistance heating, reductions of up to 60% are estimated.

This paper has demonstrated that multiple linear regression using binary variables is a powerful tool. It is relatively simple to use and produces results which are easy to interpret. It can be combined with MAC curves since both methods have the same assumptions. These methods can be very useful for practical applications such as municipal policy and planning.

REFERENCES

Bank of Canada, 2017. www.bankofcanada.ca/rates/

BC Hydro, 2016. 2016/17 Annual Service Plan Report. Available: https://www.bchydro.com/content/dam/BCHydro/customerportal/documents/corporate/accountability-reports/financial-reports/annualreports/bchydro-2016-17-annual-service-plan-report.pdf

B.C. Ministry of Environment, 2016. B.C. Best Practices Methodology for Quantifying Greenhouse Gas Emissions, 2016, Available: https://www2.gov.bc.ca/assets/gov/environment/climate-change/cng/methodology/2016-17-pso-methodology.pdf

BC. Ministry of Environment, 2012, Community Energy and Emissions Inventory Initiative. Available: https://www2.gov.bc.ca/gov/content/environment/climate-change/data/ceei

Canadian Association of Petroleum Producers, 2015 CAPP 2015 Alberta Climate Change Advisory Panel Submission, 2015, Available: http://www.capp.ca/media/issues-and-submissions/alberta-climate-change-advisory-panel

City of Toronto. 2017. TransformTO. Available: https://www.toronto.ca/wp-content/uploads/2017/11/91f6-TransformTO-Modelling-Torontos-Low-Carbon-Future-Results-of-Modelling-Gr....pdf

Constantinos A. et.al. 2007, European residential buildings and empirical assessment of the Hellenic building stock, energy consumption, emissions and potential energy savings, Building and Environment, Volume 42, Issue 3, 2007, Pages 1298-1314, ISSN 0360-1323, https://doi.org/10.1016/j.buildenv.2005.11.001.

Dall'O', G. et al. 2012. A methodology for evaluating the potential energy savings of retrofitting residential building stocks, Sustainable Cities and Society, Volume 4, 2012, Pages 12-21, ISSN 2210-6707, https://doi.org/10.1016/j.scs.2012.01.004.

FortisBC, 2017., Vancouver Island Rates, Available: https://www.fortisbc.com/NaturalGas/Homes/Rates/VancouverIsland/Pages/default.aspx

Martinez, Andrea, and Joon-Ho Choi. 2018. "Analysis of Energy Impacts of Facade-Inclusive Retrofit Strategies, Compared to System-Only Retrofits Using Regression Models." *Energy and Buildings* 158 (January 1, 2018): 261–67. https://doi.org/10.1016/j.enbuild.2017.09.093.

McKinsey & Company. 2007. A cost curve for greenhouse gas reduction. 2007, Available: https://www.mckinsey.com/business-functions/sustainability-and-resource-productivity/our-insights/a-cost-curve-for-greenhouse-gas-reduction

Municipality of North Cowichan. 2013. Climate Action and Energy Plan. Available: http://www.northcowichan.ca/assets/Departments/Engineering/PDFs/NC%20CAEP%20final%20report%20v5_reduced.pdf

NRCAN, 2015. Transportation fuel prices. Available: http://www.nrcan.gc.ca/energy/fuel-prices/4593

S.E. Chidiac, E.J.C. et al. 2011. A screening methodology for implementing cost effective energy retrofit measures in Canadian office buildings, Energy and Buildings, Volume 43, Issues 2–3, 2011, Pages 614-620, ISSN 0378-7788, https://doi.org/10.1016/j.enbuild.2010.11.002.

Travis Walter, Michael D. Sohn, 2016. A regression-based approach to estimating retrofit savings using the Building Performance Database, Applied Energy, Volume 179, 2016, Pages 996-1005, ISSN 0306-2619, https://doi.org/10.1016/j.apenergy.2016.07.087.