Aalto University

School of Science

Master’s Programme in Industrial Engineering and Management

Valtteri Kukkonen

Building energy consumption prediction using statistical methods

Master’s Thesis

Espoo, Otaniemi, November 23, 2020

Supervisor: Timo Seppälä, prof.

Advisor: Timo Karvinen, Ph.D.

AALTO UNIVERSITY
School of Science
Master’s Programme in Industrial Engineering and Management

Author: Valtteri Kukkonen

Title of the thesis: Building energy consumption prediction using statistical methods

Number of pages: XII + 92

Date: 23.11.2020

Major: Strategy & Venturing

Supervisor: Timo Seppälä

Thesis advisors: Timo Karvinen

Buildings are a globally significant consumer of energy. Thus, they are also responsible for a significant amount of greenhouse gases. Accordingly, energy efficiency is top of mind in the building management community. Ability to understand and predict energy consumption better offers new energy saving possibilities, e.g., through more advanced energy management schemes. Current artificial intelligence -empowered methods enable forecasting consumption with unprecedented accuracy for building management.

The research problem of the study is to develop an easily customizable prediction model with sufficient predictive power to be used in tasks such as anomaly detection. Two gradient boosting machine -based models are developed to predict heat and electricity consumption for a couple of dozen non-residential buildings. The models developed use a combination of historical consumption and weather data as input. Even with the same model applied to multiple buildings, both models achieve strong prediction performance 24 hours into the future.

The power of current artificial intelligence methods showcased in the study underline the need for systematic artificial intelligence strategy for companies in the building management industry. An outlook in information and building management concepts as well as in state-of-the-art prediction methods is given to guide managers working in buildings-related industries.

The research follows design science methodology. Thus, it demonstrates the applicability of design science as a method of research for information technology -empowered building management research.

Keywords: machine learning, gradient boosting machine, energy consumption, prediction, forecasting

Publishing language: English

AALTO-YLIOPISTO
Perustieteidenkorkeakoulu
Tuotantotalouden maisteriohjelma

Tekijä: Valtteri Kukkonen	
Työn nimi: Rakennuksen energiankulutuksen ennustaminen tilastotieteellisten menetelmien avulla	
Sivumäärä: XII + 92	Päiväys: 23.11.2020
Pääaine: Strategy & Venturing	
Valvoja: Timo Seppälä	
Ohjaaja: Timo Karvinen	

Rakennukset ovat huomattava energiankuluttaja globaalisti. Täten myös niiden ilmastovaikutus on huomattava. Energiatehokkuus onkin tärkeä painopistealue kiinteistönhoidossa. Energiankulutuksen tarkempi ennustaminen antaa parempia mahdollisuuksia energiansäästöön esimerkiksi uusien energianhallintamallien kautta. Nykyiset tekoälypohjaiset menetelmät mahdollistavat energiankulutuksen ennustamisen ennennäkemättömällä tarkkuudella kiinteistönhoidossa.

Tämän tutkimuksen tutkimuskysymyksenä on helposti muokattavan ennustemallin rakentaminen. Mallin on oltava kyllin tarkka, jotta sitä voidaan hyödyntää esimerkiksi poikkeamantunnistuksessa. Työssä kehitetään gradienttitehostamiseen perustuvat ennustemallit sähkön- ja lämmönkulutuksen ennustamiseen. Tutkittava rakennusjoukko koostuu muutamasta kymmenestä ei-asuinrakennuksesta. Syötteenä mallit käyttävät mennyttä kulutusta ja säädataa. Vaikka samaa mallia käytetään eri rakennusten kulutuksen ennustamiseen, molemmat mallit saavuttavat hyvän ennustetarkkuuden 24 tunnin ennusteaikaikkunalla.

Tutkimuksen havainnollistama tekoälymenetelmien tehokkuus korostaa tekoälystrategian tarvetta kiinteistönhuollon yrityksille. Työssä luotava katsaus informaationhallinta- ja kiinteistönhuoltomenetelmiin auttaa yritysjohtoa tarttumaan tekoälyn tuomiin mahdollisuuksiin. Työ hyödyntää suunnittelutieteen menetelmiä ja todistaa niiden sopivuuden informaatioteknologiaa hyödyntävässä kiinteistöalan tutkimuksessa.

Avainsanat: koneoppiminen gradienttitehostaminen, energiankulutus, ennustaminen	Kieli: Englanti

Acknowledgements

First, I thank my institution, Aalto University, for propelling me to this intriguing academic journey and for providing me with the tools needed succeed. Especially I thank my supervisor Timo Seppälä whose guidance – already during my bachelor's thesis – was instrumental for the work. Timo was always available to offer advice and point the gaps in my thinking ensuring the completeness of the work. Truly have I enjoyed our cooperation through the years.

For their assistant, I would like to thank the Granlund Consulting organisation, especially my thesis steering group Timo Karvinen, Martin Aalto, and Ville Reinikainen. Great many thanks to the experts Jukka Kohonen, Risto Lahdelma, and Ken Dooley, for giving their time to provide a commentary of the model developed. Also, I thank my father for his comments and ideas regarding the final version of the work. Special thanks to Juuso Heinonen. The boundless effort he provided in discussing the analytical tools and processes, always ready to offer his help unsparing for his own troubles, truly makes him – beside Timo Seppälä – the most important support for the completion and quality of this work.

Not to forget the people most important for the other aspects of life, I thank my companion for withstanding my absence and reduced leisure time caused by working on this thesis as well as her ability to bring joy when it could not be found in working with the thesis. All in all, I am thankful for my family and friends for all their support.

Term definitions

AI	Artificial Intelligence	EMS	Energy Management System
ANN	Artificial Neural Network	ETL	Extract-Transform-Load
BMS	Building Management System	FMI	Finnish Meteorological Institute
DER	Distributed Energy Resource	GBM	Gradient Boosting Machine
DG	Distributed Generation	GOSS	Gradient-based One-Side Sampling
DR	Demand Response	IoT	Internet of Things
DS	Design Science	MPC	Model Predictive Control
DTF	Digital Transformation Framework	VPP	Virtual Powerplant

Notation

$\bar{y}$	Mean of **y**
$\hat{y}$	Prediction for y
$\tilde{y}_i$	Negative gradient at data point x_i
a	Parameters of the weak learner, obtained in learning
β, p	Parameters controlling the overall model, obtained in learning
$\frac{\delta \mathrm{f}}{\delta \mathrm{x}_\mathrm{i}}$	Partial derivative of function f at x_i
$h(\cdot)$	Weak learner, e.g., a regression tree
$L(\cdot)$	Loss function

Table of contents

List of tables

List of figures

1 Introduction

Final energy is the energy that is consumed excluding consumption in the energy sector, e.g., energy loss in conversion (Eurostat 2020a). Industry, transportation, and households are the main consumers of final energy in the European Union (EU) (Eurostat 2020b). Energy consumption in the EU by sector is shown in figure 1. The consumption is distributed quite evenly between the three. The total consumption level has also remained quite constant.

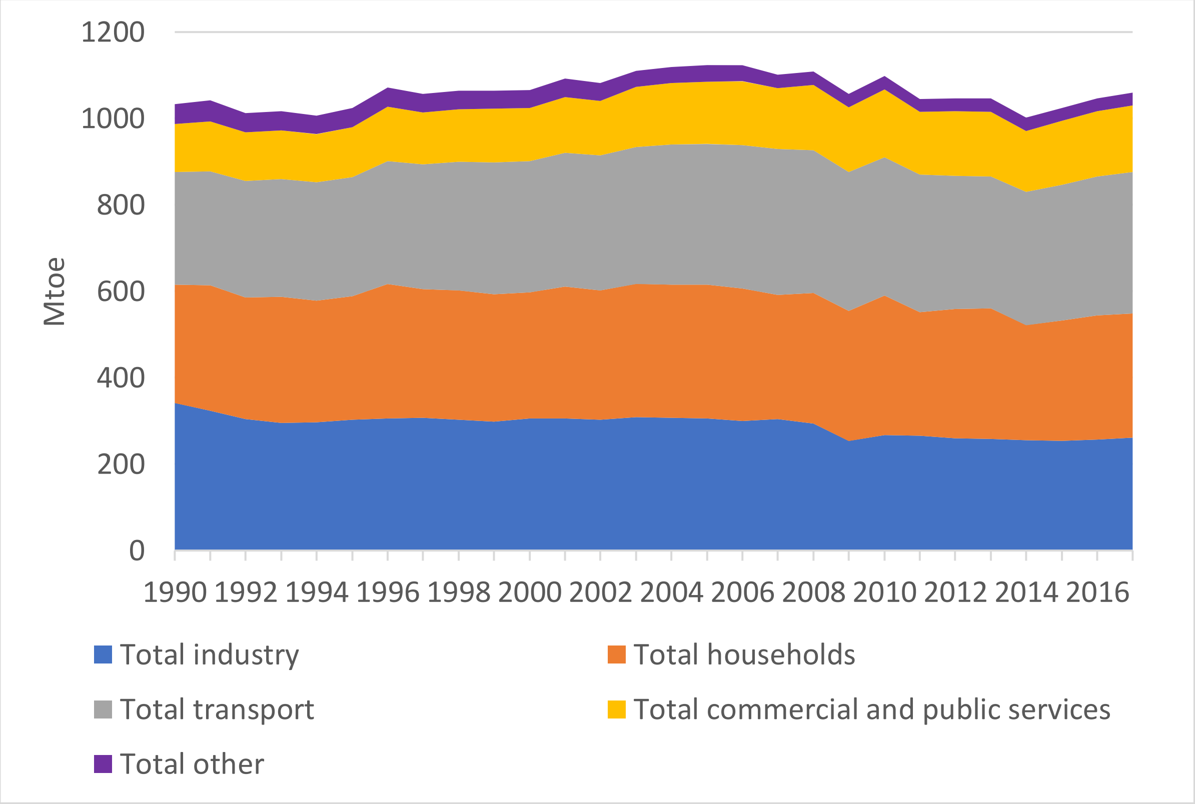


Figure 1. Final energy consumption in the EU by sector (Eurostat 2020b).

Thomas & Rosenow (2020) identify economic activity growth, energy efficiency, and the weather as the main drivers of energy consumption in the EU. The drop in consumption due to the economic slowdown between 2008 and 2012 – as discussed in Thomas & Rosenow (2020) – can be seen in the data as a downward trend in the total consumption. The mild winter of 2014 is clearly present as well. Although weather has critical impact on a year-to-year basis – depending on the harshness of the winter – weather-based deviations cancel each other out when assessing longer periods. Some researchers hypothesise that energy efficiency improvements produce rebound effects (Thomas &

Rosenow 2020). Direct rebound refers to a situation where lower energy consumption lowers energy price which stimulates consumption. Indirect rebound occurs when energy consumption increases due to heightened economic activity resulting from lower energy cost. (Thomas & Rosenow 2020.)

Let us continue with an outlook to the importance of energy consumption accountable for buildings whose impact is considerable given the large portion of energy used to heat offices and houses. For instance, heating in Finland accounted for 26 % of the total final energy consumption in 2019 (Tilastokeskus 2020). Buildings are a main driver in the household and commercial energy use presented in figure 1 as well. After discussing the critical role of buildings, the technological landscape, and the new possibilities new technologies offer are discussed. Finally, the structure of the work is presented.

1.1 Buildings' energy impact

The building sector is approximated to be responsible for 40% of the energy use in Europe (Tommerup & Svendsen 2006, Official Journal of the European Union 2010). Same is true globally (Bogdanovs et al. 2018). A study on the Swedish real estate sector estimated the contribution to be in the range of 10% to 40% of national energy use (Toller et al. 2011). The Finnish Ministry of the Environment estimated buildings to account for 40% of the national energy consumption (Ympäristöministeriö 2020). Referring to the EU Buildings Observatory data (2020), the total energy use in buildings has stayed constant throughout 2000s (please, see figure 2). Although, the absolute consumption has stayed the same, consumption with respect to floor area has decreased indicating improvements in energy efficiency (please, refer to figure 3).

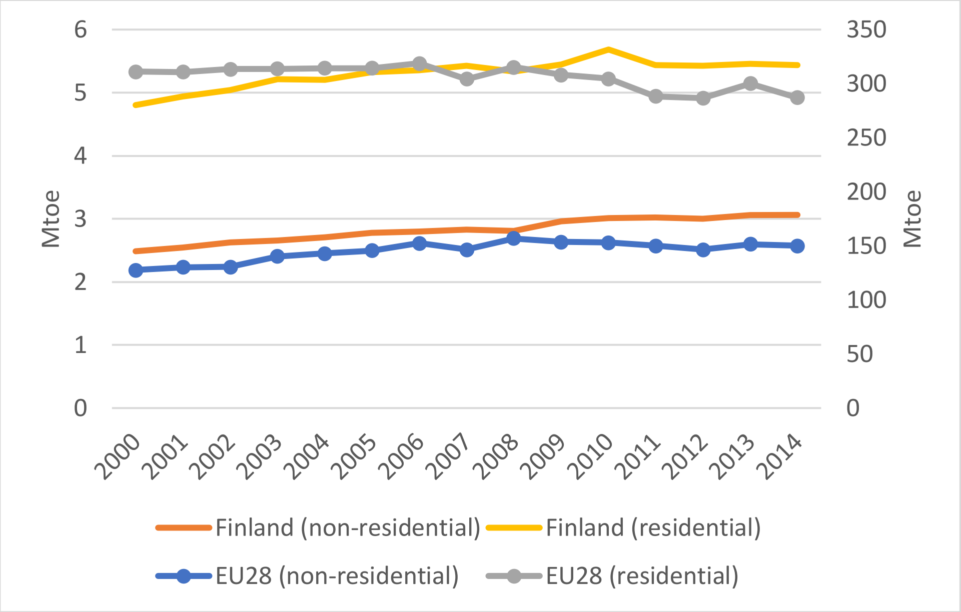


Figure 2. Energy consumption of buildings (normal climate). Please, note the secondary axis for EU-28 values. (Data: EU Buildings Observatory).

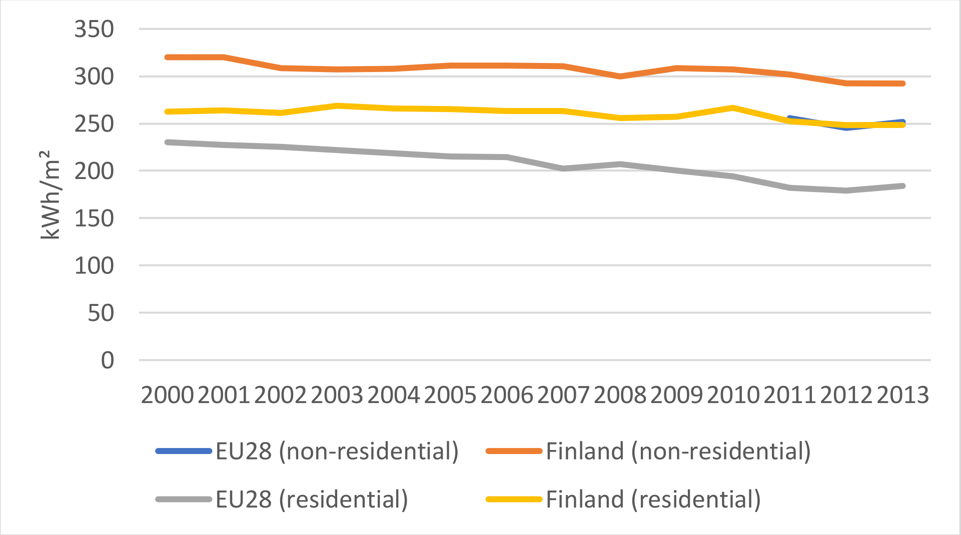


Figure 3. Energy consumption per m² (normal climate). (Data: EU Buildings Observatory).

Based on the Official Statistics of Finland: Energy Accounts, the total national energy consumption was 2 900 000 TJ in 2017. Construction accounted for 30 800 TJ and real estate for 17 400 TJ. In terms of electricity and heat specifically, construction and real estate accounted for 1 600 TJ and 14 700 TJ, respectively, from the national total of 540 000 TJ in 2017. (Official Statistics of Finland

2017.) These estimates do not contain indirect use of energy. One interesting notion is the difference in the types of energy used by real estate and construction. Real estate understandably uses primarily heat and electricity whereas fuels such as oil account for the bulk of the consumption in construction. This indicates that means to reduce electricity and heat consumption have much more potential in the operation of the building in contrast to its construction. Electricity is most used form of energy in buildings in EU. Heat energy is much more pronounced in Finland where it accounts for two-thirds of the building energy use (EU Buildings Observatory 2020). Colder climate is understandably major driver in the use of heat. Europe as a whole is much more dependent on gas, however. The building energy mix is described in figures 4 and 5 for Finland and EU-28, respectively. Finnish buildings are mainly heated with district heating. Other sources of heat are electricity and rapidly growing use of heat pumps. The energy sources used for heating can be seen in figure 6.

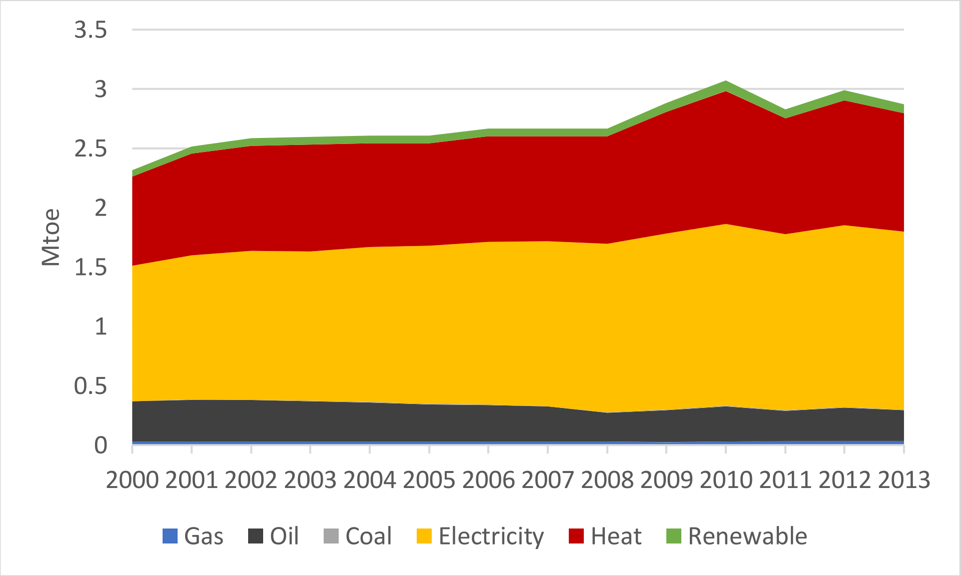


Figure 4. Building energy sources in Finland. (Data: EU Buildings Observatory 2020)

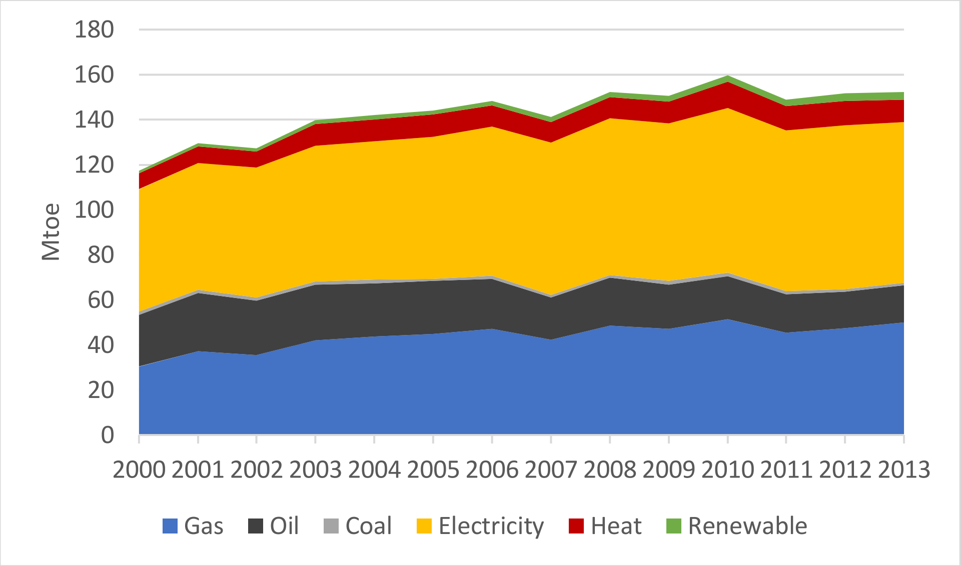


Figure 5. Building energy sources in EU-28. (Data: EU Buildings Observatory 2020)

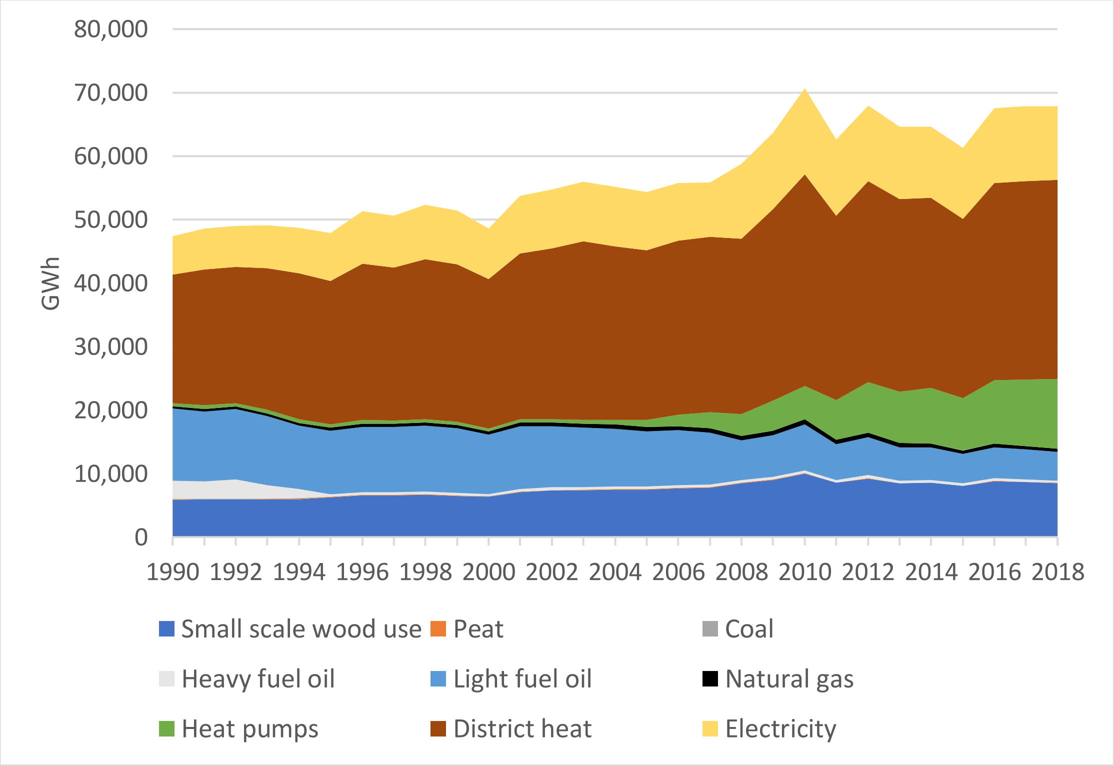


Figure 6. Sources of heating for Finnish buildings. (Data: EU Buildings Observatory).

Energy conservation in the building sector is economically lucrative but requires capital expenditures exceeding immediate operating cost savings (Nauclér & Enkvist 2009, Verhoeven 2009). Residential buildings account for around 60 percent of the energy use whereas commercial and public buildings the remaining 40 percent (Nauclér & Enkvist 2009). Energy usually is the largest operational expenditure item accounting for five to ten percent of the total for an ordinary company in the US (Minoli et al. 2017).

Considering the climate impact of buildings, they account for 30% of greenhouse gases globally (BCG 2020, Bogdanovs et al. 2018). Hence, the energy consumption in real estate is an important factor both economically and ecologically. Moreover, the rising emphasis on energy conservation is not a phenomenon specific to the real estate sector only but business leader across industries are placing more interest in managing their assets more efficiently (Gartner: Building Information Modeling 2019). There is also room for improvement; energy efficiency is estimated to account for 38 percent of the global greenhouse gas reduction potential (Nauclér & Enkvist 2009). In real estate specifically, smart buildings are estimated to reach energy savings between 20% and 40% (World Economic Forum 2016). Based on McKinsey & Company (2020) estimates it would be technologically possible to cut buildings-based emissions by 95% by 2050.

All in all, the electricity- and heat-based energy of buildings – the focus of this work – can be estimated to account for three percent of the global energy use. This estimation is, of course, crude and more indicative than precise, but it shows the global importance of energy conservation in buildings – as uninteresting it might seem. The trail of thought is depicted in figure 7.

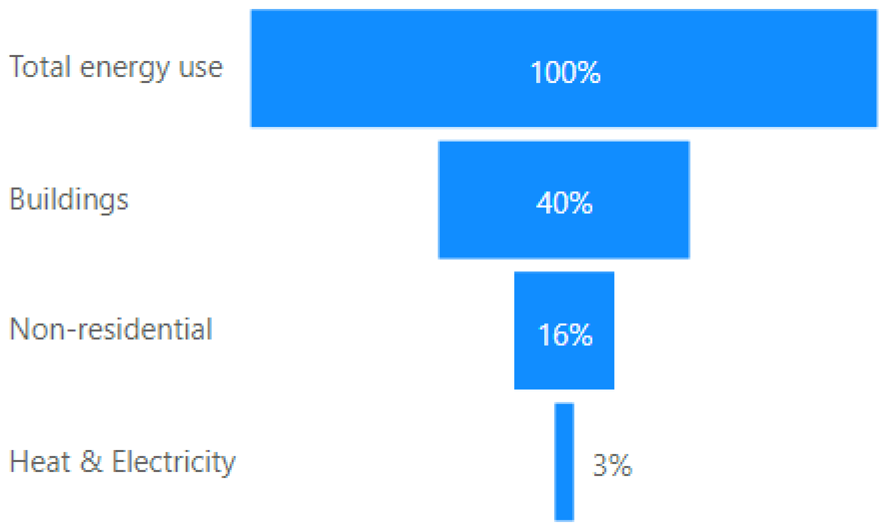


Figure 7. Size of the opportunity.

1.2 Societal pressure

Environmental sustainability has surfaced as a concern for private citizens. An indication of this mounting pressure is the finding of Nielsen N. V. (2015) – a global performance management company – that 66% of consumers are willing to pay up to 50% more for companies committed to environmental and social sustainability. As a more visible measure, the September 2019 climate protests gathered around six million people world-wide (Taylor et al. 2019). In addition to consumers, also talent is attracted by companies with environmentally friendly image. Same goes for investors as well. (Maine et al. 2020.)

The mounting pressure can also be seen in regulation. Although, the section D3 of the national building code of Finland as early as 1978 included regulation about energy economy of lighting and heating equipment as well as proper insulation (Ympäristöministeriö 1978), the requirements have become stricter through the years. For instance, the maximum allowed heat energy transmission values, i.e., the U-values, specified in the section C3 of the building code, have decreased from 0.36 W/m^2 in 1985 through 0.25 and 0.24 in 2003 and 2007, respectively, to 0.16 W/m^2 in 2010 (Ympäristöministeriö 1985, 2003, 2007, 2010). As a minor caveat, the 2007 and 2010 regulation would allow for larger values if compensated elsewhere (Ympäristöministeriö 2007, 2010). From a more global perspective, the state of New York has issued Build Smart NY program pursuing 20 percent improvement in energy efficiency in government-owned buildings (Power Authority of the State of New York 2013).

1.3 Technologies enabling the change

In addition to will one needs the means as well. The advancements in the field of data analytics – most notably the rise of machine learning methods in the domain of artificial intelligence – have introduced new possibilities to extract value from data. More powerful tools for prediction enable end-users to understand and control their consumption better (Escrivá-Escrivá et al. 2011). It is also important in the estimation of potential savings achievable by participating in demand response or by purchasing electricity from the spot market (Roldán-Blay et al. 2013). Some demand response programs also require the participants to notify the program operator beforehand about load curtailment availability. Thus, the participants need means to predict their own loads. (Krishnadas & Kiprakis 2020.)

Data is also more plentifully available through technological advancements that have driven down the cost of sensors as well as data storage and analysis (World Economic Forum 2016). As a result, the number of devices connected to the internet is expected to boom from the 11 billion of 2018 to 125 billion in 2030 (Mittal et al. 2018). Incumbents, such as real estate owners, are in a favourable position regarding data availability as they have accumulated heaps of historical data. This is pivotal for training artificial intelligence solutions. (Russo & Wang 2019.)

1.4 Other trends transforming real estate

Here trends affecting real estate that are not directly linked to prediction are briefly discussed. These topics are introduced to complement the discussion on buildings to help the reader form a comprehensive view on the building industry and better understand the viewpoint of this work.

1.4.1 Demand for new types of space

The rise in e-commerce has decreased the need for physical floor space in older shopping centres. The need for logistics property has risen due to the fact, however. The economic prosperity of developed nations also supports demand for student accommodation and hotels as more money becomes available for education and holidays. Flexible lifestyle is also becoming more popular with increasing demand for rental properties. (Knight Frank 2019.)

Office space quality is becoming more important as means to attract talent. Especially, the expansion of technology firms and coworking spaces boosts the demand for well-located offices with exciting work environments. Corporations have also expanded toward less-known locations – so called tier two cities – which has driven demand for new office space. (Knight Frank 2019.) The present Covid-19 crisis has serious effects on the labour market, however, and it already has decreased overall job demand (Béland et al. 2020). One might think this to decrease the demand for new office space. However, Barrot et al. (2020) showed the impact – at least from distancing measures – to be much smaller in well-paid industries such as computer services or consulting. Fana et al. (2020) also highlight the effects being highly asymmetric with less impact in high-value-added operations. On the other hand, they also note the popularity of telework which may be permanent. This would understandably reduce the demand for physical floor space.

1.4.2 Urbanisation

Urbanisation – with overall population growth – is an important factor supporting real estate demand. United Nations (UN) statistics predict both urban and total population to grow throughout to 2050. Urban population is expected to grow even in developed regions although the total population in these regions is estimated to plateau in 2030s. (UN 2018.) The statistics discussed are visualised in the figure 8.

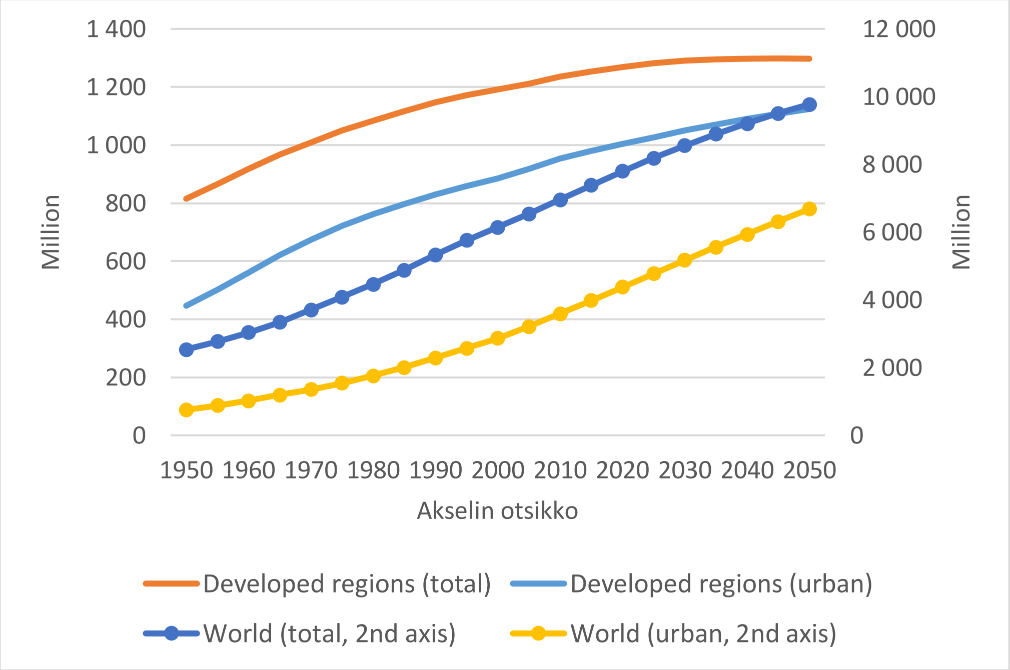


Figure 8. World population in millions (please, note 2nd axis for world totals) (Data: UN 2018).

The housing demand driven by urbanisation has resulted in a lack of affordable housing. McKinsey uses housing costs under 30 percent of household income and commute time under an hour with some basic qualities of the dwelling as the definition for affordable housing. Using these metrics there is estimated being 330 million urban households globally living in substandard dwellings or being so financially stretched by housing costs that they cannot cover for other basic needs, including food and health care. There are cities of a large variety affected; university cities, new employment hubs, capital cities, with the globally largest 100 cities accounting for two-thirds of the "Affordability Gap" identified by McKinsey's Global Institute. (Woetzel 2014.)

In addition to the urbanisation of people, Wetzstein (2017) identifies the urbanisation of capital as a driver for housing demand. Cheap money has increased mortgage lending that is driving up the prices. In addition to financial hardship, rising prices in the city centres hinders the functionality of a city by pushing "key workers" – such as police officers, teachers, nurses, and cleaners – away from the metropolitan areas. Reduced homeownership and affordable renting also inhibit upward social mobility posing risk for social harmony and citizenship. (Wetzstein 2017.) Construction costs are, of course, an important factor affecting the price of housing. The respondents of a survey study by PwC and the Urban Land Institute for European real estate companies highlight rising construction costs having a large impact in their business. Labour and material costs combined show yearly increases between five and seven percent. (PwC 2020.) McKinsey research identifies 30-percent cost reduction potential through the application of industrial approaches, e.g., off-site manufacturing, standardisation, and improved purchasing and other processes (Wetzstein 2017).

1.5 Overview of the work

This work discusses broadly the use of prediction in building management. An implementation is also developed for energy consumption prediction. The work provides practitioners with a concrete representation of the power of analytics and open source tools suitable for building management. A prediction model is developed from scratch. The code used is accessible in the author's Github (github.com/HVKukkonen). For academia, the work showcases the suitability of the design science methodology in the development of analytics for engineering applications. A thorough literature research about prediction as well as data and building management concepts is also provided.

As a start let us discuss the current technological landscape affecting buildings and their management in section two. Data management concepts related to this work and its implementation and implications are discussed in the third section. In the section four, the methods needed to develop the implementation are presented. Section five goes through the research methodology, after which the problem to be addressed and the data sources available are discussed in section six. Section seven contains thorough analysis of the data and guides the reader through the steps needed to refine the data. The development of the model is exhibited in the eighth section. Section nine covers the evaluation of the model developed. Concluding the work, implications of the model, as well as limitations and opportunities for future research are discussed in the section ten.

2 Perspectives on buildings

In this section, the technologies impacting buildings and building management are considered. Let us start by laying the foundation for the subject by a concise introduction of the main technological developments and concepts impacting future buildings. This introduction discusses buildings on two levels, namely the system level and the building level. To help keep track of the level of discussion, please, refer to figure 9. As a conclusion, the implications, and already on-going changes these technologies introduce to building management, are discussed.

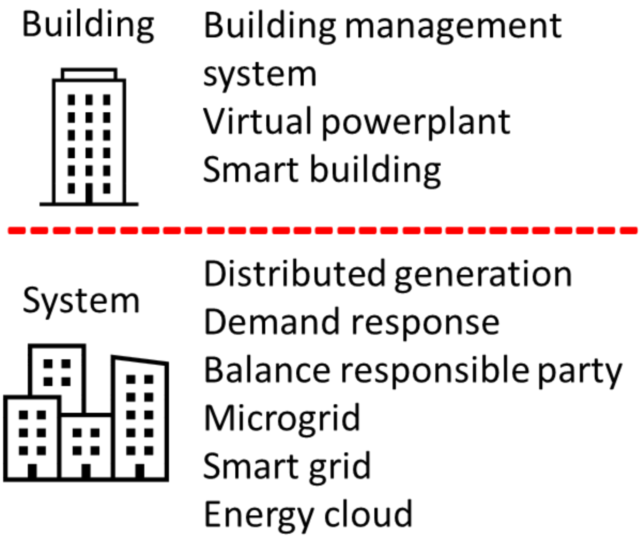


Figure 9. Levels of discussion for buildings.

2.1 Concepts defining modern buildings

In this chapter a concise overview of the most relevant concepts for modern buildings is made. This, and the following chapter form the context for the discussion about the future of buildings. The introduction is in no way exhaustive but provides sufficient technological background to assess the possibilities and requirements for prediction in buildings.

2.1.1 Building management system

Building management system (BMS) – also referred to as energy management system (EMS) or building automation system – is a computerized system running a dedicated BMS software that controls the HVAC equipment through an industrial control network (Litiu et al. 2017 Ch. 1, Sioshansi 2013 Ch. 18). Also, electrical, and mechanical systems as well as lighting can be considered in BMS. A BMS system can be considered to consist of sensors, controllers; responsible for instructing the system through output devices, communications system, data analytics and dashboard; for displaying data to and accepting commands from users. Typical operations involve operation scheduling, data gathering and analysis as well as management of demand response resources. (Minoli et al. 2017.)

Efficient operation of the BMS is important as energy savings from better use of older equipment can equal the benefits of changing to newer equipment. What is more, retrofitting required to upgrade equipment is not always possible for cultural and historical reasons. (Litiu et al. 2017 Ch. 2.) Voltage and current levels have usually provided the means of communication in electrical and electronic systems. Complex data transmission through electrical wiring makes interoperability a cumbersome engineering task, however. The IT infrastructure present in buildings can be used to provide connectivity for the system. Use of open standards in the BMS communication network is important to provide flexibility and remove lock-in to one vendor. Although increased engineering effort, this reduces the lifetime cost of the system. BACnet, LonWorks and EIB/KNX are open communication standards common in the BMS space. (Kastner et al. 2005.)

2.1.2 Virtual powerplant and smart building

Virtual powerplant (VPP) means a concept where the operation of multiple distributed generation resources is combined into one entity. A common example is a case where inherently fluctuating generation, e.g., solar or wind, is combined with controllable generation such as a combined heat and power (CHP) plant. Hence, with intelligent control of the CHP plant, it is possible to produce a steadier generation profile that resembles a conventional plant. (Wille-Haussmann et al. 2010.) Thus, the name “virtual powerplant”. VPPs need to be able to predict and model their loads reliably to be able to deliver the sold capacity (Palensky & Dietrich 2011).

Buckman et al. (2014) define a building as smart if it is able to achieve adaptability – in contrast to being reactive – through advanced information collection and usage, building controls, and advanced materials. Adaptable building is able to predict future events, e.g., a change in occupancy or weather conditions, and consequently adapt its operations and physical form to account for the events. For Litiu et al. (2017 Ch. 9), “smart” means the addition of energy efficiency, load management, and carbon footprint optimization measures to the building. Forecasting models are seen as a cornerstone enabling these functionalities.

2.2 Key concepts in modern energy systems

Here the introduction is expanded with a discussion of concepts at the system level. This – with the previous chapter – forms the basis for the discussion in the next chapter. Here again, the list of concepts is not exhaustive by any means.

2.2.1 Distributed generation, demand response and balance responsible party

Hernandez et al. (2014) distributed generation (DG) means low-power – up to some hundreds of megawatts – electricity generation that is in close vicinity to loads making them almost independent of the distribution grid infrastructure. Renewables are common DG resources. Distributed resources are by their nature more fluctuating in comparison with traditional generation. This creates demand for accurate demand forecasting. (Hernandez et al. 2014.) DG also introduces variation in voltages and backward power flows for local grids posing new requirements to grid infrastructure (Sánchez-Jiménez 2006).

Demand response (DR) means curtailment of loads by an energy consumer as a response to a signal from the grid operator. The response does not always need to be instantaneous as often grid emergencies can be anticipated. DR can be split by the type of signal to market DR and physical DR. In market DR the signal to alter loads comes through monetary incentives such as a change in price resulting in a voluntary response. In physical DR, on the other hand, the signal is binding. Physical DR requires the loads to be always alterable limiting the number of suitable processes. (Palensky & Dietrich 2011.) DR is generally market-based (Lawrence et al. 2016). Limited customer elasticity and the fact that not all physical situations are mapped in the prices necessitates the use of physical DR as well, however (Palensky & Dietrich 2011). DR is needed as the share of renewable resources in the grid increases (Lawrence et al. 2016).

Balance responsible party is a regulated party in the electricity market that maintains the balance of demand and supply in a given part of the electrical grid. A transmission system operator (TSO), responsible of the whole network, outsources the balancing of the network to a group of these actors all managing their own portfolio of resources. All consumers and producers need to have a contract with a balance responsible party. (next-kraftwerke.be.)

2.2.2 Microgrid, smart grid and energy cloud

Microgrid by Jirdehi et al. (2010) is an integrated power system of distributed resources – considering for both generation and storage – and nearby loads. An important functionality of a microgrid is its ability to operate disconnected from the main grid in an “emergency mode” (Lopes et al. 2006). From the point of view of the network operator, microgrid forms a single entity (Anduaga et al. as cited in Hernandez et al. 2014). A building itself can be considered as a microgrid (Hernandez et al. 2014).

Smart grid by Mantooth (as cited in Hernandez et al. 2014) defines smart grid as a power system that achieves autonomous responsiveness and efficiency through hardware and software added to the system. This requires devices that allow for two-way communication between end-users and grid operators or utilities such as smart meters and smart appliances (Lawrence et al. 2016). The main requirements for a future electrical grid by Ferreira et al. (2010) are reliable transmission of electrical energy, accessibility for all users including distributed energy resources (DER) such as local renewable generation, flexibility that enables responding to unexpected events and innovativeness from economic perspective enabling new business models and services. According to European Commission research, goals for smart grid development are adoption of technical standards that enable a range of potential solutions, computation- and telecommunication -based additional services, common regulatory and commercial frameworks, utilisation of existing infrastructure and development of low-cost, easy-to-deploy solutions. (Sánchez-Jiménez 2006.)

Projecting future development in the field, there is a strong belief among some industry experts that the BMS will migrate to “energy cloud” (Minoli et al. 2017). Lawrence & Vrins (2018) describe the energy cloud as a highly digitised and dynamic two-way energy system relying on cloud technology. It contains everything from buildings and vehicles to power generation and transmission systems. Power generation is seen to be based mainly upon distributed energy resources (DER) such as solar and wind power and demand-side applications, e.g., demand response (DR) and energy efficiency, thus the quest for a cleaner energy system is driving the change.

2.3 The future of building management

The three goals of building design are energy efficiency, comfort, and longevity. These goals are not always aligned – for example low indoor temperature in the wintertime is energy efficient but not comfortable – which implies that no globally optimal way to operate a given building exists. Thus, in contrast to fully automating building operations, a modern building re-engages occupants to building control. For example, some buildings adapt indoor conditions to match the preferences of different occupants. Another example is the creation of multiple temperature zones to office spaces. (Buckman et al. 2014.)

Building needs to adapt to changes in its environment. Better adaptability offers significant energy efficiency improvements with additional comfort, for example, as the outside temperature changes. For a building to be adaptable, however, it needs to predict future events. This requires the adoption of external data sources, such as room booking system data, for building operation. This is data that usually exists but is not integrated with the BMS. (Buckman et al. 2014.) As a case in point, Bogdanovs et al. (2018) integrated weather forecast in the heating control system achieving eight-percent savings while simultaneously increasing indoor temperature stability. Prouzeau et al. (2018) focused on the dashboard to visually aid the technical manager to get insight from the building data. In addition to the pivotal role of the user interface, they discuss the possibility to enhance energy efficiency and occupant comfort through accurate inside temperature prediction. As the examples show, learning is an important feature of a smart building. The building learns to spot patterns from occupancy history and building usage to constantly improve its operations. (Buckman et al. 2014.)

Model predictive control (MPC) is a control scheme making its way to buildings from the process industry that builds on the idea of adaptability. The main idea of MPC is to include a dynamic model of the building into the control scheme with prediction capabilities. An MPC is able to predict the main disturbances affecting the environment beforehand, to assess the impact of these disturbances to the system and to optimise its actions according to pre-set goals. A natural application for MPC in the heating control of a building. As an illustration, one can consider an office building when outside temperature drops suddenly. A conventional control scheme, e.g., the proportional-integral-derivative control, reacts to the situation only as the inside temperature falls below a certain setpoint. Idea in MPC, on the other hand, is to first predict the drop in the outside temperature, i.e., utilising a weather forecast. Then to assess the impact to the inside temperature using the building model. Consequently, MPC would start heating the building even before the outside temperature drops. Hence, MPC is able to reach much more stable indoor climate than conventional control schemes. (Killian & Kozek 2016.)

MPC is a prime example of the possibilities offered by the combination of prediction with controls. MPC is extremely useful considering the system balancing effort with renewable sources discussed earlier. With MPC, buildings with large thermal capacity can be utilised as energy storages by pre-heating or -cooling them (Killian & Kozek 2016). From the smart grid point of view, buildings transform from unpredictable loads to balancing resources. Many experimental implementations of MPC have been developed and have showcased good performance, commercialisation has not yet occurred, however. The main obstacle for an MPC implementation is its reliability on an accurate building model. No plug-and-play solutions for the modelling exist which renders the modelling building specific. What is more, the lack of easy-to-use tools for modelling impose a need of talent with expertise in modelling. This is a bottleneck in the building industry. The lack of data about occupancy and indoor environment in residential buildings imply non-residential buildings as the natural starting place for MPC adoption. Moreover, MPC does not require replacement of conventional building control but can be adopted as an additional supervisory layer. (Killian & Kozek 2016.)

Smart technologies not only enable the optimisation of energy consumption on the building level but also on the system level. Smart buildings can participate in balancing the energy consumption profile in the grid by controlling its loads as a part of a demand response program. Smart buildings can have their own generation resources, e.g., solar or wind power, or be designed to store electrical or thermal energy. (Lawrence et al. 2016.) As a case in point, Li (2019) created a BMS dubbed "economical BMS" that automates the purchase of electricity for a building using AI-enabled methods to optimise the use of existing storage and consumption assets. Cost saving potential up to 61 percent was reported. In contrast to Li (2019), the introduction of smart capabilities usually happens, not in the BMS level – as few small and mid-sized buildings even have centralised BMS – but in the component and subsystem level. This lowers the barrier to improve the functionalities of legacy buildings. The introduction of demand response functionality is one of the most important capabilities enabled by smart buildings.

Smart technologies are not about only hardware, however. Algorithms to coordinate building operations are needed as well. Control and coordination are pivotal in integrating smart buildings into the smart grid. (Lawrence et al. 2016.) In addition to technology, demand response also requires a pricing system that dynamically reflects the balance of supply and demand (Palensky & Dietrich 2011).

Possibilities offered by advancements in the fields of automation and artificial intelligence (AI) offer possibilities beyond just making individual buildings smart. In the energy cloud framework BMS migrates from the building level to cloud. (Lawrence & Vrins 2018.) The energy cloud enables the analysis and comparison of multiple sites within one system in comparison to traditional site-specific BMSs. It also enables the use of external data sources to aid in the analysis. Instead as with building-specific systems, a single cloud-based system offers harmony and standardisation to site management and enables focusing development resources. Harnessing data from multiple sites also increases the reliability of analysis through data quantity and enables benchmarking between similar operations. (Sequeira et al. 2014.)

Given the emergence of the energy cloud, value will migrate from production of energy to additional services. Hence, data surpasses the actual energy as the most valuable resource exchanged. Also, energy is seen to flow both ways as individuals and especially companies with large office spaces start small-scale energy production, e.g., through photovoltaics. These developments will disrupt the whole industry. Incumbents have the customer relationships and data making them well positioned for the transition, however. Operational efficiency is seen as the natural starting place to apply advanced analytics with companies already reporting savings of 20 or 30 percent in efficiency. Nevertheless, possibilities to create new offerings or improve customer segmentation and targeting need to be pursued to stay competitive. (Lawrence & Vrins 2018.)

As a more concrete representation of the possibilities of cloud technologies in the energy space, Sequeira et al. (2014) created an energy management system (EMS) for an industrial setting. The authors gathered time series measurement data in real-time from energy meters, analysed this to spot irregularities or inefficiencies and provided this information for potential site managers through customizable visual dashboard. Their system effectively contains the elements described by Minoli et al. in the case of BMSs. Here again, data visualisation is utilised to carry the information refined by analytics to the user.

Giordano et al. (2019) implemented energy cloud in the University of Calabria Campus, connecting sites with photovoltaics production and/or energy storage to form a microgrid at the campus. They assigned "community energy provider", a non-profit entity, to aggregate the energy resources in the microgrid and manage the allocation and purchasing of energy. The management of energy occurred in two general phases. First, the demand and supply of energy from the energy resources was estimated and a time-of-use tariff determined. In the second phase, demand response (DR) program is applied to distribute the energy among participants through an auction. The work of Giordano et al. offers a real-life implementation of a cloud- and renewables-based energy system in a setting that resembles a real city. It is clear how the system described is dependent on predicting consumption and demand both from the individual participants' perspective and as aggregate.

Also, Newaz et al. (2014) developed an energy cloud platform to analyse energy consumption and control appliances. They highlight the importance of accurate demand forecasting in smart grids. By the authors, accurate modelling enables significant capital and operating expense savings. For example, reliable and accurate demand prediction would remove the need to invest in energy storage for supply reliability.

3 Perspectives on data management

Given the vast amounts of data present in modern building technologies data management needs to be considered. Considering data processing, a birds-eye-view into the data processing pipeline of an organisation is sufficient for us. Organisations have multiple data sources that impact their operations directly or that can be used for analysis purposes. The first step of data processing is the ETL process – introduced in subchapter 3.1.1 – that governs the extraction of data from the data sources and the required transformation to enable its storing into the organisation's systems. The stream processing enabled by the Lambda architecture's speed layer can be utilised to ensure the quality of the incoming data by adding anomaly detection algorithms that operate on the layer (Kiran et al. 2015). The stored data can be utilised for strategy and other business development of the organisation or it can be a vital part of the company's operations. As discussed in the relational databases subchapter (3.1.1), the storage facilities are often separated by the use case of the data they contain.

To draw insight from data, however, it needs to be analysed. Analysis can describe any "detailed examination of anything complex in order to understand its nature or to determine its essential features" (Merriam-Webster 2020). Prediction – the relevant form of analysis for our purposes – will be discussed in the next section. For the analysis to have an impact, the resulting insight needs to be incorporated into the operations of the organisation. In case of models, e.g., prediction models, this is called model deployment. The overview of this process is presented in figure 10.

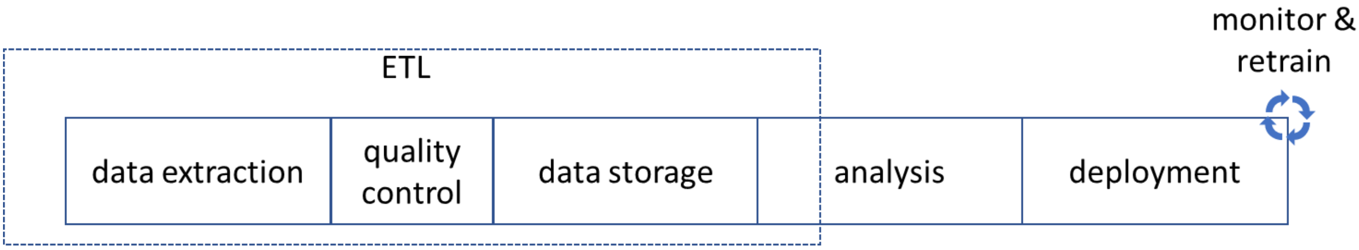


Figure 10. Overview of data processing.

The intelligence cycle used by the law enforcement and intelligence agencies offers an analogy to the data processing pipeline presented here. Intelligence cycle consists of data collection, its refinement into information and analysis on this information to form intelligence (Warner 2013, ch. 1). This shows how applicable the process is to all information processing. Zachman (1987) exhibits similar search for analogies by formulating software architecture based on concepts in classical architecture.

Let us start by an introduction to the technological concepts most relevant to us. Building on this introduction, a link to buildings-related concepts discussed in section two is formed in chapter 3.3. Also, the concepts redefining the industry are discussed.

3.1 Key concepts in data management

This chapter starts by a concise introduction of some of the technical concepts present in the field of data engineering and data management. Concepts as fundamental as relational database and extract-transform-load process are discussed in the first subchapter. More recent developments such as cloud computing, machine learning, and blockchain are discussed in the second subchapter. Here again, an exhaustive list is not pursued but only a sufficient introduction to the technologies relevant for the discussion in the following chapter.

3.1.1 Building blocks of data management systems

Database means a computer application that stores and retrieves data, potentially in large quantities (Amos & Amos 1999, p. 82). Another definition by Hernandez (2013), defines database as an organised collection of data used to model an organisational process. In a relational database, data is stored in tables. The system managing a relational database is called relational database management system (RDBMS). (Cammack et al. 2006, p. 577 and Hernandez 2013.) Information in a relational database can be stored in multiple places with the relationship information saved in the database (Tomsic 2000, p. 337). The relational database concept was developed by Dr. Edgar F. Codd in the late 1960s with the first implementation in 1969 (Hernandez 2013). From the database management point of view, databases are either operational or analytical. Operational databases store dynamic data to represent a state of something, e.g., inventories of a retailer. Analytical databases, on the other hand, store historical data for business development purposes. (Hernandez 2013.)

Data warehouses are data collections for decision support applications. They provide historical and summarised data consolidated from operational databases. Accordingly, they are much larger as well. Data warehouses serve ad hoc, complex queries to answer specific business questions. Data marts are similar structures with the exceptions that they are organised around a specific organisational subset. Data marts are faster to roll-out due to their smaller size but can lead to integration problems if data from multiple marts is needed. (Chaudhuri et al. 2001.) Data warehouses are traditionally implemented as relational databases (Chaudhuri et al. 2001 & 2011). Bill Inmon is commonly regarded as having developed the concept (Hernandez 2013).

Business intelligence means the tools and practises to collect, analyse and present large quantities of data to help in corporate decision making (Dayal et al. 2009). Chaudhuri et al. (2011) note that business intelligence applications usually run on data kept in a data warehouse. These applications usually take the form of spreadsheets, query portals, and dashboards (Chaudhuri et al. 2011).

Extract-Transform-Load – more commonly referred to with the acronym ETL – is a process governing the preparation of data from various sources and its loading to storages to be used in business intelligence applications (Bansal & Kagemann 2015, Chaudhuri et al. 2011). By Chaudhuri et al. (2011) and Bansal & Kagemann (2015), ETL tools ensure data quality through discovering and correcting faults. They can enforce uniqueness of certain attribute combinations, e.g., a name-address combination. ETL also governs the capturing of wanted data and its efficient loading into the storage. (Bansal & Kagemann 2015 and Chaudhuri et al. 2011.) Efficient loading is vital as the amount of data is large. The systems also need to be resistant against crashes, e.g., by keeping a log of the changes (Chaudhuri et al. 2011.)

The ETL process can be executed in batches where the process is initiated periodically. It can also be done online – also called streaming – as soon as new data arrives. A batch process queries the whole dataset whereas an online process operates on message-per-message basis. (Kiran et al. 2015, Marz & Warren 2015 and Vanhove et al. 2016.) Batch processing enables handling larger data sets but introduces latency into the system (Marz & Warren 2015 and Vanhove et al. 2016).

3.1.2 Overview of current concepts in data management

Big data is not any specific form of data. On the contrary, big data refers to all the different formats of data that are produced – and need to be utilised – whose volume increases rapidly. Hashem et al. (2015) consider big data as data that is diverse, complex, and extremely sizeable. They also consider the techniques needed to operate this data in their definition. Originally Gartner developed the "three Vs" to characterise big data, namely volume, variety, and velocity (the speed of data transfer) (Hashem et al. 2015). The three Vs mean that current data processing and storage solution will not work in the world of big data. For example, the unstructured nature of big data, i.e., the variability, means that the data cannot be stored in relational databases. (Bansal & Kagemann 2015.)

Internet of Things (IoT) is a term introduced by Kevin Ashton in 1999 (Ashton 2009, Ravulavaru 2018, p. 7). By the – quite broad definition – given by Ravulavaru (2018, p. 7), internet of things means the trend where electronic, electrical, and mechanical objects start to communicate with each other. The cost of sensors as well as data storage and analysis has declined rapidly (World Economic Forum 2016). As a result, the number of IoT devices is estimated to reach 125B in 2030 (Mittal et al. 2018).

By the NIST definition of cloud computing (Mell & Grance 2011), cloud computing refers to the use of pooled computing resources that are provisioned and released dynamically among users. The reallocation process of resources is automated to a great extent (Mell & Grance 2011). The NIST definition offers four deployment models for the cloud, namely private, community, public, and hybrid cloud. Private and public being self-explanatory, community refers to a cloud that is used only by a specific group of organisations. These organisations usually have similar requirements such as those related to security and compliance. Hybrid cloud includes all possible compositions of the above. (Mell & Grance 2011.)

Clouds simplify system administration as the addition of new resources occurs automatically without a need of interference from the IT staff. The on-demand nature of the capacity also ensures optimal hardware utilisation. Hence, cloud infrastructure is cost effective for the organisation. (Marz & Warren 2015.)

Machine learning is the best-known subfield of artificial intelligence (AI) with the largest amount of commercial applications (Alaybeyi et al. 2020). Machine learning comprises of statistical methods that extract patterns and information from data. They main idea is that machine learning methods infer the – usually complex – relationships of inputs and outputs in contrast to traditional software engineering where the relationships are programmed explicitly. (Alaybeyi et al. 2020, Géron 2017 and Guido & Müller 2016.)

Supervised learning is the most utilised subdomain of machine learning (Alaybeyi et al. 2020, Guido & Müller 2016). In supervised learning the algorithm is given a set of inputs and outputs whose relationship it models. Unsupervised learning is a completely different task where the algorithm casts structure to the data, e.g., clusters of similar datapoints. (Alaybeyi et al. 2020, Géron 2017, Guido & Müller 2016.) Reinforcement learning, on the other hand, has few commercially viable use cases (Alaybeyi et al. 2020). It is based on the exploration of an environment by an agent whose actions are governed by a predetermined value function it tries to maximise given the feedback of its actions (Alaybeyi et al. 2020, Géron 2017.)

Blockchain is a technology concept introduced by (an) anonymous author(s) under the alias Satoshi Nakamoto in 2008. It forms the basis for an electronic payment system in which two parties can transact without any trusted third parties. It is a distributed timestamp server that proves the chronological order of transactions using cryptography. (Nakamoto 2008.)

In blockchain transactions are grouped as blocks and then verified by nodes in the network following a consensus procedure. These blocks form a chain where the hash values – used to validate a block – of all the successors of a given block are dependent on it. Thus, changing a block would require changing all its successors' hash values. The determination of these hash values is made CPU-expensive on purpose. (Musleh et al. 2019, Nakamoto 2008.) Hence, as long as honest nodes control a majority of the network's CPU capacity, disagreements can be settled by choosing the longest chain of valid blocks (Nakamoto 2008).

In addition to persistency and audibility of transactions, blockchain offers privacy as the cryptographic keys used are anonymous. This cannot guarantee complete privacy, however, as leaking the identity of the owner of one transaction could reveal other transactions made by her. (Nakamoto 2008, Zheng et al. 2017.)

3.1.3 The Lambda architecture

As noted in the subchapter 3.1.2, big data requires new data architectures. An example of this is the commonly used Lambda architecture introduced in Marz & Warren (2015). On the concept level, the idea of Lambda architecture is to combine online and batch processing into one framework. Thus, it can serve both low-latency and big data use cases. (Kiran et al. 2015, Marz & Warren 2015, Vanhove et al. 2016.)

The Lambda architecture has three layers. The lowest layer, the batch layer, queries the underlying raw data precomputing views – or query results – that are stored in the next layer, the servicing layer. The servicing layer is a specialised, distributed database that enables random reads on the views of the batch layer. Its structure is optimised for random reads and updates without support for random writes. This makes the database architecturally simple. The need to wait for the batch process to finish before new data is available introduces latency for the two-layer architecture. The final layer, the speed layer, however, can handle tasks where low latency is needed. It maintains a view that is updated as soon as new data arrives. Hence, it is able to provide low-latency service. Its drawbacks are the complex structure which makes it much more prone to errors, and the limited size of the input it can process. Thus, the speed layer is only used for tasks specifically demanding quick responses. The system is synchronised such that the speed layer discards the data as soon as it reaches the servicing layer. Thus, any errors made in the complex speed layer get revoked. (Kiran et al. 2015, Marz & Warren 2015, Vanhove et al. 2016.) Munshi & Mohamed (2018) implemented the Lambda architecture in the context of smart grids. They were able to handle large amounts of data including images and video showcasing the potential of the Lambda architecture.

3.2 The business impact of modern data solutions

McKinsey – based on their research – note the adoption of AI accelerating among companies. They find that for companies to be successful in AI they need a bundle of capabilities. Early adopters of AI have already nurtured their core and advanced technologies, and pool of talent needed to succeed in AI on a larger scale. Core technologies include cloud computing, and web and mobile technologies. Advanced technologies are big data capabilities and advanced analytics. (Bughin & Van Zeebroeck 2018.)

The Boston Consulting Group (BCG) (2020b) highlight three dimensions that organisations need to develop to incorporate AI into their business. First, end-to-end processes for creating AI solutions need to be established to truly harness AI. (BCG 2020b.) Without clear processes, AI methods are only used in point solutions without creating systemic capabilities for the organisation (Ransbotham et al. 2018). By 2018, only 10 percent of companies had sought to adopt AI in the company scale (Bughin & Van Zeebroeck 2018). In addition to tools and processes, efficient use of the AI talent in the organisation needs to be ensured by the organisational structure by BCG (2020b). AI expertise should be charted on the company level and pooled to support the AI development in the organisation especially on platform decisions, data governance, and cybersecurity. In addition, floor level AI experimentation should be encouraged. Third, people are the key dimension for successful AI implementation in the organisation. These are not only AI experts hired outside the organisation, but also existing staff reskilled to adopt AI into their work. (BCG 2020b.) Not surprisingly, companies with strong background on digital capabilities have been first to found success with AI (Bughin & Van Zeebroeck 2018).

First-movers in AI have started the adoption from cost-saving projects. They are regarded as simple to justify due to the easily pinpointed nature of cost savings. (Ransbotham et al. 2018.) Pioneers who have successfully pursued these projects, can build on these efficiency gains to release funds for successive AI-fuelled improvements. Thus, they are stacking up their lead. (Bughin & Van Zeebroeck 2018.) Especially, Chinese organisations have emphasis on cost-saving whereas the c-suite from other regions see revenue creation as the future of their AI endeavour. Elsewhere, the rationale behind favouring revenue growth projects is that they are regarded as passing onetime opportunities whereas efficiency improvements for internal processes can be postponed without much harm. (Ransbotham et al. 2018.)

The MIT Sloan research reveals that the gap between AI pioneers and the more hesitant is widening. The pioneer group – already invested in AI – has only picked up the pace by leading their peers when ranked by increase in their AI spending. Moreover, these organisations also feel to have strengthened their understanding about AI the most. But who are these pioneers? The pioneer group in the study was characterised by having the most developed data management capabilities including data lakes and company-wide data management systems. This only underlines the increased importance of data as the key resource for future – and current – organisations. Management support is important in fostering the first projects and enabling the organisational shift. Organisation also learn to pick the right AI projects as they get more experience. AI requires a dedicated AI strategy to systematically develop it throughout the organisation. Multiple cases in the Sloan research support the idea of blending the AI experts with business professionals and making them collaborate to identify the best use cases for AI-powered solutions. (Ransbotham et al. 2018.)

Organisations can approach AI adoption using the Digital Transformation Framework (DTF) by Matt et al. (2015). Planning digital transformation – as strategic planning in general – concerns defining the strategy elements and the resources allocated to pursue the chosen strategy. The DTF defines four dimensions which to consider in the planning process. The ability of the company to adopt new technologies forms the first dimension. (Matt et al. 2015.) Strong firms can seek to establish their technologies as industry standards. This can create persistent competitive advantage, e.g., through making other companies dependent on the technology or producing lock-in to customers. However, pursuing technology leadership is risky as success cannot be guaranteed. Adoption of new technology may lead to changes in value creation which is the second dimension. New technology offers avenues to create new offerings. Pursuing these new possibilities often requires new competences making it harder to execute for the organisation, and consequently, increasing the risk for these projects. (Matt et al. 2015, Porter 1985.)

Such changes also require structural changes, the third dimension of the DTF. New activities need to place within the corporate structure. Minor changes in the value creation logic may allow for the additional operations to be integrated into the existing corporate structure whereas larger ones may be better to separate into their own entities. The last of the four, the financial aspects, places the boundaries for the other three. Financials need to be carefully assessed before making the changes. On one hand, deteriorating financial situation justifies, demands even, changes to be made in the organisation. On the other hand, poor financial situation restrains the organisation from the resources it needs to implement changes. Successful creation of digital business requires all four dimensions to be well aligned. (Matt et al. 2015.)

Besides the DTF, managers can leverage the resource-based view developed by Barney (1991). It considers the resources of a firm as the source of its competitive advantage. Here resources include everything from physical assets to capabilities and processes a firm possesses. The value, rarity, inimitability, and non-substitutability of a resource determines whether it is a source of competitive advantage. Value means that resource can be used to neutralise threats or pursue opportunities in the firm's environment. Rarity means that the resource is not possessed by many in the marketplace.

For the advantage to be persistent – referred to as sustained competitive advantage – the resource giving the advantage has to be inimitable. This is because otherwise competition would be quick to copy the resource, no matter how valuable or rare it was. Barney (1991) finds three reasons for inimitability. First, the resource might be dependent on time. This view proposes that some resources, e.g., a firm culture induced by the values of the historical era present in the early years of the firm, can only be acquired at a specific point in time. Another reason is causal ambiguity which refers to a situation where nobody – also considering the management of the firm – knows what resources possessed by the firm give its competitive advantage. It is important to note that knowledge, even if only present in the firm, will eventually leak outside the organisation. Thus, the source of the advantage cannot be known to anyone for it to remain sustainable. Social complexity is the last reason given. It means that the resource is such a complex social phenomenon that it cannot be systematically managed by a firm. (Barney 1991.)

The final source of sustainable competitive advantage, the non-substitutability, means that there must not be strategically equivalent resources that are imitable or not rare. Barney (1991) identifies information management systems specifically as a potential source for sustained competitive advantage. Technology per-se is not identified as the source of advantage but the close interplay of personnel and machines.

3.3 Importance of data management in modern buildings

As discussed in section two, there are energy- and building-related concepts that address systems, not individual buildings. Hence, data management would need to be discussed on the system level as well. Such literature is regrettably limited, however, as illustrated by the fact that the query TITLE("demand response" AND "information management") in Scopus returns no articles.

Fortunately, Baek et al. (2014), present concrete conceptualisation and prototyping of data systems needed in a large-scale smart grid. Their discussion is on the country-level which exhibits unique problems arising from data volume and system complexity. The concept of Baek et al. relies on cloud computing – discussed in subchapter 3.1.2. They highlight scalability and resource utilisation as the main support for their choice. Their idea – on the conceptual level – is to handle the smart grid on three hierarchical levels. By their view, the data produced and broadcasted by IoT devices operating on the lowest level is gathered and processed in the regional and the "top" level utilising capacity from private and public clouds. The data from the regional level is aggregated and processed in the top level to offer a view over the whole system. The regional level data centres are in charge of initial data processing and controlling of the lower level IoT devices. Third parties would be allowed to provide software-as-a-service type of offerings. The system would also provide data-as-a-service, e.g., energy consumption statistics, for its users. Baek et al. propose a centralised information flow management service to coordinate the influx of vast amounts of information. They also emphasise cyber security as being critical for the system and provide a security scheme to enable communication.

The ETL process is also present in smart grids as by Baek et al. (2014) the gird's information management consists of information gathering, processing, and storage. Data integration is seen as a major obstacle given the number of different devices. Contrasting with the centralised view of Baek et al., Musleh et al. (2019) discuss the utilisation of blockchain in smart grids as the core framework enabling the information management. They refer to Strasser et al. (2014) according to who centralised management of smart grid participants transactions would be very costly and complex. Energy trading, charging of electric vehicles, and DER integration are highlighted as the most prominent applications enabled by blockchain. Blockchain technologies offer the security, anonymity, and robustness needed in these applications (Musleh et al. 2019). They also enable monitoring the grid’s state in real-time and the use of decentralised control centres (Su & Huang 2018).

4 Prediction

In this section prediction in buildings is discussed. First, a general introduction to prediction in the building context is given. Then a deeper dive into specific methods used in this study follows. Emphasis is placed on the gradient boosting machine, especially, as it is the method of choice in this study.

4.1 Introduction to energy consumption prediction in buildings

Prediction models can be divided into white, black, and grey box models. White box models are physics-based models primarily used in predicting heating-based energy consumption. The prediction is based on solving systems of equations describing heat transfer in a building. The name "white box" describes the deterministic nature of these models. As these models are based on principles of physics, the results are easier to interpret. (Foucquier et al. 2013, Li et al. 2014, Bourdeau et al. 2019.) Moreover, the impacts of the assumptions made regarding the system are assessable. On the other hand, as Li et al. (2014), Foucquier et al. (2013) and Pan & Zhang (2020) remark, white box techniques require detailed information about the building which limits its usability as this information rarely is available. Moreover, this renders these models poorly scalable as the model needs to be rerun for every building. The amount of computation needed is also large for such models.

In contrast to physics-based white box models, black box models are powered by statistical methods. These methods require little or even no information about the building characteristics but deduce relationships between given inputs and outputs. These inputs usually include building information, of course, as well as historical consumption, and environment data. The notion of "black box" characterises the difficulty of explaining and interpreting the results that often arises with these models. This is because the actual impact of a given input value is often hard to pinpoint. This is especially the case with artificial neural networks (ANN). Although, these models do not require such detailed information, they usually need heaps of it. Model selection and tuning are also on the emphasis when working with statistical models. (Bourdeau et al. 2019, Foucquier et al. 2013, Li et al. 2014.) These models also need to be retrained if they are to be used for buildings with different characteristics compared to the one the model was trained on. (Wang & Srinivasan 2017.)

In terms of the types of statistical models used for energy consumption prediction in buildings, both Ahmad et al. (2018) and Wang & Srinivasan (2017) found regression, artificial neural networks

(ANNs), and support vector machines being the most popular accounting for 26 %, 41 %, and 12 % of the studies reviewed, respectively. Ease of implementation favours ANNs and regression. Regression methods are also easier to interpret than ANNs but tend to preform worse. (Ahmad et al. 2018, Wang & Srinivasan 2017.)

Grey box, or hybrid models combine methods from both categories. A hybrid model can ease the demands on data quality and quantity posed by white and black box methods, respectively. For example, a statistical model can be used to estimate some physical parameters needed in the physics-based model. (Foucquier et al. 2013, Li et al. 2014.)

4.2 The decision tree -based methods

Decision trees are divide-and-conquer type machine learning methods used in classification and regression (Myles et al. 2004), also referred to as classification or regression trees based on use case (e.g., Friedman 2001, Loh 2011). Decision trees partition the feature space into smaller subspaces (Papadopoulos et al. 2018) by a set of if-then rules (Touzani et al. 2018). This partitioning is done iteratively as described in Morgan & Sonquist (1963), the first regression tree algorithm (Loh 2014). This rule-based structure helps in interpreting the input-output relationships in the model (Friedman 2001). Decision trees are nonparametric models meaning that their parameters are not determined prior to training. This gives them flexibility needed to capture complex dependencies but on the other hand makes them prone to overfitting. (Géron 2019.)

The algorithm by Morgan & Sonquist (1963) was developed for the analysis of survey data and works as follows. First the sample is divided into two groups based on a feature in the data. The division is done by estimating how much the error variance would be reduced by the segregation over all possible segregations. Then the split imposing the highest reduction in error variance is chosen. This division into smaller and smaller subgroups is continued to a point where the possible gain of splitting any group is estimated to be too low. Morgan & Sonquist proposed a two-percent threshold where the process is stopped when no subgroup accounts for more than two percent of the (sum of squares) error over the whole sample as a split is only done to a chosen subgroup.

A key weakness of the algorithm described is the need to scan through all subgroups for all splits over all features. Moreover, this exhaustive search is present even in many of the modern algorithms (Ke et al. 2017). Exhaustive search is problematic from efficiency standpoint. The computational complexity for the search is high, meaning that the number of computations increases quickly as the number of features and observations increases. In addition to efficiency concerns, this unrestrained search favours variables with more splits introducing a bias in the feature selection process. This hinders the ability to reliably deduce input-output relationships by examining the tree structure. (Loh & Shih 1997.)

4.3 Ensemble models

Ensemble model is a term used for a model combining multiple "learners" which are usually simple models (Touzani et al. 2018) such as decision trees (Papadopoulos et al. 2018). The predictions of these learners are combined to form the final prediction of the model (Drucker 1997). The idea is that an ensemble of simple models is able to overperform a single, more complex model. For example, in the context of decision trees, Friedman (2001) notes that averaging over many smaller trees effectively combats inaccuracy and instability, the common drawbacks of individual small and large trees, respectively. What is more, decision tree -based ensemble models can handle categorical inputs without pre-processing and are robust to outliers and missing data. (Friedman 2001, Touzani et al. 2018.) These algorithms are also computationally efficient event with large amounts of data (Friedman 2001, Touzani et al. 2018) in contrast to support vector machines – a popular machine learning method – for instance (Pan & Zhang 2020).

Ensemble models can be divided into homo- and heterogenous models. Homogenous models are further divisible by the dependence between the learners. Learners being independent enables their training to be parallelized. (Wang et al. 2019.) One of the advantages ensemble models have over traditional black box machine learning models is their interpretability. As ensemble models are a combination of usually simple learners, the importance of different features in the data can be described. (Pan & Zhang 2020.) In the case of gradient boosting machines – discussed in more detail later on – Friedman (2001) highlights the relative influences of individual inputs on the variation of the model over all inputs as means of describing the dependencies captured by the model. The exact influence metrics vary by the software used, however. In addition to influence metrics, Friedman proposes the use of partial dependence plots where the predicted value of the model is depicted against values of a given feature regarding others features constant. Only low-level dependencies can be visualised, however, as the interplay of multiple features does not get captured when changing the values of only one feature at a time.

Bagging and boosting are both methods used to combine learners to form ensemble models. In bootstrap aggregating, i.e., bagging, a subset of the training data is chosen randomly to train each of the learners. These differing inputs will result in differing predictions between the learners, hence, strengthening the overall prediction. As the sampling of the training data for the learner is done randomly, the learners are independent of each other and, thus, can be trained in parallel. (Breiman 1996, Drucker 1997.) In the case of boosting, the learners are dependent. This is because after the first learner is trained similar to bagging, the entire training set is run through this one-learner model. Next, the probability of the observations to be included in the training of the next learner is determined by the prediction error from the one-learner model. Hence, training the next learner focuses on improving the cases where the former model did poorly. This process is continued iteratively to create an ensemble of learners focused on the errors of their predecessors. (Drucker 1997.) Hence, boosting reduces the bias, or systematic error, of the model resulting in shallow, dependent trees who influence the final prediction disproportionately. (Touzani et al. 2018.) The weights controlling the influence of different trees, i.e., the composite model, are learned in model training as discussed later in detail. Boosting has been experimentally shown to be superior to bagging (Freund & Schapire 1996).

4.4 Gradient boosting machines

Gradient boosting machines, or GBMs, are boosting-based ensemble learning algorithms combining gradient descent and boosting. Gradient descent is a numerical minimization method where gradient of a chosen criterion – commonly referred to as the loss function – is calculated with respect to the target function, e.g., a prediction model. Using this gradient, the parameters of the target function are updated by subtracting the gradient, hence, “stepping” toward the minima of the loss function. (Friedman 2001.)

Gradient boosting can be represented as the following algorithm (Friedman 2001):

1. $F_0(x) = argmin_p \sum_{\mathrm{i}=1}^{\mathrm{N}} L(y_i, p)$
2. $For\ m = 1\ to\ M\ do$:
3. $\tilde{y}_i = -\left[\frac{\delta \mathrm{L}(\mathrm{y}_\mathrm{i}, F(x_i))}{\delta \mathrm{F}(\mathrm{x}_\mathrm{i})}\right]_{F(x)=F_{m-1}(x)}, i = 1, N$
4. $a_m = argmin_{a,\beta} \sum_{\mathrm{i}=1}^{\mathrm{N}} [\tilde{y}_i - \beta \mathrm{h}(x_\mathrm{i}; a)]^2$

```
5. $p_m = argmin_p \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + ph(x_i; a_m))$
6. $F_m(x) = F_{m-1}(x) + p_m h(x; a_m)$
7. $End\ for$
```

In the algorithm, one starts by obtaining a value *p*, the initial guess, who minimizes the loss $L(y_i, p)$ over all observations using a line-search (1). This initial prediction is improved in "boosting" iterations *M* times (2). The boosting round starts by evaluating the gradient of the loss function with respect to the prediction given by the function *F*, for every point in the data (3). Next, a weak-learner – such as a decision tree – is trained such that it most highly correlates with the negative gradients obtained earlier. This is achieved by minimizing the squared error – used by the metric of difference – between the gradients and the decision tree (4). Then a new line-search is performed to learn the parameter p_m in a way that the loss over the new model – obtained by adding the new weak-learner (6) – is minimised. The learning is done stagewise meaning that parameters in F_{m-1} are unchanged. (5.) (Friedman 2001.)

Gradient boosting decision trees are widely used in a plethora of machine learning tasks. They are regarded as being accurate, efficient, and interpretable. (Ke et al. 2017.) These algorithms have their weaknesses, however. Looking at the algorithm described in Friedman (2001), one can see that the problem of exhaustive search in the training of $h(x_i; a)$ – in the case where $h(x_i; a)$ is a decision tree – isn't addressed. When handling big data, this bottleneck becomes a problem for these algorithms such as commonly used XGBoost and pGBRT (Ke et al. 2017).

LightGBM is a fairly recent algorithm developed to combat this. LightGBM combines two novel strategies, *Gradient-based One-Side Sampling (GOSS)* and *Exclusive Feature Bundling* to help decrease the amount of computation needed in the decision tree training. The information gain from a given observation depends on the absolute value of the gradient in this data point. Thus, the training is improved by selecting only a subset of the data based on the gradients. However, including only observations with large gradients changes the data distribution resulting in lower accuracy. GOSS counters this by first sampling a subset of data among the lower valued gradients and then introduces a weight to increase the importance of these observations. Thus, training focuses on under-trained samples but preserves the data distribution.

In exclusive feature bundling, the features are compressed to *exclusive feature bundles* meaning a single composite feature that captures the information of multiple features. Hence the dimensionality of the data can be reduced. The ability to compress the features results from the notion that high dimensionality of the data results in a sparse feature space meaning that many features are mutually exclusive and thus can be bundled. LightGBM is able to reach speed improvements up to 20 times sacrificing little accuracy. (Ke et al. 2017.)

4.5 The role of prediction as enabler of the future

Prediction methods are an important part of many building and data systems. For example, Krishnadas & Kiprakis (2020) use machine learning empowered prediction for DR capacity scheduling. Sequeira et al. (2014) develop an anomaly detection system for their proof-of-concept energy cloud implementation. Newaz et al. (2014) highlight accurate demand forecasting as a key research area in smart grid research. As the examples given demonstrate, research in prediction models is needed to unlock the potential of a variety of energy-focused applications for buildings. This work illustrates the development of such a model.

As discussed in section three, big data poses new challenges for data management. Machine learning methods can be utilised to overcome these challenges. For example, deep learning methods have proved effective in handling the variety of big data. These methods have been used to integrate data from heterogenous data formats. (Qiu et al. 2016.) On the other hand, many of the challenges posed by big data also have implications for prediction models. For instance, many machine learning algorithms have been designed with the assumption that the entire data set is held in memory. Naturally, this becomes inappropriate with big data. (L'heureux et al. 2017.) Thus, prediction algorithms that cater for the characteristics of big data need to be developed. For example, algorithms that support parallel programming have been successful in handling big data (Qiu et al. 2016).

5 Methodology

This section introduces the reader to the research methodology used to conduct the work. More specifically, design science and systematic literature review are introduced. Also, the reasons to choose these methods are discussed.

5.1 Design Science

Design Science (DS) or Design Science Methodology is a practice-oriented way of conducting research. It focuses on manmade artefacts in contrast to natural phenomena. Solving problems is in the essence of DS. Thus, the utility of the solution is pivotal when considering the quality of the research. (Dresch et al. 2015.)

Hevner et al. (2004) defines seven criteria for guiding DS research. The research must produce an artefact of a functional form. This derives from the emphasis on solving a concrete problem. The problem solved also needs to be considered relevant. The solution needs to solve the specified problem demonstrated by rigorous evaluation. The research also needs to contribute to the scientific areas relevant to the artefact. The research needs to be conducted with academic validity and rigour. The research process should respect the iterative, even experimental, nature of design science with emphasis on functionality and utility over optimality. The findings need to be communicated in a way that is approachable for both technology- as well as management-oriented audiences. (Dresch et al. 2015.)

Information science is the domain where DS approach has historically been utilized the most (Dresch et al. 2015, Hevner et al. 2004, Peffers et al. 2007). It is also suitable for both academic and organisational contexts (Dresch et al. 2015). Hence, DS is a suitable framework to guide the research process, as the objective is to develop a statistical model and a software tool to solve an actual problem. What is more, the current work is conducted for a commercial party who values the useability and customisability of the artefact.

The DS process used in the thesis corresponds to the outline given in Dresch et al. (2015). First, the problem is identified and defined after which the awareness of the problem is established. Awareness of the problem governs all aspects of the problem and extensive understanding. In addition to the problem, the functionalities and requirements of the artefact need to be charted. Literature review is an important step for establishing the level of understanding discussed (Dresch et al. 2015, Thornhill

2009). Additional sources of information should also be utilized to minimize potential biases related to different sources of information (Dresch et al. 2015).

After awareness of the problem has been established, the potential artefacts suitable for addressing the problem need to be identified. The process for creating the pool of potential solutions in governed by the awareness of the requirements for the artefact. This underlines the importance of the knowledge gathering. The set of potential solutions needs to be then analysed further and the options weighed against each other to select the most lucrative artefact to pursue. (Dresch et al. 2015.)

After the artefact is produced, it needs to be evaluated against the identified requirements to determine if it satisfies the problem. The evaluation can be performed in the actual or in an experimental environment. The evaluation can lead to the artefact being rejected as not satisfiable in which case a new artefact needs to be developed. This iteration is enabled by the already identified set of potential solutions. The learnings from the development and evaluation process can also be used to narrow down the set of potential artefacts. Thus, the learning should be a deliberate step in the process. (Dresch et al. 2015.)

After an artefact passes the evaluation, it is concluded as the solution. In the conclusion, the results achieved, and the decisions made need to be documented and formalized explicitly. Also, the limitations need to be considered. Even though the artefact has been developed for a specific use case, it needs to be next generalized for a class of problems. Hence, the research is able to advance science, a demand of DS. Last, the results need to be communicated. Effective communication is vital in order to harness the utility of the artefact. Afterall, utility is pivotal when considering the quality of DS research. (Dresch et al. 2015.)

5.2 Systematic literature review

Systematic literature review is an important step in an academic research process (Seuring & Gold 2012) and as such is also included in the DS methodology. Its purpose is to compile a holistic picture about what is already known in the literature about the research problem the current research is trying to address. (Dresch et al. 2015). It advances the understanding of the researchers about their research subject and helps in grounding and interpreting the results of their study (Dresch et al. 2015, Seuring & Gold 2012).

Kugley et al. (2016) note that it is usually not desirable to address all sides of the actual research question in the literature review. This implies the need to formulate a separate question, the review question, which governs the content, breadth, and depth of the literature review. In DS the literature review aims to form a framework through which to address the artefacts and the problems discussed in the research. (Dresch et al. 2015).

Systematic literature review is conducted to build an understanding of the problem and potential solutions. The review question to be answered in this literature review is “What statistical prediction methods have been used to forecast energy consumption in buildings that are applicable to northern climate?”. Hence, the objective is to build thorough understanding of the usage characteristics and requirements of statistical forecasting methods related to energy consumption in buildings excluding studies not applicable to colder, less humid climates.

The research strategy focuses on the use of the Scopus bibliographic database as well as the bibliographic search engine Google Scholar. The search involved the following search phrases and controlled vocabulary expressions; "(TITLE-ABS-KEY ("energy prediction") AND KEY (buildings)) AND PUBYEAR > 1999", “(TITLE-ABS-KEY ("gradient boosting" AND energy) AND KEY (buildings))”. All the articles are listed in the appendices one and two for excluded and included articles, respectively. Initial screening was done to the articles found using the abstracts and titles as instructed by Dresch et al. (2015) and Kugley et al. (2016). Hence, 189 articles in total were excluded due to not found relevant in the initial screening or not being accessible. The inclusion criteria used was the following:

1. Focus on energy consumption
2. Use of a statistical forecasting method
3. Applicable to dry and cold climate type

16 articles from the total of 205 were included in further analysis. The articles found relevant were codified as discussed in Seuring & Gold (2012) and Dresch et al. (2015). The category tags used were AI (artificial intelligence), STAT (deterministic statistical method) derived from the type of models used as well as the prediction objective-based HEAT (heating), ELEC (electricity), CONS (consumption), LOAD (load). These tags are also present in appendix 2. Artificial intelligence -based methods were found to dominate the literature. Also, electricity was emphasised over heat. The

distinction between load and consumption forecasting was not found useful as the concepts are so closely related.

In addition to secondary literature sources, primary expert sources were used in a form of a workshop. This helps to reduce the potential bias emerging from the use of academic literature only (Dresch et al. 2015, Kugley et al. 2016).

5.3 Semi-structured interview

Interviews are commonly used methods to acquire qualitative data. Interviews can be structured, semi-structured, or unstructured. Semi-structured interviews – used in this work for model validation – have a format that is planned beforehand, usually as a set of open-ended questions. These questions are not supposed to constrain the interview, however, but to facilitate interaction. Thus, some questions can be neglected during the interview and some can be added on the spot. (DiCicco-Bloom & Crabtree 2006, Myers & Newman 2007.) Semi-structured interviews are usually scheduled and located outside of everyday events (DiCicco-Bloom & Crabtree 2006).

The semi-structured interview method is used in the thesis as it best fits the purpose of model validation. This is because limited time requires the use of a pre-planned question set to ensure that all relevant aspects are covered. On the other hand, the interviewees being experts in their field, they undoubtedly have points of view that are highly relevant but unknown to the interviewer. The semi-structured nature enables discussing these viewpoints.

By DiCicco-Bloom & Crabtree's (2006) view, different stages can be observed during a semi-structured interview. First, the goal of the interviewer is only to get the interviewee talking. This is to make the interviewee feel comfortable and more relaxed. Broad and open-ended questions should be used. This builds rapport between the interviewer and the interviewee, essential in order to get the interviewee open up on more dire issues. Next, interview progresses to exploration phase where more in-depth questions are used to advance interviewer's understanding on the subject. Final stage is called co-operation stage where the actors collaborate to make sense of the interviewee's world. To reach this final stage may require multiple interviews.

Myers & Newman (2007) discuss common problems and pitfalls present in semi-structured interview. They remind that interviews are by nature artificial events as interviewer and interviewee are usually complete strangers. Still, the interviewee has to form opinions under time pressure. Moreover, she can lack the trust required discuss matters she regards sensitive. This may result in uncomplete or unreliable results. Interviewer cannot also be considered entirely neutral but has effect on the data acquired through wording and other, even unconscious, signals. Interviewee may also hesitate to voice opinions she is not sure of not to appear unknowledgeable.

By Myers & Newman (2007), interviewer needs to situate herself before the interview. This means acknowledging her position and background with respect to the interviewee. Interviewer can try to match the interviewee with language and dressing to minimise the social dissonance. The interviewer needs to understand that she is constantly interpreting the input of the interviewee, not just objectively absorbing it Myers & Newman (2007). The interviewer should mirror the answers of the interviewee when she would like further elaboration. This is to avoid leading the interviewee. Confidentiality of disclosures needs to be discussed explicitly and potential recordings destroyed after the research process. (DiCicco-Bloom & Crabtree 2006, Myers & Newman 2007.)

The author had been in contact with some of the interviewees before the research process. On one hand, this existing rapport enabled advancing – referring to (DiCicco-Bloom & Crabtree 2006) – to later stages of the interview more quickly which sped information gathering. On the other hand, this may have influenced the expression of negative aspects. To counter this, the interviewees were asked to speak freely. What is more, their names being known inhibits them from sugar-coating their responses not to seem incompetent. With the interviewees not known beforehand, a single one-hour interview may have been too short to reach co-operation. This may have limited the information gathering. Special attention was payed to phrasing the questions in a way as objective as possible. Sentiment of the interviewees was also asked at the beginning of the interview to uncover potential dispositions regarding the model that might have affected the opinions voiced. Explicit consent was asked for recording the interviews. The recordings were destroyed after use. The template used in the model validation interviews can be seen appended in appendix 3.

6 Problem definition and context

The problem the work tries to address, is discussed in the section with available data sources. Let us start by defining the problem for which the research is conducted. Then the possible data sources available are charted. The section is concluded by devising a set of concrete requirements against which the model can be assessed.

6.1 Problem definition

The identification of the need for an energy consumption prediction model surfaced in internal discussions with Granlund Consulting personnel who have identified the client need implicitly and explicitly. Granlund Consulting experts pointed out a potential use case for prediction being cost optimization regarding energy purchasing. In addition, real estate owners have interest in their emissions as their environmental impact is under scrutiny from regulators and the public. As energy consumption is a major source of emissions for these actors, prediction is a valuable tool for them to manage the "carbon risk". (Please, see workshop documentation in appendix 4.)

The objective of the work is the development of a prediction model. Referring to the data processing pipeline discussed in the section three, the focus of this work will be on the data analysis part excluding any ETL steps required to prepare the data. This is because Granlund already has the ETL process in place. Model deployment into operational use is also excluded. The reason for this choice of scope is the considerable workload needed to first develop a functional model and then configure the whole pipeline needed to utilise it. An initial plan for deployment is drafted, however, and discussed in chapter 10.6. The choice of scope with respect to a general data processing pipeline is illustrated in figure 11.

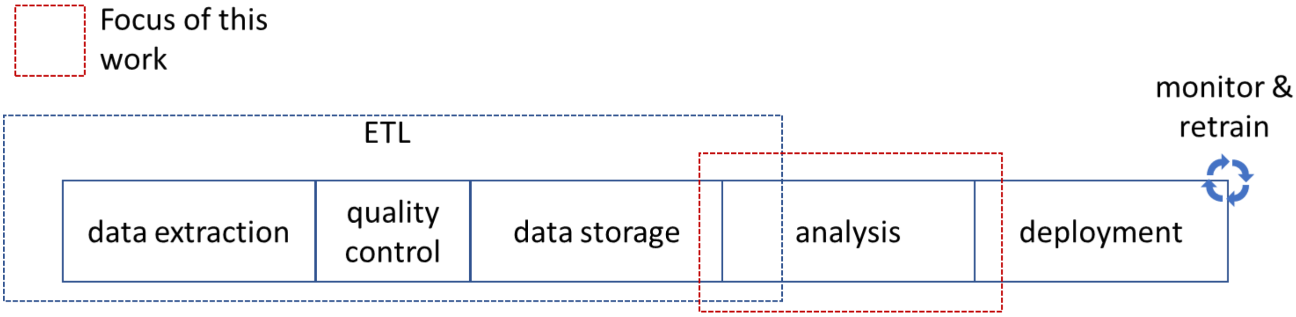


Figure 11. Data processing pipeline with the choice of scope.

In terms of a specific use case, anomaly detection arose as the direst problem solvable by prediction in the discussions with Granlund. Granlund monitors the energy consumption in its clients' building stocks and alerts them about discrepancies in the consumption which can be indications of system failures or misuse. In this case, anomaly is defined as a rapid, clearly differing incidence that progresses quickly over a couple of hours. In the organisation, anomaly detection work is still done by human eye pairs using simple heuristics such as constant percentage deviations from last month or same month last year. This is a source of considerable expense for the company. What is more, such monotonous work is prone for human error. This is the primary use case to be addressed in this work.

Workshop methodology was utilized in defining the identified problem (please, see appendix 4 for workshop documentation). This corresponds to Dresch et al. (2015) who emphasize that the researcher utilizes multiple sources of information to form a wide perspective about the problem and the requirements for the artefact. The solution requirements were explicitly addressed in the workshop.

Additionally, however, Granlund is interested in the use of new statistical methods more broadly. Hence, they hope the model to be versatile-enough that it can be customised to other tasks as well.

6.2 Data sources

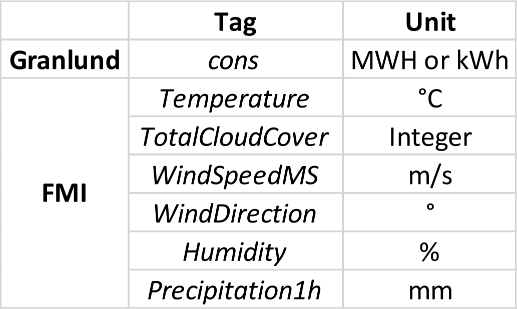
Energy consumption data can be divided into real, simulated, and benchmark data. Sources for real data – the focus of this work – include metering, energy bills, and onsite surveys. (Bourdeau et al. 2019.) Wang & Srinivasan (2017) classified the input data used in AI-based prediction models into meteorological, occupancy, and "other" data. Bourdeau et al. (2019) considered other methods of prediction as well and identified indoor conditions, operation of building systems, and building characteristics, in addition to the former.

The focus of this work is real meteorological and consumption data. The data accessible at Granlund consists of electricity and heat consumption data gathered hourly either by smart meters or is provided by the energy companies of Granlund's clients. The data processing at Granlund results in latency up to 24 hours. In addition to the consumption time series data, the Finnish Meteorological Institute (FMI) offers both past and forecasted weather data. The data is openly accessible and machine readable through their application programming interface. Given the forecasts, the outdoor conditions can be utilised to increase the prediction accuracy. The Python scripts used to access the data can be accessed in the author's Github (github.com/HVKukkonen). These tools have also standalone value as they enable automated weather and weather forecast querying from the FMI's service.

The consumption dataset used in the study consists of district heating and electricity consumption data gathered from non-residential buildings located in the Finnish city of Turku. The data consists of hourly consumption readings over a timespan of two years. The data is sourced by energy companies and does not include submetering data. In the dataset there are unique identifiers for different buildings (tag *ObjectID*) and meters (tag *MeterID*), consumption value (tag *cons*) in MWh for heat and kWh for electricity as well as timestamp (tag *Time*) for when the measurement took place. The heating dataset consists of 720 k rows gathered by 29 main energy meters from 20 buildings. The electricity data has 990 k rows from 31 buildings sourced from 35 meters.

The weather data from FMI has hourly temperature (tag *Temperature*, unit °C), cloud cover (tag *TotalCloudCover*, unit integers 0 - 9), wind speed (tag *WindSpeedMS*, unit m/s), wind direction (tag *WindDirection*, unit °), humidity (tag *Humidity*, unit %) and precipitation (tag *Precipitation1h*, unit mm) measurements. This data is available 48 h into the future. All input data sources used in the model can be seen in table I.

Table I. Model input data.



	Tag	**Unit**
Granlund	*cons*	MWH or kWh
FMI	*Temperature*	°C
	TotalCloudCover	Integer
	WindSpeedMS	m/s
	WindDirection	°
	Humidity	%
	Precipitation1h	mm

6.3 Solution requirements

The use case affects the characteristics of the model. Short-term forecasts – usually regarded as ranging from minutes to 24 hours – are needed for the real-time management of building energy systems (Fan et al. 2019), for instance, for selling and buying power (Ahmad et al. 2018) or pursuing energy conservation (Fan et al. 2019), whereas long-term predictions are useful for planning purposes, e.g., in urban district heating system projects (Koschwitz et al. 2020).

In discussions with Granlund, 24 hours was chosen as the primary prediction horizon for the model. This is because Granlund's internal data processing introduces latency up to 24 hours. 24 hours can be considered suboptimal for pure anomaly detection, however. This is because when assessing a given hourly consumption, the past hours' measurements are already known. Thus, forecasting for anomaly detection could, in principle, use only one-hour horizon. However, Granlund sees value in the possibility to expand the horizon to timespans up to 10 days.

The model should be flexible to easily accommodate the change of forecasting horizon. The model chosen should also be robust to outliers and missing values present in both Granlund's and FMI's data streams. As Granlund is interested possibilities beyond anomaly detection, the model is assessed using the 24-hour horizon. Naturally, the prediction task is easier the shorter the horizon. Hence, the assessment with 24-hour horizon offers a lower-bound estimate of the model's performance for shorter time spans. As a key yardstick, if the prediction accuracy for the 24-hour horizon is sufficient to be used in anomaly detection, the model overall is considered to be sufficient. The assessment of model performance against the forecasting horizon would require training multiple models and is consequently excluded from the scope of this work. The development of prediction models for both electricity and heat, with a performance that is 24 hours into the future sufficient for anomaly detection is the research problem of this work.

7 Initial data analysis

Initial data analysis is a pre-modelling step in the data analysis process. It is concerned with ensuring the data being accurate and complete. (Huebner et al. 2016 & 2018, Smitha Rao et al. 2019.) Huebner et al. 2016 & 2018 emphasize that it should be conducted strictly independent of the research questions. This is to avoid false positive results. The aim is to achieve full awareness of the data properties, a prerequisite for correct analysis and result interpretation (Huebner et al. 2016). There is no standardized procedure (Huebner et al. 2018, Smitha Rao et al. 2019). However, Smitha Rao et al. (2019) provide a three-step process for initial data analysis consisting of initial data diagnostics, data cleansing and data transformation.

In the section, the characteristics and quality of the data are assessed. Thorough examination is done for both weather and consumption data. Section is concluded by showcasing the cleansing and feature engineering steps taken to refine the data for the model.

7.1 Initial data diagnostics

In initial data diagnostics, patterns in the data are identified. This step is concerned with the identification of errors, outliers, and missing values in the data. Summary statistics and data visualizations are common tools used in this process. (Smitha Rao et al. 2019.) An additional objective is to understand the properties of the data and test researchers' expectations, e.g., regarding its distribution. This understanding is needed to validate the applicability of chosen statistical methods and assess reliability of results. (Huebner et al. 2018.)

Let us start with the energy consumption data. There exist quite many duplicates. These are a consequence of the fact that some meters affect multiple buildings which results in duplication in the Granlund's database. The shares of duplicates are 47% and 43% for the heating and electricity datasets, respectively. This results in quite considerable shrinkage of the dataset as the duplicates are removed, leaving us with 390 k rows of heat and 570 k rows of electricity consumption data. Neither one of the datasets contains any missing values. Assessing the data quality visually, however, one can notice some meters taking only discrete values (please, note figure 12). This is a problem with data granularity, probably resulting from rounding or measurement inaccuracy. Such abnormalities need to be removed as these meters behave systematically different from the rest of the meters. Thus, keeping such meters would hinder the training. A separate model would need to be developed in order to account for these meters which is excluded from the scope of this study. The behaviour discussed is much more common in heat meters than in electricity. To quantify this phenomenon, a test period is chosen and differences of values between every time step pair are calculated. Next, zero differences are counted meter-wise. Zeros accounting for more than half the differences over the period is used as the indication of the behaviour discussed. The idea is that consecutive same values result in lots of differences being zero. Using this cut-off, one arrives at 230 k rows for heat and 560 k rows for electricity passing the half zeros rule. The procedure also removes faulty meters as they usually produce constant zero measurement.

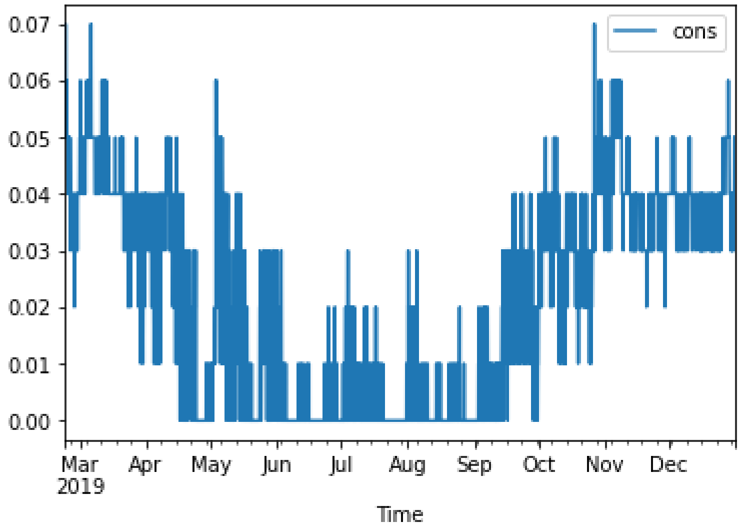


Figure 12. Successive duplicate values.

Before analysing consumption data further, the environment of the buildings should be considered. This can be done through the weather data. In the weather dataset there are missing values for all measured variables. The share of rows with at least one missing value is quite small, however – 0,5 % to be precise. There are correct amounts of unique values for cloud cover, i.e., 10, and wind direction. Minimum, maximum, and mean values are reasonable throughout the data set as well. All in all, the data seems to be of high quality. As temperature is the most important driver for heating energy consumption, it can be assessed visually in figure 13. The visual analysis does not indicate errors. There is quite large intra-day variation noticeable, however.

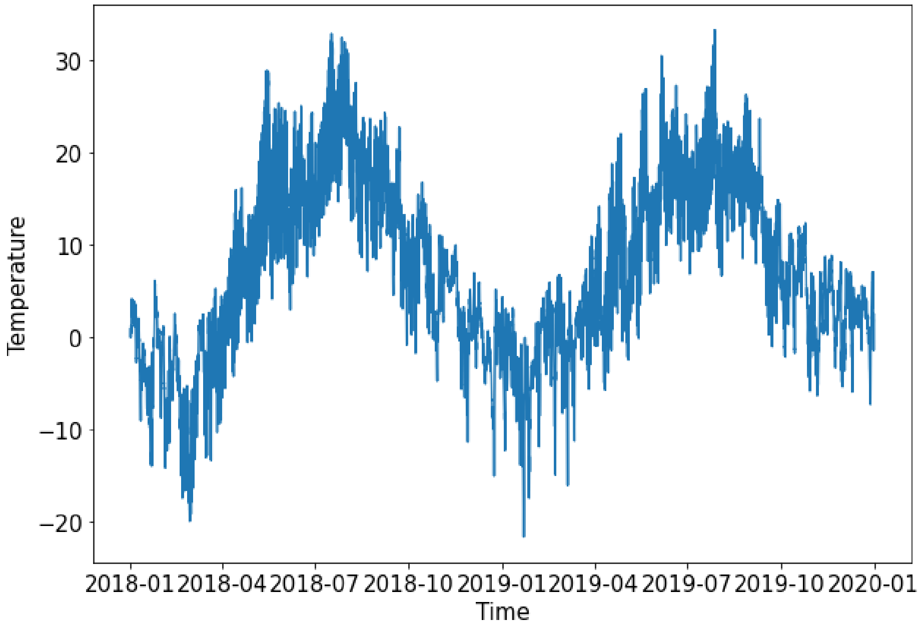


Figure 13. Hourly ambient temperature of Turku.

Coming back to the consumption data, there is great amount of variance noticeable in the heat consumption data (please, note figure 14). Oddly, the winter 2018 seems to have much higher peak consumption than the following measurement. Moreover, coming back to the temperature graph, no difference between the coldest days of 2019 and 2018 can be observed. In search of the reason for the odd behaviour, the distributions of heat consumption data per meter are visualised in figure 15. Only one of the meters – the meter with ID 15305 – seems to account for the maximum values. Given that the meter does not show similar behaviour in the following years even with similar conditions and that the magnitude change is so large, the meter is excluded from the data as faulty. Removing only this single meter results in much smoother profile as seen in figure 16.

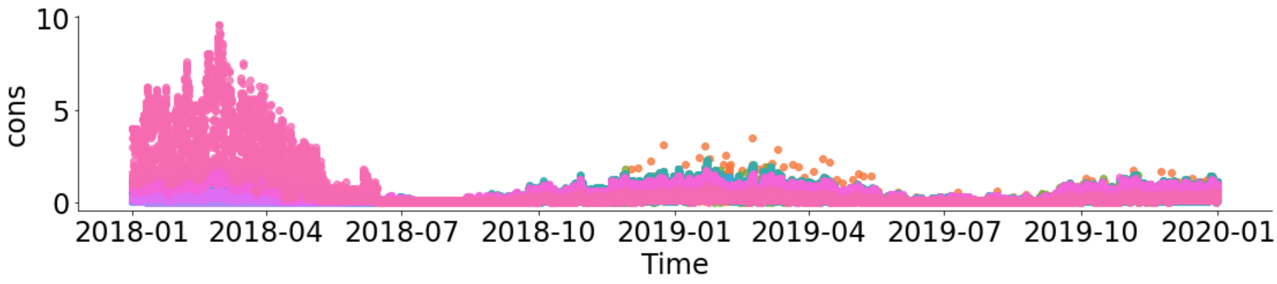


Figure 14. Heat energy consumption as a function of time with meters separated as colours.

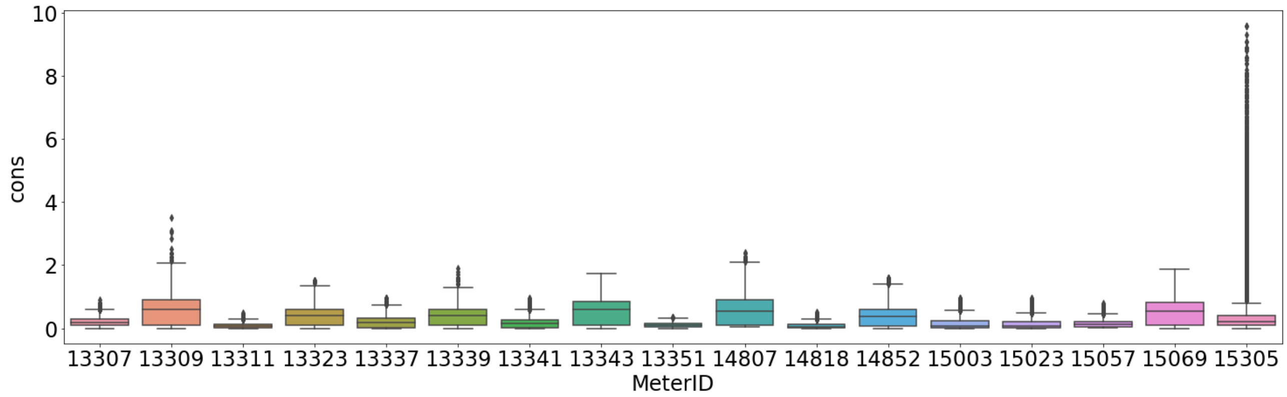


Figure 15. Distribution of heat data between meters.

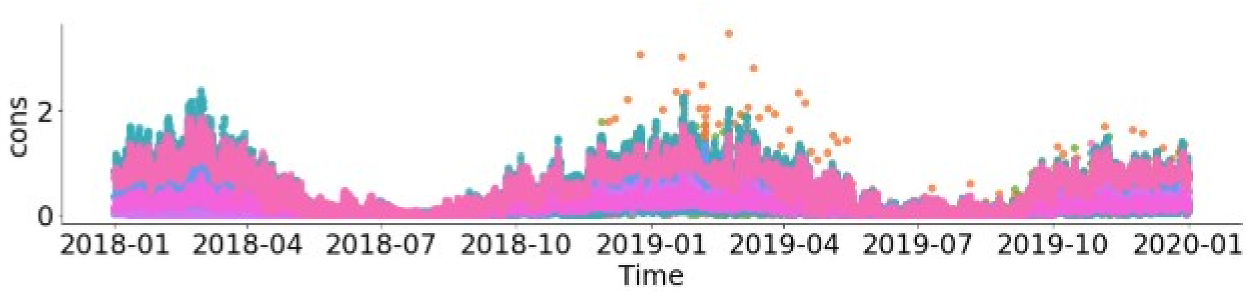


Figure 16. Heat energy consumption as a function of time without the meter 15305.

To quantify, the standard deviation for the heat consumption data is 0,32 MWh and the mean hourly consumption is 0,28 MWh. The maximum value is 3,5 MWh and the 75th percentile is at 0,41 MWH. Clear seasonality is present in heating energy demand due to seasons of the year. In addition, there are multiple meters measuring different buildings which introduces variance. The values in figure 16 also seem continuous without apparent outliers. The scale of the values also seems plausible for large non-residential buildings.

The picture is more tranquil with electricity consumption. The mean for the hourly electricity consumption is 123 kWh with the maximum of 999 kWh and a standard deviation of 152 kWh. The figure 17 tells the same story visually. However, there are larger differences in the mean values between meters as represented by horizontal lines in the box plot in figure 18. In other words, some buildings consume much more electricity than others although their heating energy consumption is much more on par. Thus, mean consumption level is an important feature especially for the electricity prediction model.

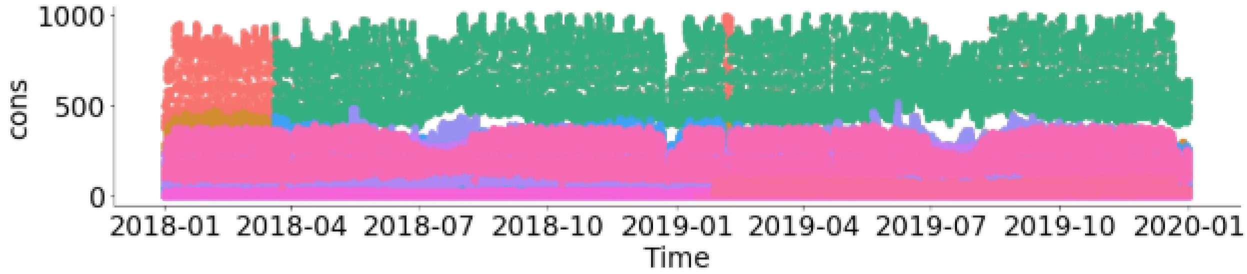


Figure 17. Electricity consumption as a function of time with meters separated as colours.

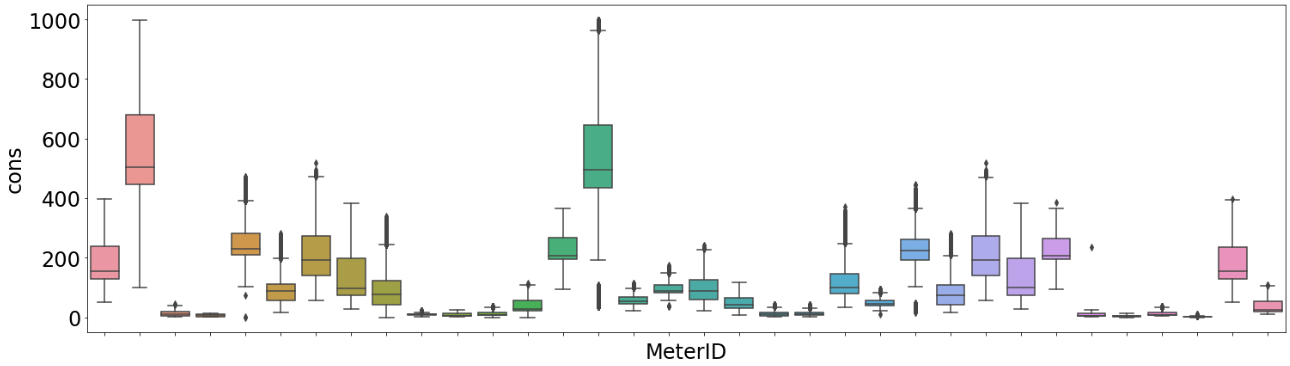


Figure 18. Distribution of electricity data between meters.

Diving deeper into the fabric of the data, seasonality is an interesting question. Based on the analysis discussed above, there is intra-year seasonality in the heat consumption with peaks in the winter and valleys in the summer. However, one should find intra-week and intra-day seasonality as well as surely more energy is desired during the working hours.

In figure 19, the hourly heat consumption patterns for different days is presented. Mornings see the peaks in consumption as the building needs to be heated up for the day whereas only a base level needs to be maintained in the evenings. Also, the intra-week seasonality is apparent with much more steady profiles in the weekends. The 90 % confidence intervals for hourly mean consumption form the bands in the figure. Hence, mean consumption indeed changes statistically significantly based on time of day which is quite rational. One needs to be mindful of the fluctuation caused by the comparison of different meters and, thus, buildings, however.

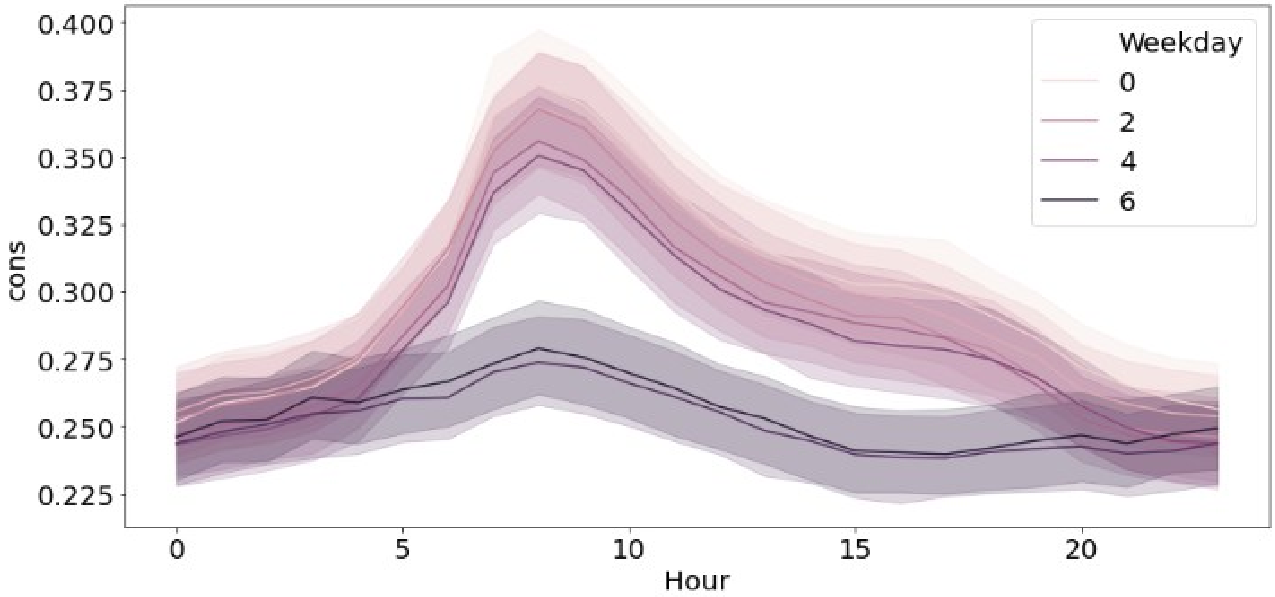


Figure 19. Hourly heating energy consumption by day with confidence intervals.

Similar profile is present with electricity as illustrated in the figure 20. The electricity consumption is a bit more rounded, though, with peaks a few hours later at noon. It seems that electricity goes more hand-in-hand with occupancy rates, whereas heat energy is needed as a shorter burst in the mornings to heat up the building and can then be toned down with the building structures transmitting the heat back throughout the afternoon. Without occupancy data, however, this hypothesis cannot be tested in this study.

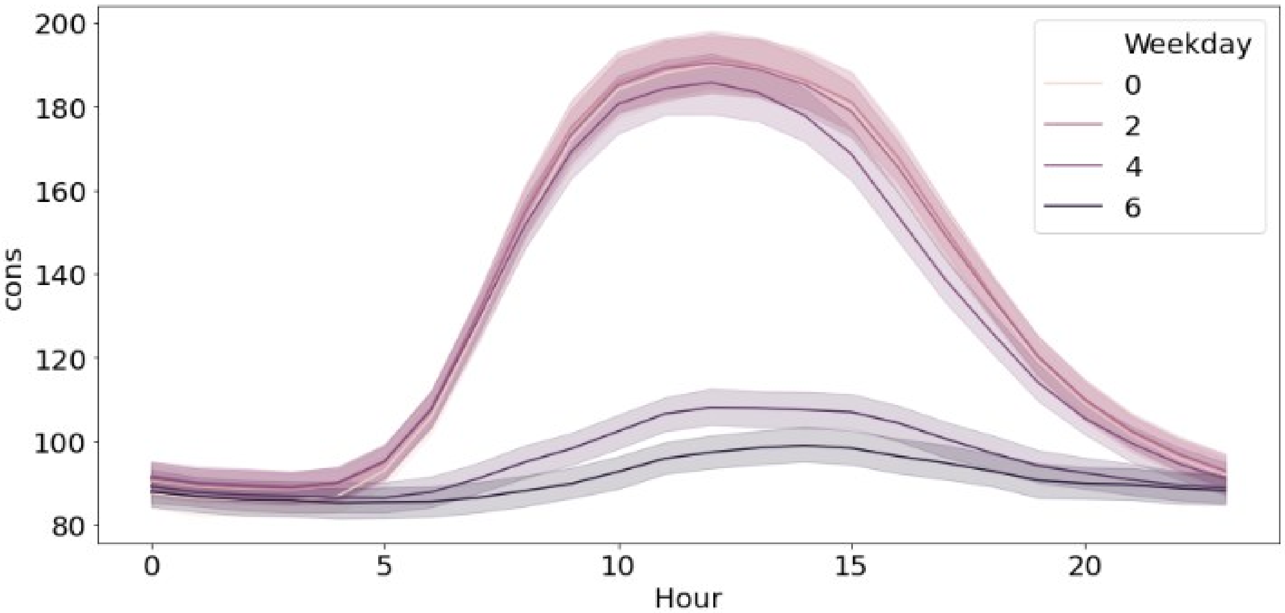


Figure 20. Hourly electricity consumption by day with confidence intervals.

To assess the expected relationship between outside temperature and energy consumption, the energy consumption against temperature for both heat and electricity is plotted. Figures 21 and 22 show the mean of consumption over meters against temperature for heat and electricity, respectively. In the case of heat, there is clear negative correlation with colder temperatures whereas for electricity such a trend is absent. This effect disappears after 10 degrees Celsius. This is probably because inside heating is turned off in the summertime. Consequently, purely outside temperature -based model would not work well for heating consumption prediction in the warmer months.

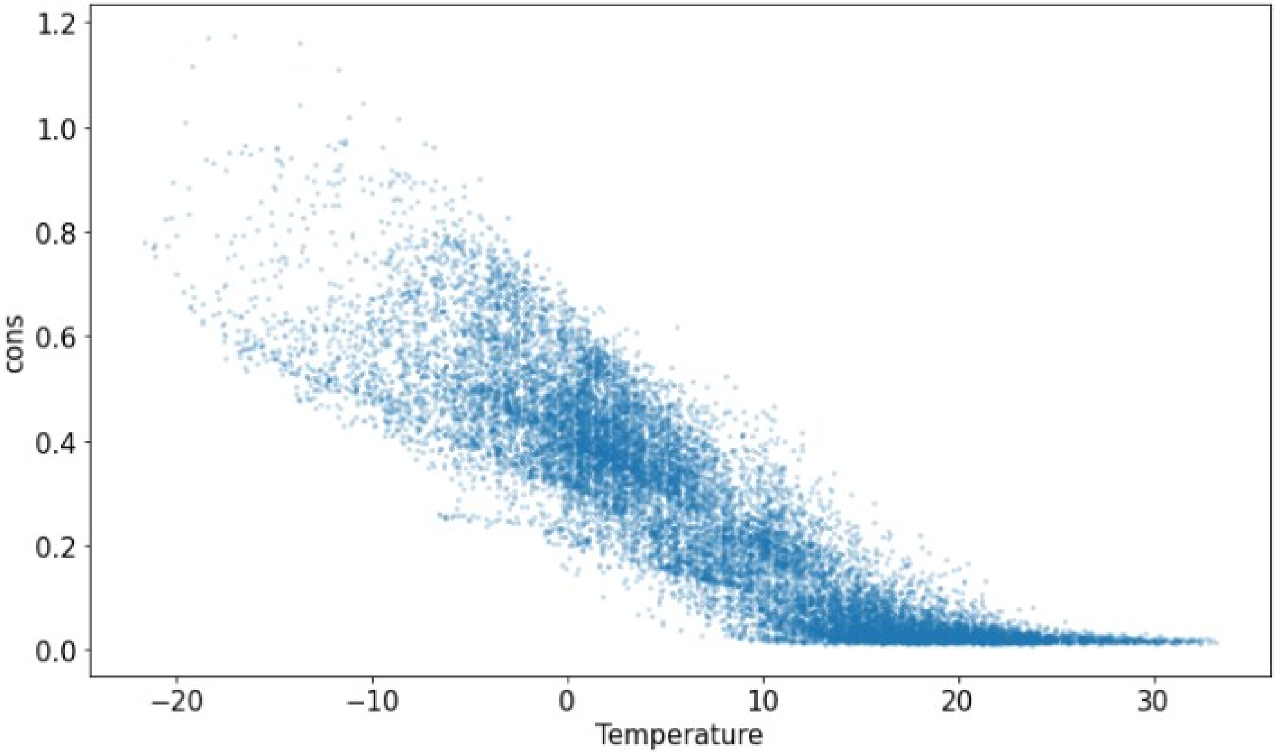


Figure 21. Heat consumption against temperature.

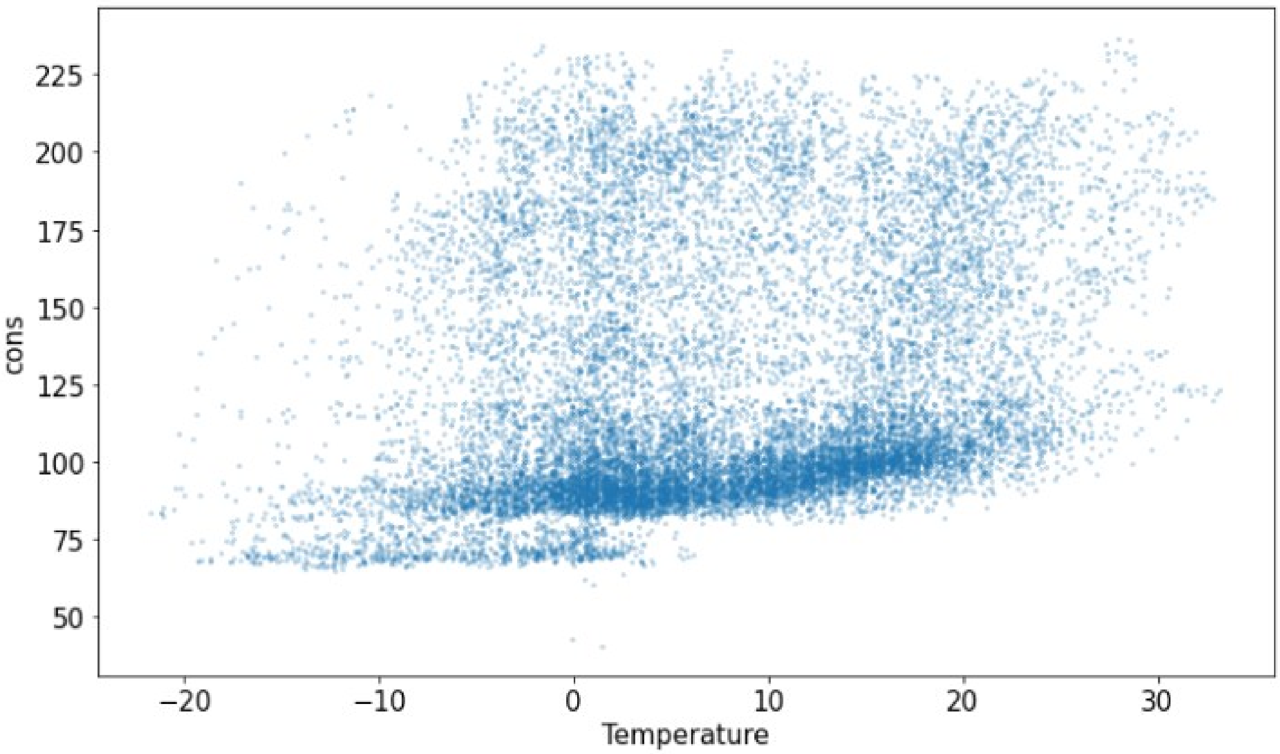


Figure 22. Electricity against temperature.

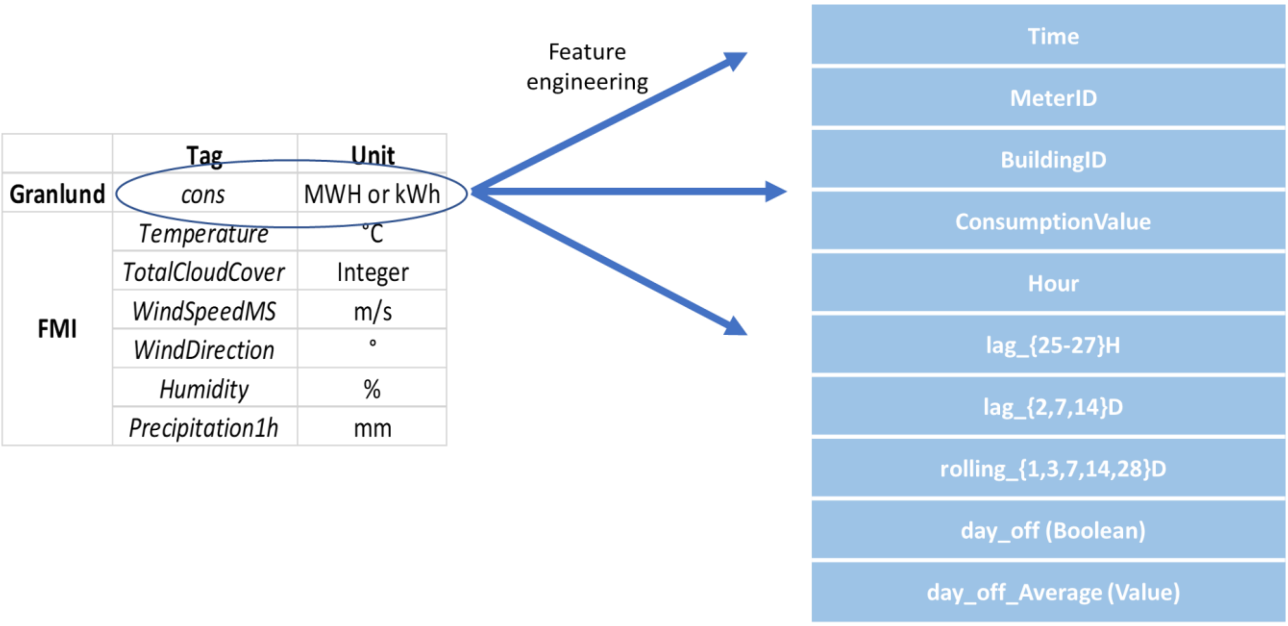
7.2 Data cleansing and transformation

In the data cleansing step, the errors found in the initial data diagnostics step are dealt with. The meters with periodic constant values are removed from the consumption data following the earlier discussed difference procedure. Weather data was also clean except for the missing values. The model should be designed to be robust to errors as it may encounter new types of data errors when in production. To achieve this robustness the errors in the weather data are left in place. Moreover, the decision tree -based ensemble methods used are robust to errors, as discussed.

Data transformation contains data rescaling and feature synthesis. Feature synthesis involves feature engineering, i.e., enriching new features from the old, and feature selection where some of the features can be discarded to save processing power. (Smitha Rao et al. 2019.) Rescaling can be skipped as the modelling methods used do not benefit from it, as discussed earlier. In addition, due to rather low dimensionality and reasonable dataset size, no feature selection is performed. Feature selection would require additional analysis unnecessary considering the expected computation time being low as it is.

From feature engineering perspective, the consumption dataset in enriched by introducing lag values for the consumption. Specifically, consumptions from 25 to 27 hours past and from two, seven and fourteen days past are introduced in addition to rolling averages from the last one, three, seven, fourteen and twenty-eight. These are the past values the model uses to predict the future consumption. In addition, information about off-days, i.e., weekends and holidays, is added, more specifically, concerning if the day is off-day and what has been the average consumption past off-days. The feature engineering process is illustrated in figure 23. Lag temperature values are also introduced in the weather data from one to five hours past. Wind direction, originally given in integer degree values between 0 and 360, is binned to eight bins of 45 degrees in width.

Both heat and electricity datasets are divided into even-sized chunks with monotonic time index. The first chunk, or fold, is first used as the training set and the second as the test set. Then first two are used in training and the third in testing. This procedure – a modification of the k-fold-blocks cross-validation introduced by Touzani et al. (2018) – enables us to assess the model against multiple test sets while ensuring that test data is always unforeseen to the model. The growing size of the training set should improve model performance towards the end of the procedure. Thus, different training-testing rounds are not directly comparable.



	Tag	Unit
Granlund	*cons*	MWH or kWh
FMI	*Temperature*	°C
	TotalCloudCover	Integer
	WindSpeedMS	m/s
	WindDirection	°
	Humidity	%
	Precipitation1h	mm

Figure 23. Feature engineering of the consumption data.

8 Modelling

Separate models for heating and electricity are developed. Both models developed are decision tree - based gradient boosting machines. This architecture is chosen due to the advantages in robustness, efficiency, performance, and explainability discussed earlier. Gradient boosting trees are also prominent in Kaggle prediction competitions when there is structured data. A case in point is the gradient boosting model by Taieb & Hyndman (2014) developed for the Load Forecasting track of the Kaggle Global Energy Forecasting Competition 2012. From implementation standpoint, the chosen architecture is great due to little pre-processing required which simplifies the adoption to production and upkeep of the model.

After choosing the overall architecture of the model, the specifics are determined through fine-tuning of the hyperparameters (Géron 2019, ch. 2). Correct tuning is extremely important for machine learning models (Mohammadi et al. 2019). Learning rate is highlighted as the most important hyperparameter to optimise by Friedman (2001), in addition to the number of components, or trees in this case. Friedman identifies a relationship where the learning rate affects the optimal value of iterations – note that one component is added every iteration – and vice versa. More specifically, larger number of iterations calls for lower learning rate to avoid overfitting. In order to maintain low training time for the model, the number of iterations is fixed to 1000 and the learning rate optimised based on this.

Randomized search is used to test for different hyperparameter values. Randomized search is advisable when the search space is too large to be searched exhaustively using grid search (Géron 2019, ch. 2). The hyperparameter optimisation is carried out using the Optuna framework (Akiba et al. 2019). Optuna is easy to implement and lightweight in addition to being efficient as it uses pruning – the ability to terminate trials not meeting predefined conditions (Akiba et al. 2019) – making it ideal for model tuning in this case.

An alteration of the cross-validation procedure – introduced as the predictive sample reuse method by Geisser (1975) – is used to compare the different models produced in randomised search. In cross-validation, data is divided into chunks of one or many observations. Usually one chunk is left out for testing while the others are used to train the model. Then the chunk to be left-out is changed. The idea of cross-validation is to arrive at an estimate of the true error of the model by averaging over these iterations (Varma et al. 2006). The modification done to the procedure described comes from the use of time series data. As time series data is characterised by dependence of future observation on the past observations, one needs to ensure that training data is always older than testing data. This is achieved by splitting the data monotonically with respect to time and using the oldest chunk for training and the second oldest for testing after which the two oldest are used to trained and the third to test and so forth. Varma et al. (2006) and Reunanen (2003) highlight a common source of bias related to cross-validation. Researchers tend to optimise the model parameters using the same data they use to evaluate the final model. This introduces a bias in terms of the error being underestimated as the model is optimised to deal with data it is not supposed to have encountered before the actual model evaluation. To remedy this, the dataset is divided into training and testing sets and only the training set is used in model tuning and training. This hold-out test set comprises of 20 percent of the data. Sufficient amount of data in both heat and electricity sets enables us to hold out a test set of sufficient size for reliable assessment of model performance.

The model tuning was initially performed with a wider array of tuned parameters, but only learning rate and the number of leaves were found to be of impact. This is in line with the LightGBM 2.3.2 documentation (2020, p. 57) where the number of leaves was highlighted as the main parameter controlling the model complexity. The importance of learning rate has been discussed earlier. Hence, the models have a learning rate of 0,1 with 242 leaves for heating load and 0,01 learning rate with 114 leaves for electricity consumption. Rest of the parameters are at default values. Instead of utilising gradient one-side sampling, an ordinary gradient boosting decision tree was used. This is because with the current amount of data, the training cycle lasts under a minute offering no incentive to trade accuracy for speed by using GOSS.

9 Findings

The models developed in section eight are assessed here and their performance validated. Statistical validation is performed using commonly used performance metrics found in industry literature. Chapter 9.1 contains the statistical validation. Additionally, expert interviews are used to validate the results. The interviews are discussed in chapter 9.2.

9.1 Statistical evaluation

Bias and variance are the two sources of error that characterise the performance of a machine learning model. Bias means persistent error in the model performance. Model with high bias systematically produces false predictions. In other words, the model is underfitting, i.e., not learning the data. Variance on the other hand, is the result of overfitting. An overfitting model fits the random noise in the data in addition to the true signal. This results in high variance as the model outputs highly differing predictions. (Gutierrez 2015.)

Two metrics to evaluate the performance of the models are used here, the coefficient of variation of the root mean square error (CV(RMSE)) and R^2 defined as $CV(RMSE) = \frac{\sqrt{\frac{1}{N}\sum_i^N (y_i - \hat{y}_i)^2}}{\bar{y}}$ and $R^2 = 1 - \frac{\frac{1}{N}\sum_i^N (y_i - \hat{y}_i)^2}{var(y)}$, respectively (Touzani et al. 2018). In CV(RMSE) the root mean squared error (RMSE) is compared to the mean of the data. This makes the measure unitless making it easier to compare and assess. CV(RMSE) is the most used performance metric in building energy consumption prediction studies (Amasyali & El-Gohary 2018). Under 30 percent CV(RMSE) is the ASHRAE guideline for prediction accuracy in energy forecasting (ASHRAE Guideline 14–2014). R^2 is considered as the proportion of variance in the data that the model explains. Hence, it is always between zero and one making it also unitless and easy to interpret. (Gutierrez 2015.) The metrics discussed for the models are 23.3 % CV(RMSE), 91.0 % R^2 and 19.3 % CV(RMSE), 97.6 % R^2 for heat and electricity, respectively (please, see table II). Based on the metrics, both models reach great performance and would be suitable to be used, for instance, in anomaly detection.

To better assess model performance, simple linear regression models are trained, one for each energy type. The models take the latest available consumption values and temperature for the hour. Although the models being simple, they reach quite satisfactory results. The heat model even passes the 30-

percent-CV(RMSE) guideline. The benchmark performance values are 28.4 % and 40.2 % CV(RMSE), and 86.0 % and 89.7 % R^2 for heat and electricity, respectively (please, note table II). Compared to these models, our models still clearly outperform them.

It is interesting how good the R^2 values are – especially for electricity. As one can see from the definitions of these metrics, R^2 has variance as the denominator whereas CV(RMSE) has mean. And as our data has rather large variance – as discussed in 7.1 – it boosts the R^2 metric. This variance explains, the large R^2 values. It also implies us to focus more on the CV(RMSE) for performance evaluation.

Table II. Model performance metrics with linear regression model as benchmark.

Model	CV(RMSE) (%)	R^2 (%)	Reg_CV(…) (%)	Reg_R^2 (%)
Heat	23.3	91.0	28.4	86.0
Electricity	19.3	97.6	40.2	89.7

Although, the promising results, the residuals – i.e., the differences between predictions and actual values – need to be checked for patterns indicating any systematic error. Here an observation is understood to emerge from a process of the form:

$$y_i = M(W_i, C_i) + \epsilon$$

Here the specific observation is denoted by i, W_i is a vector of weather metrics, C_i is a vector of past consumption information, and ϵ is white noise. The deterministic part of the process, M, is thought to depend only on past consumption characteristics and weather and is believed to be captured by our model. For our belief to be true, the residuals should resemble white noise with a mean of zero if the model is truly able to capture the deterministic process underlying the data. In the figure 24 the residuals of the heating prediction model, in fact, resemble white noise quite well with a mean around zero. More specifically, the mean is 0.02 MWh with a standard deviation of 0.07 MWh. Such well-behaved residuals confirm the claim about great model performance. Please, note that the residual distribution should not be normal as we are assessing residuals over a set of multiple buildings with different mean consumption levels. A plot of the residuals with respect to the mean of the heating data is provided in figure 25 for comparison. Analysis on the electricity model residual shows similar results (please, see figure 26). The mean of electricity residuals is 0.87 kWh with a standard deviation of 24 kWh.

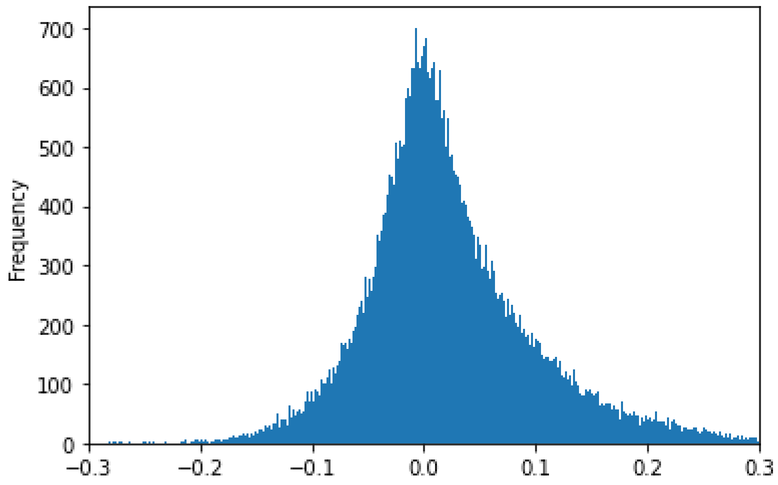


Figure 24. Heating load model residuals.

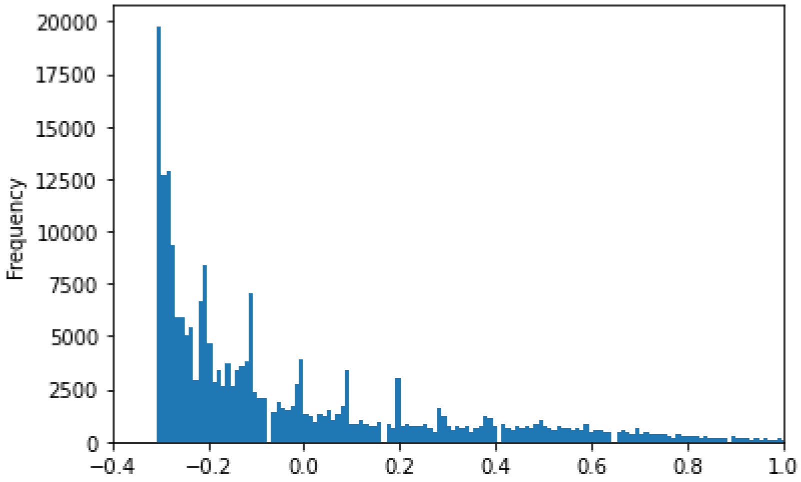


Figure 25. Heating load mean residuals.

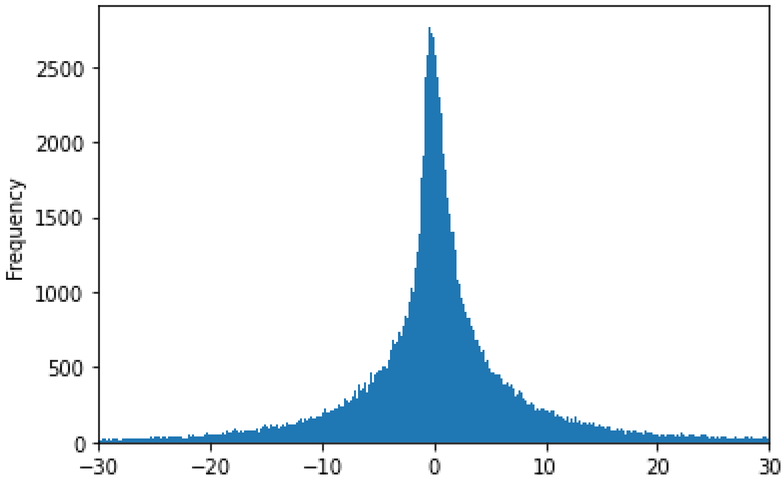


Figure 26. Electricity load model residuals.

Assessing the residuals through time, they seem to be homoscedastic, i.e., having even variance (please, see figures 27 and 28). Hence, the predictive power of the model is constant over the time dimension as one would expect for a successful model.

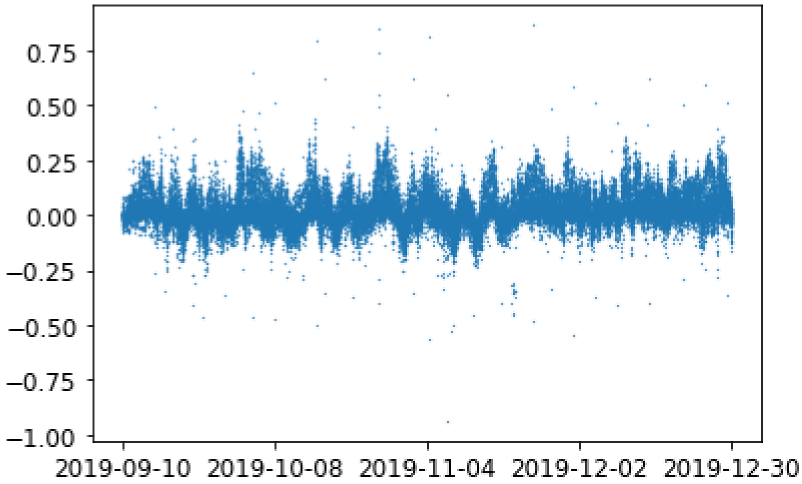


Figure 27. Heating load model residuals through time.

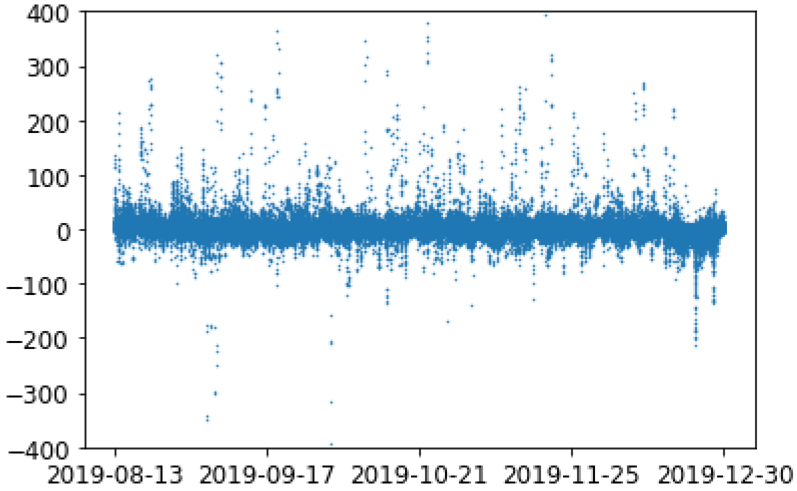


Figure 28. Electricity load model residuals through time.

Let us take a sample of a few buildings and assess the residuals building-wise. Here, the residual should resemble normal distribution. Of course, what is even more important is the mean of zero, as discussed earlier. The distributions of residuals for individual buildings visually seem to resemble the normal distribution more than the residuals over all buildings, as one would expect. Of course, the distributions are not strictly normal. Also, there appears to be bias in the predictions for some of the buildings. Hence, the model is not optimal. This could possibly be improved by training individual models for all the buildings. Overall, the residuals seem to have a mean around zero for most of the buildings, however. Hence, the model seems to perform well from this point of view as well. Plots for six individual buildings for heat and electricity can be seen in figures 29 and 30, respectively.

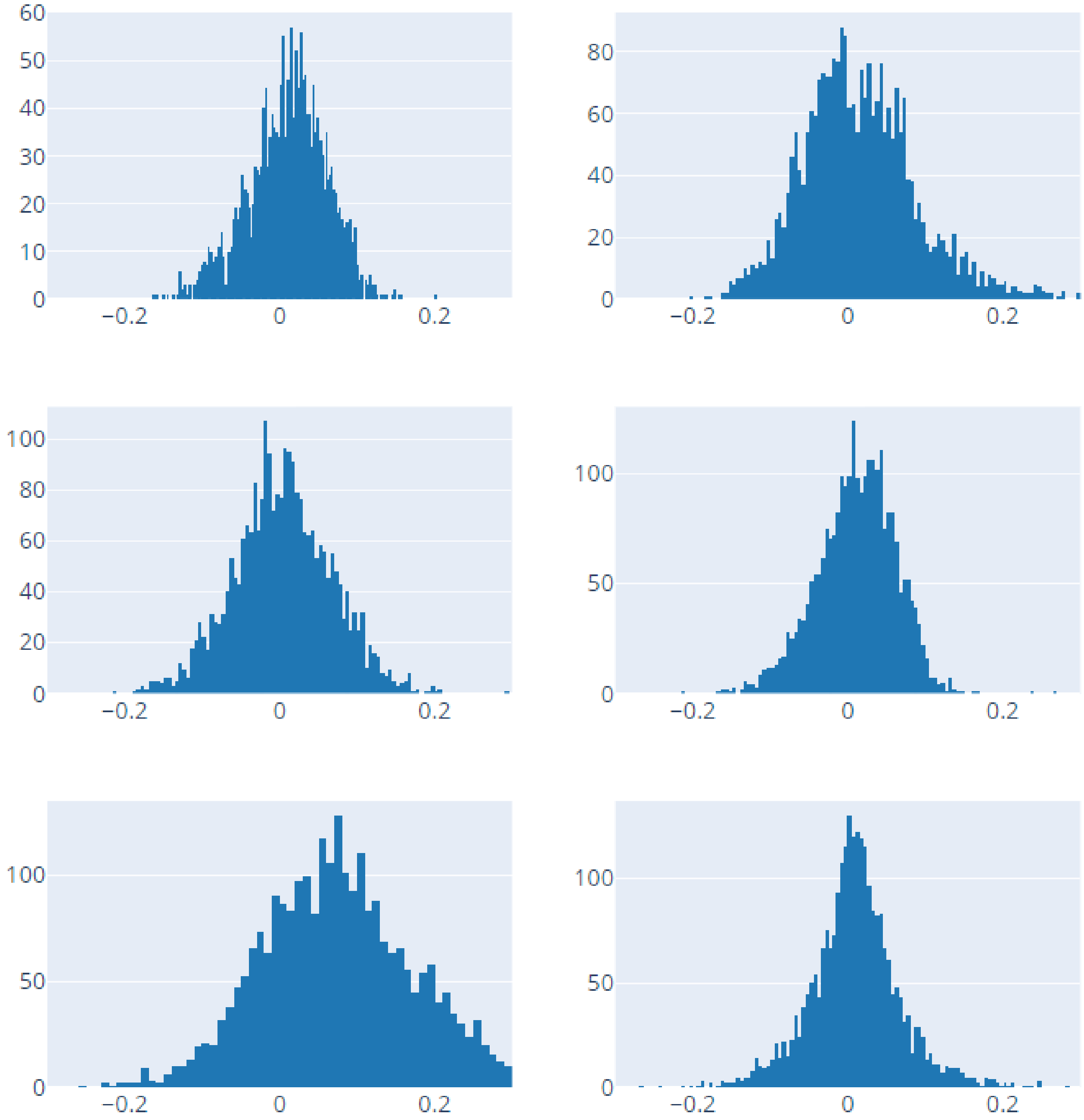


Figure 29. Meter-specific heat residuals.

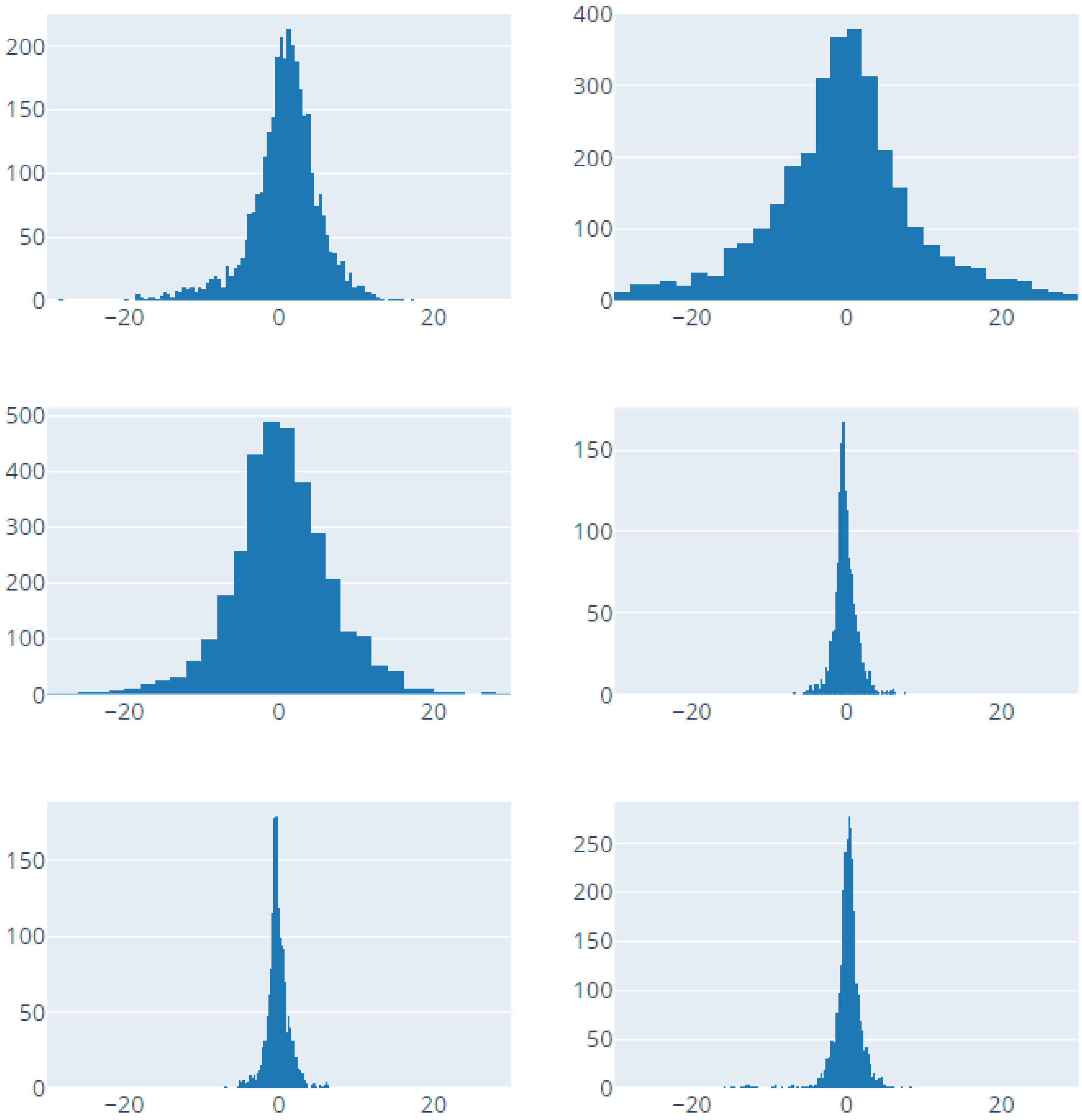


Figure 30. Meter-specific electricity residuals.

Normality assumption offers a way to construct confidence intervals (DiCiccio & Efron 1996) that – in our case – can be used to determine if a measured value should be flagged as an anomaly or not. In practise, however, the correct error sensitivity is found by trial and error and is highly dependent on the exact use case and context. A data centre manager is surely much more concerned about deviations in energy consumption than a landlord would be, for instance. Analysing the magnitude of the residuals with respect to consumption values, proportionately larger deviations with lower consumption values are found. Such plot for electricity can be found in figure 31. This is expected, though, as larger values affect the used error metric more the model prioritises higher consumption values in training. This is useful to note, however. Training should be altered if one would be particularly interested about lower consumption values. The model has adequate performance for lower consumption values as well, however.

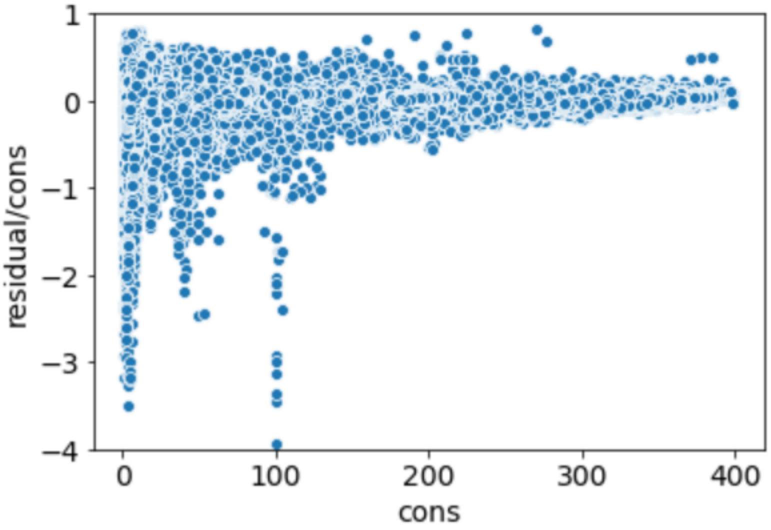


Figure 31. Electricity residuals with respect to consumption.

A scatter plot with predicted values against real values can also be used to assess model performance. The desired result is, of course, one-to-one correspondence between the values. Such plots for heat and electricity can be seen in figures 32 and 33, respectively. The plots are as expected with some discrepancies with electricity. Especially, the rather large overpredictions that are tightly packed around 100 kWh raise questions. However, such values – as highlighted in figure 33 – there are only around 80. Thus, no further interest is given to these values as they have no practical impact.

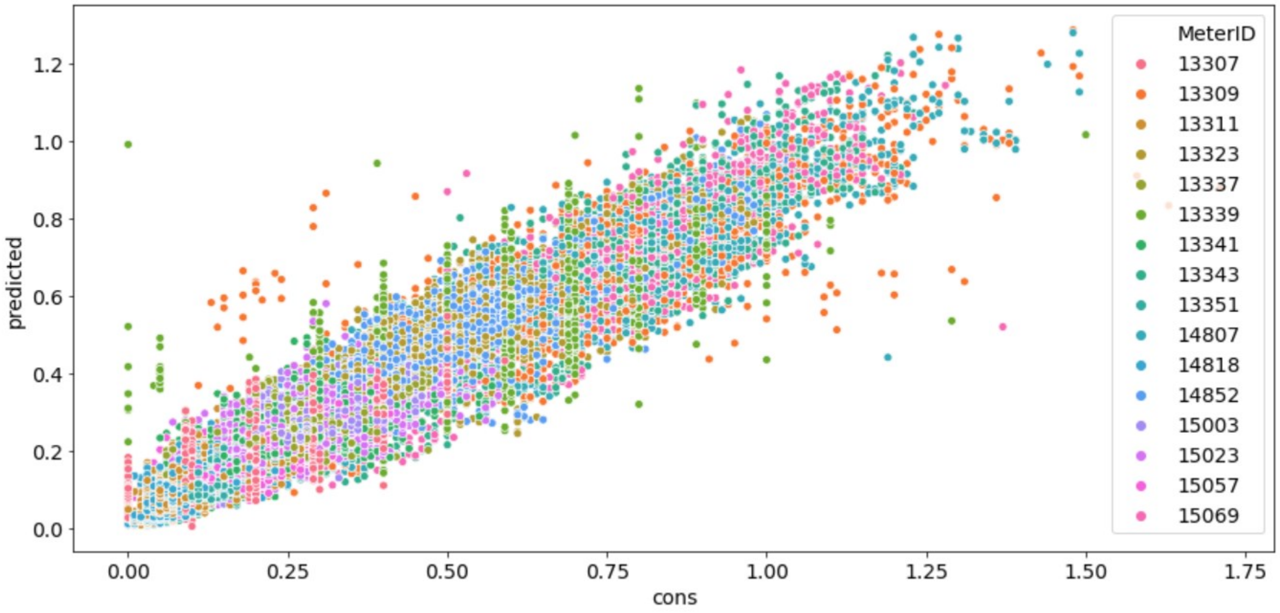


Figure 32. Heat actual consumption to predicted

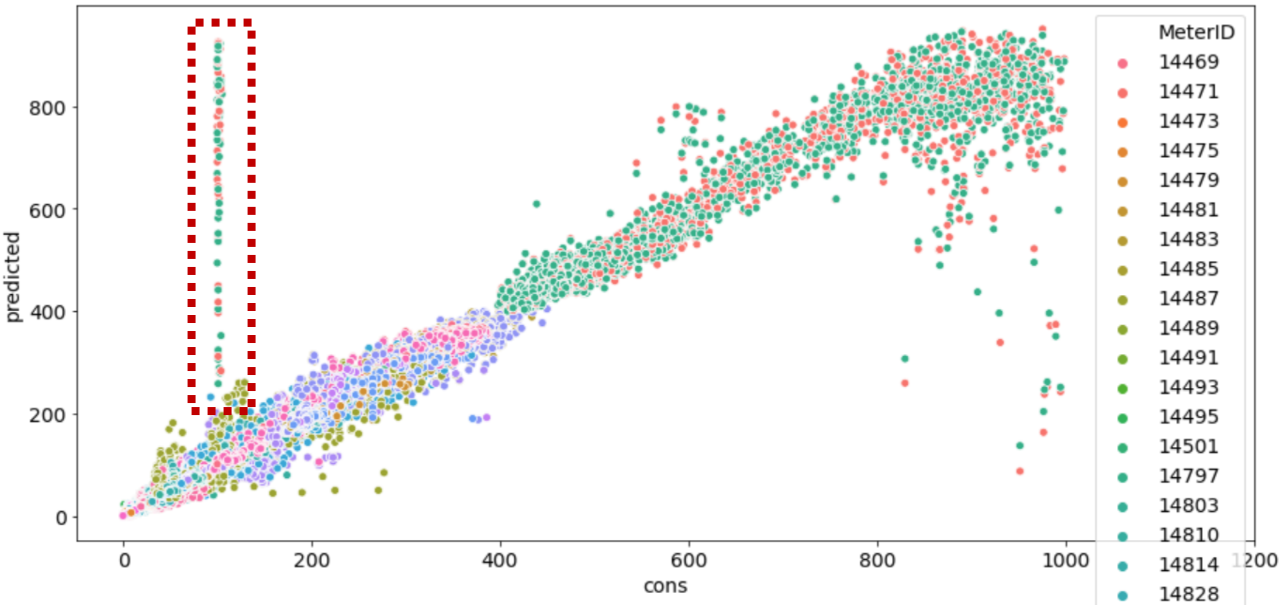
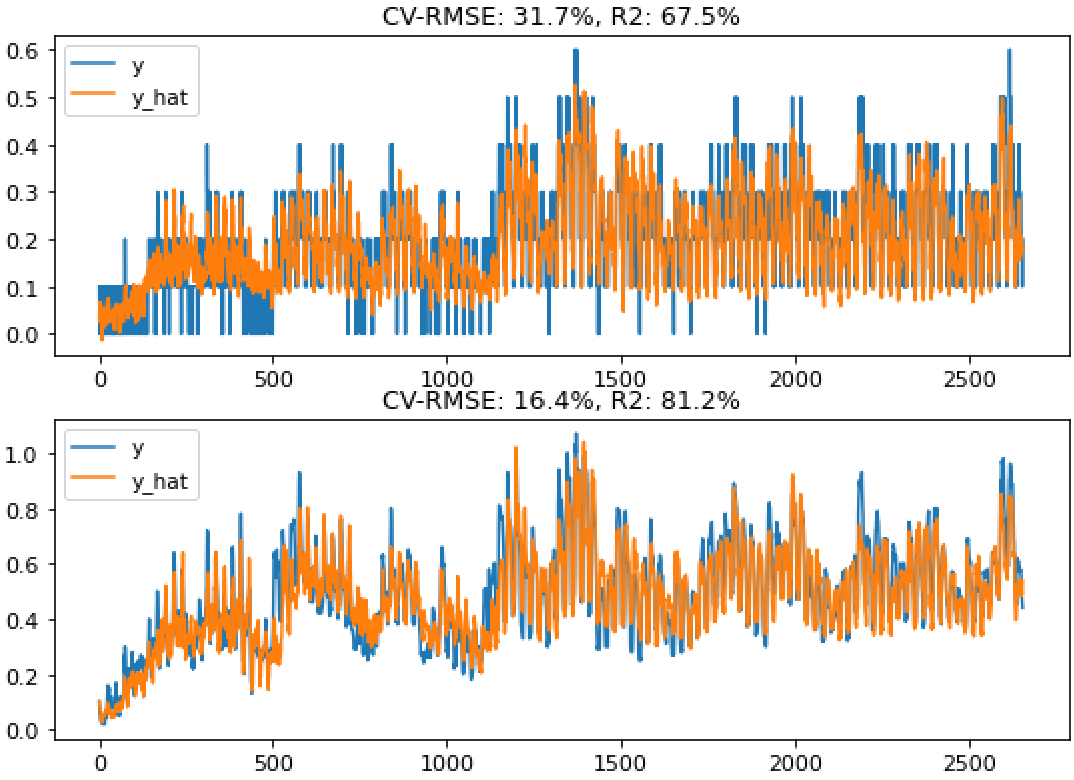


Figure 33. Electricity actual consumption to predicted with highlighted divergence.

The predictions of the models can be compared to actual heating and electricity consumptions in figures 34 and 35, respectively. Model is clearly able to forecast the consumption in great accuracy

24 hours into the future. Here one can observe the lower R^2 values resulting from the variance being lower for meter specific data than over all meters.


CV-RMSE: 31.7%, R2: 67.5%

CV-RMSE: 16.4%, R2: 81.2%

Figure 34. Heat consumption and the 24-hour prediction.

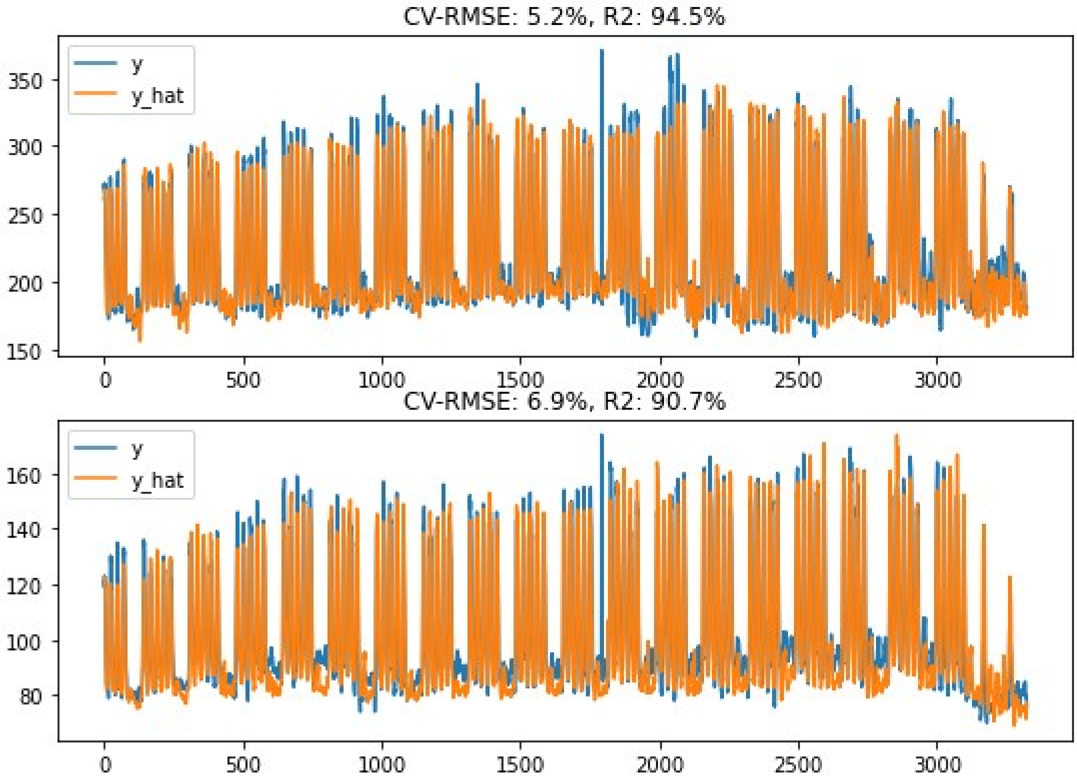


Figure 35. Electricity consumption and the 24-hour prediction.

9.2 Expert validation

Following the Design Science methodology, the model is validated by expert opinion. Five experts from both academia and business experienced in statistics, software, buildings, or usually some combination of the former were interviewed. The interviews conducted were semi-structured – please, refer to chapter 5.3 for more information – each lasting an hour. Interview details can be seen in table III. The insights from these discussions are elaborated next with comments attributed to different interviews by the interview number. The interview question used can be seen in the third appendix.

Table III. Interview details.

Interview number	Date	Medium	Name, title	Expertise (in years)
1	31.08.2020	Zoom (video call)	Jukka Kohonen, Ph.D., University Lecturer	Statistics: ~10 y. Software: ~30 y.

2	04.09.2020	Zoom (video call)	Risto Lahdelma, Dr.Tech., Prof.	Statistics: ~25 y. Software: ~20 y. Buildings: ~10 y.
3	08.09.2020	Teams (video call)	Ken Dooley, Technology director	Software: ~5 y. Buildings: ~18 y.
4	09.09.2020	Teams (video call)	Timo Karvinen, D.Sc. (Tech.)	Statistics: ~2 y. Buildings: ~2 y.
5	09.09.2020	Teams (video call)	Martin Aalto, Manager	Statistics: ~1 y. Software: ~3 y. Buildings: ~2 y.

The interviews focused on three main areas, i.e., the mathematical side of the model performance, usability and special characteristics in the building-context, and software view on to the model and its implementation. Also, other potential use cases were discussed. To start with the mathematical side of the evaluation, the model was deemed to exhibit great performance by all the experts interviewed.

The model architecture is regarded suitable for the task given the context. No reason to suggest change of architecture. (1-5.) "We haven't got the accuracy you have here [with our consumption prediction models]. I would say what you have done is really well." (3.) Usually, models tend to the mean. Here, however, spikes are also predicted and seem to be about right most of the time which is very good. The work overall seems thorough given it is a master's thesis. There is material for successful Ph.D. research as well. (2.) Residuals have zero mean which is good. The distributions do not seem quite normally distributed, however. (1, 2.) The tails of the distributions could be analysed further (1-3), e.g., to look for differences in performance between buildings (1, 3). Some simple model, e.g., linear regression, could be developed for benchmark. The model is quite a black box. For validation purposes, the tree created could be analysed to understand more deeply how it forms the prediction, and if it makes sense from physics point of view. (1, 2.)

The experts pointed out important aspects especially regarding model validation. To increase the credibility of the results and clarify the performance of the model with respect to other means of analysis, a simple linear regression model is developed in chapter 9.1 as instructed by the experts. Due to time constraints, additional analysis on the tails of residual distribution, i.e., where the model errs the most, is not supplied. To produce an in-depth view of the model structure would an interesting task. Future research could be done to see if simpler – and potentially even more robust – model with similar performance could be developed.

Regarding model usability in the building context, the experts felt confident that the model will perform with similar buildings and main metering data. However, they pointed out critical differences different kinds of and differently used buildings from the energy point of view.

Extrapolation to different kinds of buildings might turn out to be problematic (1-5). Also, going to the submetering level would likely cause trouble (3). The use of longer measurement period should also be considered as only two years are present in the data currently (1). Heat consumption in the winter is also different as space heating is turned off in the summertime. Time of the year information could be thus incorporated. Water consumption is also important for heat consumption and is not in the data currently. (2.) Water consumption could also be used as a proxy for occupancy. Occupancy is important for electricity consumption (3, 5), but occupancy may be regular enough to be incorporated indirectly, e.g., by a holiday calendar (3). The great performance of the model with electricity without occupancy may be because the buildings in the dataset have low occupancy and the users don't use that much electricity (3) The model can be tuned when the building environment changes. On the other hand, it might be quite adaptable already as change in the building is reflected in the historical consumption data it is fed in. (1.) Anomaly detection offers great value in building energy (3). An anomaly that gradually changes energy consumption might be neglected by the model, however (1).

In a scenario where the model is deployed for production, thorough testing needs to be done beforehand to validate sufficient performance for the specific building stock concerned. The modelling done in this work seeks for good performance with minimum data processing and operational requirements. This is to ensure usability in business context. Acquiring and incorporating occupancy data in this use case would be against this principle as the performance – especially for electricity – is good already. The author believes the model to be adaptable to handle different weather conditions resulting in good year-to-year performance. Different buildings with different types of use will surely require retraining without guarantee of similar performance.

Although, the other data processing needed to configure an automated workflow for an operational application is excluded in the work, the draft of such a pipeline was discussed in the interview. The draft made sense to all experts. End-user usability (4) and data quality related problems (2) were highlighted as common pain points with such applications. Open-source code libraries are also susceptible to become unsupported as time progresses which requires the availability of personnel capable of rewriting some of the solution code (5). These notions are useful and need to be considered when incorporating the model into production. In terms of other possible application areas, scenario building (1) and energy system optimisation (2) were identified in the discussions.

10 Discussion

The implications of the findings from the study and the model developed are discussed in the context of building management. Also, the disruptions from smart technologies more broadly are discussed. After the business-oriented discussion, learnings from the study are discussed in the context of building management academia. Discussion of model limitations and further research avenues concludes the section.

10.1 Model implications

The interpretability – aforementioned strength of decision trees – is showcased in figure 36 for the heating load prediction model and in figure 37 for electricity. The feature value indicates which features the model finds the most useful. More specifically, the importance metrics in figures 36 and 37 are Shapley values resulting from the TreeExplainer, a method developed by Lundberg et al. (2020) to infer Shapley values for interpreting tree-based models. In the figures, every observation, i.e., data row, is represented by a dot. Red dots indicate a high feature value, e.g., high hourly consumption value, and blue dots low. The most important features get listed first. The direction and the magnitude of the impact is indicated through the Shapley values on the x-axis.

Assessing the gains in the heat model, the 25-hour lag, i.e., the latest available consumption reading, is – expectedly – the most important feature. Electricity, however, reveals a different story; the seven-day lag is the most important feature the 14-day one coming second. This implies there to be a strong weekly seasonality in the electricity consumption. Heating demand, on the other hand, exhibits stronger path dependence, i.e., recent observations predict future events.

Moving on to weather, the three-hour past temperature is found to be the most important for heating prediction. Its effect is also negative – as one would expect – meaning that lower temperatures predict higher heating consumption. For electricity, weather conditions seem to have little impact.

All in all, the main implication is that historical consumption is clearly more important than weather data even for heating load which one would think of being highly dependent on the temperature. Consequently, reliable prediction without the weather forecast would seem to be possible at least for short-term predictions with this type of buildings. In combination with the usefulness of longer lags, this opens an avenue to possibly lengthen the prediction horizon while maintaining reasonable accuracy. This pursuit is out of the scope of this study, however.

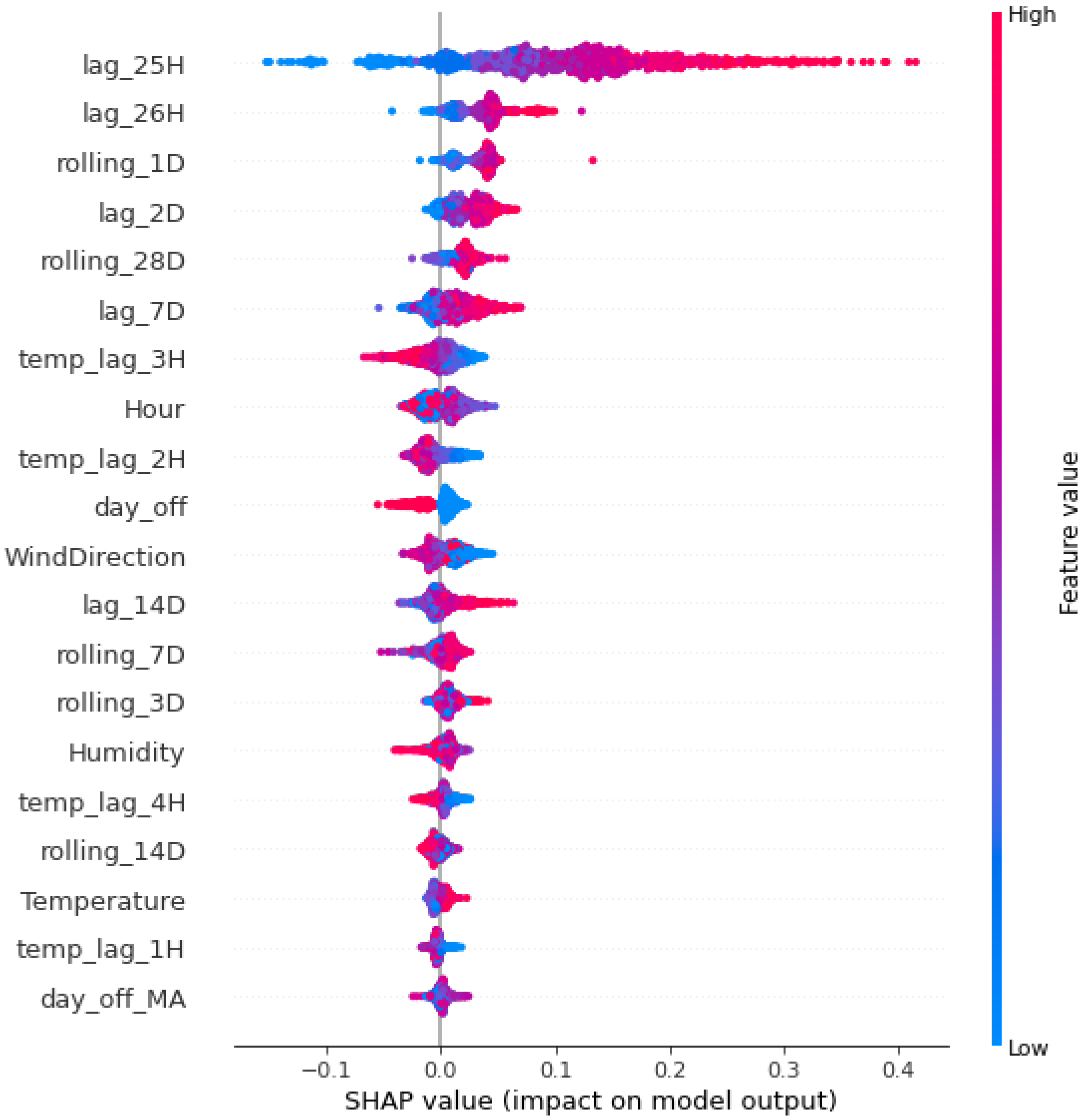


Figure 36. Feature values in the heating load model as Shapley values.

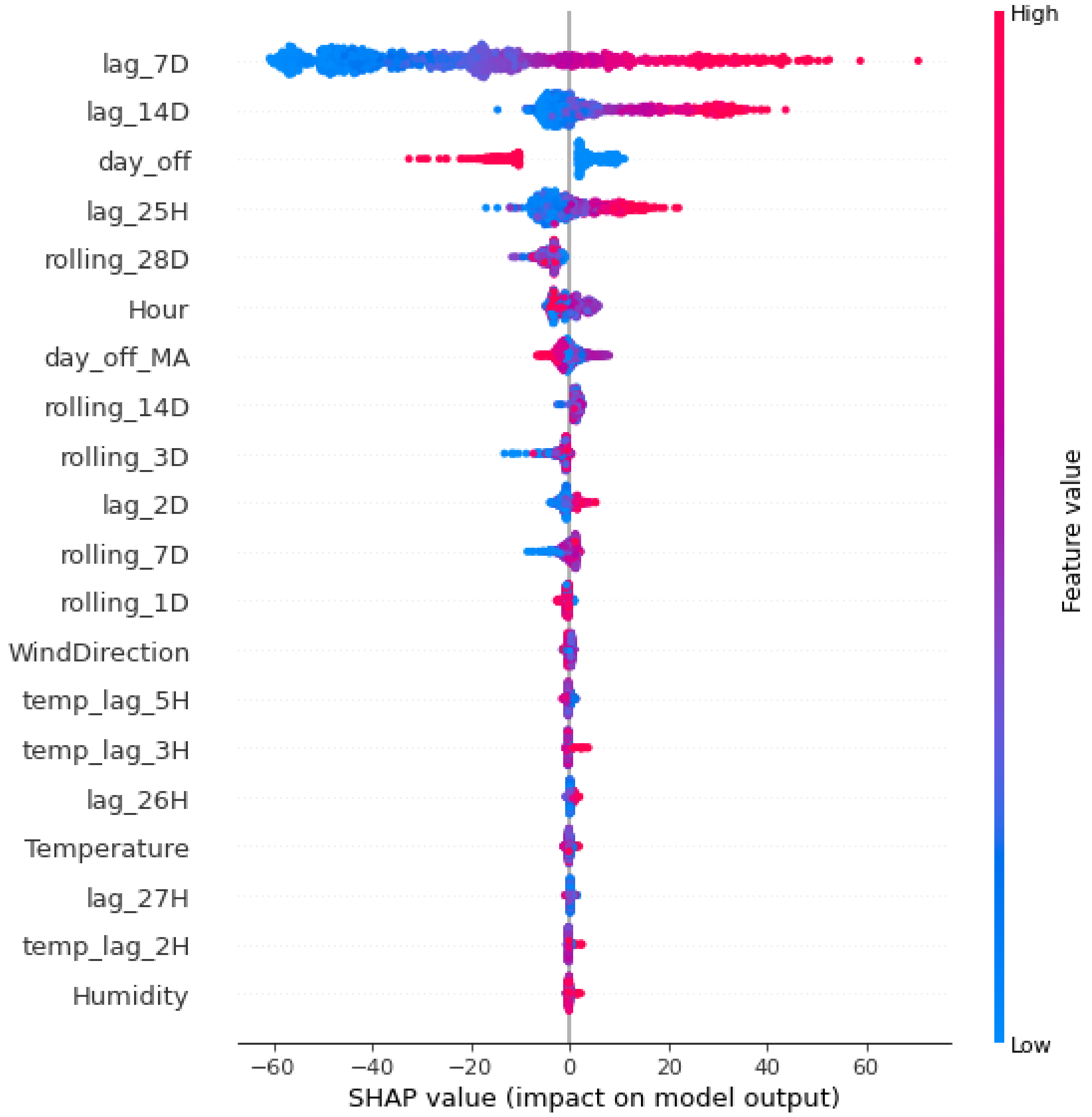


Figure 37. Feature values in the electricity load model as Shapley values.

As noted earlier in chapter 3.1.3, Kiran et al. (2015) propose using anomaly detection in the Lambda architecture's speed layer to assure data quality. The model developed in this work would fit this task nicely. It is efficient enabling the use of large data streams and has the required performance for anomaly detection. Such implementation would be extremely useful for Granlund as data quality is an issue in the energy consumption monitoring service Granlund offers. The model would need to be integrated into Granlund's existing ETL pipeline. Similar model would be useful for utilities as well as they need to correct the errors in their data which requires manual work. An interesting future research proposal would be to study possibilities to correct the anomalies found automatically. An anomaly detection tool built on top of the model developed here would be the first link in an automated anomaly correction pipeline. Such a project – how interesting it might be – is, of course, out of scope for this study.

10.2 Building management implications

As discussed, the whole energy system is in transition. This change is driven by the large-scale introduction of renewables as a consequence of the pursuit for cleaner energy system. Renewable generation has inherently variable generation profile. This instability needs to be dealt with to enable the transition to renewables-led energy system. This stabilization is a complex endeavour as it requires the participation of demand-side players as well. Means to coordinate all grid participants do not yet exist but require the application of state-of-the-art technologies and concepts in multiple disciplines. For example, new pricing models need to be applied in the electricity markets to enable the participation of distributed generation. This work discusses the transition only from the standpoint of buildings, for discussion about needed changes in electricity markets see, for instance, Kukkonen (2017).

Focusing on the building perspective, the building management community needs to change perspective. Currently, the question in the minds of building managers is how to maintain the specified indoor environment with minimum cost. This mindset is correct in a system where energy flows only one way. In a world with constantly tightening energy conservation regulation, the relevance of this question is only underlined. However, an additional perspective needs to emerge. As the future energy system is in dire need of demand-side services that buildings utilising smart technologies are able to offer, building management needs a mindset change to seize this opportunity. Future buildings are not only consumers of energy but “prosumers” that are able to generate energy through renewables and offer peak shaving services to balance the grid. Given the increased rate of innovation in the building sector, building management needs to be proactive to keep up with the latest developments. They need to be openminded to experiment with new business models and service offerings.

To capitalise on the possibilities offered by new technology, new kind of expertise is needed from building management professionals. Building management has traditionally been regarded as a field for engineers with building-specific expertise only. However, the transition to the smart building era

– not to mention the whole energy cloud development – requires more multidisciplinary approach. Talent with business knowledge is required to develop the new business models enabled by smart technologies. In addition, as the smarts of a smart building is in its software, not hardware, software and mathematical knowledge are needed to apply these technologies. To fully understand the possibilities that smart technologies offer in the building space, close cooperation between experts from all these disciplines is needed. Multidisciplinary cooperation is needed already when integrating multiple building systems, e.g., HVAC and lighting, under one BMS (Kastner et al. 2005).

Retrofitting building equipment induces heavy capital expenditures. Moreover, BMSs are investments with long lifetime. Thus, the sector has not been known for experimentation, but international standardisation of technology is often waited before adoption. (Kastner et al. 2005.) Capital expenditure considerations are on the emphasis when introducing smart technologies as well. However, especially with small and mid-sized buildings, smart technologies are often introduced in the component and subsystem levels, as discussed earlier. This brings these technologies in reach of managers with more stringent capital expenditure requirements. Cost sensitivity is a feature of the building industry (Kastner et al. 2005). Modularity is important especially when applying smart technologies in a piece-wise manner. This is to ensure system evolution through the building's lifetime. Smart building needs the interlinkage of its energy resources and sensors. Thus, the smart components and systems need standardised means of communication.

The earlier discussion about smart technologies highlights the importance of prediction. Prediction is utilised in new control schemes, e.g., MPC, used to minimise the cost of purchasing energy, optimise the use of DG resources in a microgrid, estimate the amount of available demand response capacity, and in this work, to detect anomalies such as mismanaged building equipment. The importance of occupation information for building operation is underlined in prediction. This is because occupant behaviour has a considerable impact for building energy consumption and indoor environment. Different users also appreciate different conditions, e.g., temperature (Chen et al. 2009). Hence, occupants need to be included in the operation of a building. Instead of trying to minimise the impact of occupants, two-way communication channels need to be established between the building and the occupants. For instance, future occupation information, e.g., through room booking systems, should be included in the prediction of future energy need. Also, the preferences of the users should affect the building, for example, people in the wintertime probably appreciate a bit lower indoor temperature than in the summer due to difference in clothing. Additionally, the building should also communicate with its users. For instance, as preferences vary, different temperature zones can be maintained in common areas such as libraries. The building needs to then communicate the existence and location of these areas to its occupants.

This study showcases a practical application of statistical – or black box – prediction model in building context. First, the applicability of gradient boosting machines specifically is shown for building energy consumption modelling and prediction. Although classified as a "black box" model, the name is a bit misleading as one of the gradient boosting machines' strength is their interpretability as shown earlier. GBM is great statistical method for situations where data includes categorical features as they do not require pre-processing. The robustness of GBM allows its application in situations where the input cannot be entirely controlled as the case usually is with practical applications. Thus, the model can withstand unseen outliers resulting from measurement errors that are commonplace in the building energy context.

Second aspect of the work – and even more important than the application of GBM specifically – is the development of a model that combines energy data with environment data from external sources, namely the FMI weather data. This is in line with the notion of Buckman et al. (2014) that there are usually useful external data sources that just are not currently being utilised. In case of demand response, Chen et al. (2009) highlight the need of real-time energy pricing information. Online analytics utilising energy price combined with the status information of the building are needed to make energy purchasing, selling, and conservation decisions. Representing and constantly updating a comprehensive status of a building is a challenging task from a software development standpoint. Hierarchical spatial division of the building, e.g., room, floor etc., allows for the creation of complex behaviours as an aggregation of simpler policies making the system more affordable and flexible (Chen et al. 2009).

10.3 Buildings as DER providers

Accenture (2018) identifies two new players emerging in the energy business through smart technologies, the demand aggregators, and the energy managers and consumption optimizers. Burger & Luke (2017) call the latter EMS providers for the fact that their business is to provide their customers with monitoring and control equipment. They also highlight the fact that aggregators usually provide their customers with EMS equipment needed to allow them to offer energy resources.

Hence, aggregators' business goes beyond only providing EMS, to actually enabling the customer to economically benefit from the resulting energy resources by enabling the sales of these resources to other market participants (Ikäheimo et al. 2010). Ikäheimo et al. (2010) as well as Ma et al. (2017) characterise aggregators by the type of resources they manage to demand aggregators, i.e., demand response, and generation aggregators, i.e., virtual power plants. Aggregators can also combine these resources (Lu et al. 2020). Aggregation is not always needed but consumers can offer their flexibility resources directly to the market. This is common for large industrial consumers. (Accenture 2018.) Participation in DR programs requires large-enough loads (Lamprinos et al. 2016, Lu et al. 2020). Hence, aggregation is required for smaller consumers to participate (Lu et al. 2020, Ma et al. 2017). Moreover – at least in Finland – industrial sector loads have already been exhausted (Ikäheimo et al. 2010).

The fundamental purpose of an aggregator is to financially expose consumers to the state of the grid at any given time. Currently, small, and mid-sized consumers are insulated from wholesale market fluctuations by time-invariant pricing schemes. (Ikäheimo et al. 2010.) Retailers, balance responsible parties, and third-party companies could, in principle, all act as aggregators (Ikäheimo et al. 2010, Lu et al. 2020). Retailers are already well connected to the electricity market and have an existing relationship with the customer. Hence, they are best positioned to become aggregators. (Ikäheimo et al. 2010.) Retailers need to expand their pool of talent to have the required expertise to develop the DR business (Ma et al. 2017). Thus, a deliberate effort is needed from their part to enter the business. Ma et al. (2017) find an implicit – i.e., a voluntary, price-based – model with retailers as aggregators being the best choice for the Nordic electricity market. The development of the business models is strongly influenced by the regulatory environment, e.g., the electricity market structure (Burger & Luke 2017, Ma et al. 2017).

The aggregation market is still immature making the participation for buildings – especially small ones – difficult. The characteristics of the building and its use, of course, affect the types of programs suitable for the building. Focusing on non-residential buildings, direct control programs might be suitable for smaller buildings whereas larger buildings could participate in implicit DR. Large-enough buildings could also provide their flexibility resources directly to the market. Participation through a specialised aggregator would probably still be more convenient, however. (Ma et al. 2017.)

Analysis by Rubel et al. (2017) foresees two distinct markets to emerge for DER provisioning, a market for complex, customized solutions, and a market for standardised solutions. Conventional office spaces are probably best served by the cheaper, standardised solutions. These solutions can be distributed through a network of installers by the equipment manufacturers. Energy retailers can also act as intermediators connecting their customers with manufacturers (Rubel et al. 2017) and acting as an aggregator for the resulting DERs as discussed in Burger & Luke (2017).

Building management can be predicted to evolve more centralised through the emergence of aggregators and optimisers. This is because the connectedness of smart technologies enables remote operation with little on-site presence. This in turn improves staff utilisation and enables higher degree of specialisation. Actual on-site maintenance is probably outsourced. Even more importantly, however, smart technologies bring algorithms and analytics in the forefront of building management. Human actors are going to reposition to supervisory functions whereas operative decision making, e.g., electricity purchasing and DER management, is automated. The development of these algorithms requires notable upfront investment whereas the integration of an additional BMS – given it is using standardised communication protocols – is practically free. Thus, the importance of economies of scale as a key success factor increases considerably. The development of such a software system requires resources, e.g., access to talent, available to only companies of a certain size. The model developed in this work could work as a component of such a system. Building management needs to embrace, not fight, this development. Even though the actual on-site managers lose some control over the building, centralised planning enables more efficient use of the building. Additionally, prediction enables more stable indoor environment and frees on-site resources for other work. Open collaboration with utilities and retail energy companies experimenting with aggregation programs provides learning pivotal for unlocking the potential of smart technologies. On the other hand, prototype implementations involve more hassle and with the absence of standards may introduce lock-in to a certain provider. Buildings – especially larger ones – can harness their DERs using open standards technology even without third-party involvement.

10.4 Implications to data strategy

This work – although yet being on the proof-of-concept level – is a typical cost reduction pilot. Special attention has been dedicated to document the process from the raw data to the final model as thoroughly as possible. That being said, the existence of this report on its own does not make the organisation any more capable. Especially for this being a pilot project – it is important to deliberately diffuse the learnings from this work into the organisation. Hence, a power point presentation – catered for building management professionals – is prepared. Effective communication of the findings is also emphasised in design science as already discussed in chapter 5.1.

Even given the combination of a power point and a word document, the development of capabilities on the organisation level requires strong support starting from the c-suite level as identified in the MIT Sloan research in Ransbotham et al. (2018) (for complete discussion, please, refer to the chapter 3.2). They also identify the need for dedicated AI strategy (Ransbotham et al. 2018). The Digital Transformation Framework by Matt et al. (2015) with its four dimensions; ability to adopt new technologies, changes to value creation, structural changes, and financial aspects, would be a great starting point for AI strategy formulation. Thorough evaluation of Granlund's capabilities and characteristics would require access to business sensitive data and would be a task clearly outside the scope of this research. However, to project the learnings from this project on the DTF, Granlund clearly has abilities to adopt new technologies. This ability has only been visible, however, in point solutions such as this work.

As discussed in chapter 3.2, BCG research identifies the need for end-to-end processes for creating, deploying, and monitoring AI applications. Granlund is at the point where the lack of these processes inhibits it to create systematic capabilities into the organisation from these point solutions as discussed 3.2 and in Ransbotham et al. (2018). Management level effort is needed to start developing these processes. First wins – such as cost reduction in the data reporting service by applying the model developed in this study – are important in justifying the required top management resources. Cost reductions also release financial resources for the development.

Speaking of financial aspects, the initial investment has to be sufficient to support a dedicated development team. On the other hand, sustained effort is needed to develop the core and advanced capabilities identified by McKinsey research discussed in 3.2 and Bughin & Van Zeebroeck (2018).

Capabilities-wise, Granlund – even with existing web application and database expertise – is lacking knowhow in cloud-based technologies and advanced analytics. McKinsey identifies the bundle of these capabilities being vital to succeed in AI-driven competition. Talent management and acquisition is important as employees with these capabilities are highly sought of. Implications to organisational structure and ways of value creation become relevant when the needed organisational capabilities have been developed.

The development of statistical prediction models is extremely data intensive, as discussed earlier in section four. Hence, companies with access to historical consumption data are in a strong position regarding model development. This includes utilities, electricity retailers, and those technical service providers who have access to their customers data.

Assessing data as a resource from Barney's (1991) resource-based view, data clearly has value as it enables development of statistical models. Rarity, on the other hand, is harder to assess. Access to vast amounts of data is seen to give competitive advantage – arguably even in the proportions that regulators should be interested – to technology giants such as Alphabet, Microsoft, and Amazon (The Economist 2017). On the other hand, data is cheap to copy and distribute. An example of this in the Finnish context is the national electrical transmission system operator Fingrid's Datahub project. Datahub by Fingrid (2020) is "a centralised information exchange system for the electricity retail market". Datahub will store electricity consumption data and make it easily accessible for parties authorised by the consumer. Thus, new entrants do not have the need to setup their own system to gather and store their customers data to provide analytics or other, e.g., DR services – enabling DR is a pronounced endeavour in the Datahub project. (Fingrid 2020.) For Granlund, this clearly undermines the advantage it possesses from having its own data processing and storage capabilities. Of course, as companies can utilise only the data they have been given explicit permission to access, Granlund can have an advantage – given that it has negotiated ownership for the data it gathers for its customers.

Considering the characteristics affecting the sustainability of an advantage – namely inimitability and non-substitutability – simulated data can be used to substitute real. Simulations are only approximations of reality, however, meaning that even a hypothetical model's performance is dependent on the correctness of this approximation. White-box models – discussed in section four – can be used to imitate the utility of data-hungry black box models. Statistical methods have reached performance levels far beyond the reach for other approaches in multiple applications, however.

10.5 Contribution to existing knowledge

Design science principles guided the thesis process. The work demonstrates the use of design science in a modelling setting. Moreover, it shows how fitting the methodology is for an engineering thesis setting which often blends both academic and business aspirations. In addition to acting as a case in point about design science in engineering, I hope and believe that those readers of this work who are in the midst of their own thesis and consider applying design science methodology can use this work as a guideline. For this, extra thought was put to showcase the analytical process leading to the development and validation of the model developed. Also, emphasis was placed on explaining the choices made and the resulting trade-offs as transparently as possible following the Design Science principles.

For building management practitioners, the work acts as a case in point about the possibilities offered by analytics in building management. It highlights a process from problem identification to solution and testing, as well as provides concrete, code-level representation for a real-life prediction solution using state-of-the-art open source software. The model is developed to be easily customizable for a class of similar prediction problems, e.g., by offering easily changeable prediction horizon. All in all, the gradient boosting machine framework is especially flexible for different kinds of prediction situations. I hope managers find spark – and financial commitment – to start analytics experimentation in their corporations. Fitting the theme of prediction, implications for the future changes in the building management landscape are portrayed. This work exhibits of the power of open source technologies accessible to anyone with a computer and an internet connection. A pivotal aspect to understand about the present state of technology. The work also builds synthesis of data management and energy systems literature. Hence, the gap with lacking system level data management literature – discussed in chapter 3.3 – is narrowed. This domain is still in dire need of future research, however.

This work is among the few AI projects done in the whole Granlund organisation. As such it offers a concrete viewpoint into the possibilities offered by AI. The work highlights future project possibilities that can be built directly on top of the model developed to reach concrete business value. For example, incorporating the model into the energy consumption monitoring workflow at Granlund automated anomaly detection can be achieved. Even more important than point projects, however, is the develop of an AI strategy. The Digital Transformation Framework by Matt et al. (2015) is discussed in Granlund's context and concrete steps are laid out for the management to get on their way to start building AI strategy needed to systematically harness the power of AI in the organisation. Resource-based view by Barney (1991) is also introduced as a tool to evaluate the value of Granlund's resources regarding data. Thorough elaboration on the code used is provided in the Jupyter Notebook file found in the author's Github (github.com/HVKukkonen). The work yet to be done is also outlined and discussed in the chapter Implementation plan next.

10.6 Implementation plan

As discussed in section six, this work provides only the statistical model – the "brain" so to say – for a prediction-based application. Developing the workflow needed for such an application is out of the scope for this study. That being said, a preliminary implementation plan for the solution is developed in figure 38. This plan should be regarded only as an illustration of a potential set-up. It is primarily directed to people with business background for them to better understand the solution context and help in planning future work.

The model is written in Python with Jupyter Notebook. Notebooks are interactive documents for developing, documenting, and executing code and communicating the results. They can include computations, text, images, and mathematical formulas. (Jupyter Team 2020.) The notebook environment suits well model training and verification of results. The training should be done with data from the actual use case, e.g., a dataset with consumption from the buildings for which the model is used. However, to incorporate the model into a workflow of an application, it should be extracted from the notebook as a Python script. Python, or other tools, can be utilised to create the workflow around the script. The tools planned to use are also incorporated in figure 38. After the system is deployed, its performance needs to be monitored and retraining commenced should there be drop in performance. Model deployment among the intended usage of the model and its potential limitations were discussed in a meeting with the Granlund project management team and other personnel.

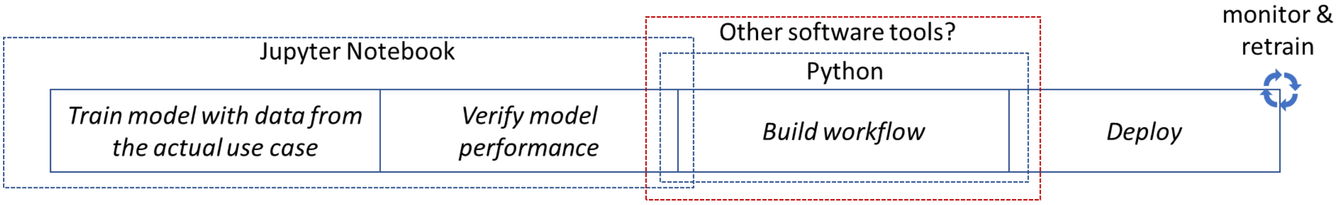


Figure 38. Illustration of future work needed for model deployment.

10.7 Limitations and future research

This work only scratches the surface of what analytics has to offer in the building management space. As only main metering data is used, the model is applicable to the overall energy consumption of a building and not the individual sources. Further studies could research an adaptation of the model for submetering data, e.g., a HVAC system. Similar methodology should be applicable to such systems. The need to retrain the model at least for different kinds of systems needs to be noted, however.

Separate models – in addition to different forms of energy – could also have been develop for different buildings. This is not regarded necessary although model performance could be enhanced, however. This is because drawing conclusions from a plethora of models would have been more ambiguous. The amount of training data available for a model would have shrunk proportionately also. Given the current performance this was not regarded fruitful. Experimentation with such an approach could be appropriate if even better performance is needed, however. For example, this should be tested if the proportionately larger error present with lower consumption levels – please, note the discussion in 9.1 – turns out to be a problem.

Although there is substantial amount of data in terms of data rows, this data originates only from a couple dozen of buildings – 20 and 31 for heating and electricity, respectively, to be exact – all located in the same municipality. Thus, the performance of the model might be different given different environment. That being said, historical consumption clearly dominates environmental characteristics in explaining the model behaviour as discussed in 9.1. Hence, the model is likely to perform also in different environment. Of course, a drastic change in the environment could potentially change the relationship of past and future consumptions. Similar discussion can be had about occupancy characteristics. The buildings are all in non-residential use. Thus, applying the model to a residential building might result in lower performance. Moreover, no occupancy-related features in the data are present who might be used to improve performance. The argument about the historical consumption driving the prediction applies here also, however. Renovation on of a building could similarly affect the suitability of the model for that building.

Considering the forecasting horizon of the model – which is fixed 24 hours – one might argue it to be suboptimal for anomaly detection as discussed in chapter 6.3. This is considered in the model development by making the horizon setting to depend on only one user-customizable variable. Hence, the model is easily optimised for anomaly detection or any other application requiring different prediction horizon. The longer horizon was asked for by Granlund to assess the predictive power over longer timespans. The analysis of the results with 24-hour horizon proved the predictive power sufficient for shorter timespans as well.

Anomaly detection as a concept is still rather vague as the difference between an anomaly and normal fluctuation is dependent on the use case. As discussed in section six, an anomaly in the building energy consumption context is defined as a rapid change, clearly distinct from normal use that progresses quickly over a couple of hours. The model is developed and verified to have sufficient performance to detect such changes. Should the change be different, e.g., progress slowly over the weeks and months, the model would not be suitable for such a case. Thus, in model deployment it needs to be confirmed that the objective corresponds to a case for which the performance is verified. Should this not be the case additional investigation is needed before deploying the model into production.

In Dresch et al. (2015) the researcher is advised to create a pool of potential solutions as a part of the Design Science research process that can be compared and ranked by the solution criteria. This would have surely been useful in this study as well. However, due to time constraints this was deemed to bring too little value given the good and thoroughly verified performance of the model developed. Moreover, during model training and construction a hyperparameter search was done which in fact creates multiple models and compares their performance. This comparison is, of course, done only among gradient boosting machine models, however. On the other hand, given the use of model such expressive as the gradient boosting machine model developed here, it is unlikely that remarkably better results would have been obtained with something else.

Although the integration of the model into a workflow was left out of scope for this work, an implementation plan was formulated. The plan was also found valid and useful by the experts interviewed. In addition to the considerable workload carrying out the implementation would have required; the author did not have access to modify internal systems. The model is developed mindful of the efficiency requirements regarding the size of data flow in an actual production environment. Hence, the actual model needs little development. The findings of this study also support the work of other teams at Granlund working on prediction.

The discussion about the managerial implications of artificial intelligence -powered prediction only scratched the surface of the subject. That being said, the discussion in chapter 3.2 offers a pivotal view on the relevance of the AI in today's business. Moreover, it gives the potential readers higher in the organisational ladder context and frameworks which they can utilise to consider the allocation of organisational resources. This subject clearly requires additional research due to its importance and timeliness. Moreover, the references in 3.2 consider only large global organisations. This introduces a limitation to the trustworthiness of the implications given the context of this study as resources such as capital and talent are surely scarcer in Finnish companies than in global enterprises. Moreover, the potential special characteristics of the Finnish building sector are not addressed. The market can be considerably less developed in term of advanced analytics. A thesis of similar length could be dedicated to the subject by itself. All in all, this offers an avenue for future research.

Bibliography

Accenture, 2018. Flex and balances - Unlocking value from demand-side flexibility in the European power system.

Ahmad, T., Chen, H., Guo, Y. & Wang, J., 2018. A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review. *Energy and Buildings,* Volume 165, pp. 301-320.

Akiba, T. et al., 2019. *Optuna: A next-generation hyperparameter optimization framework.* s.l., s.n., p. 2623–2631.

Alaybeyi, S., Linden, A. & den Hamer, P., 2020. 3 Types of Machine Learning for the Enterprise. *Gartner.*

Amasyali, K. & El-Gohary, N. M., 2018. A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews,* Volume 81, pp. 1192-1205.

Amos, S. W. & Amos, R. S., 1999. *D-Type Connector to Dynode.* s.l.:Elsevier.

ASHRAE, 2014. *ASHRAE Guideline 14–2014, Measurement of Energy, Demand, and Water Savings.* s.l.:ASHRAE Atlanta.

Ashton, K., 2009. That 'Internet of Things' Thing. *RFID JOURNAL.*

Baek, J. et al., 2014. A secure cloud computing based framework for big data information management of smart grid. *IEEE transactions on cloud computing,* Volume 3, p. 233–244.

Bansal, S. K. & Kagemann, S., 2015. Integrating big data: A semantic extract-transform-load framework. *Computer,* Volume 48, p. 42–50.

Barney, J., 1991. Firm resources and sustained competitive advantage. *Journal of management,* Volume 17, p. 99–120.

Barrot, J.-N., Grassi, B. & Sauvagnat, J., 2020. Sectoral effects of social distancing. *Available at SSRN.*

BCG, 2020a. Engineering, Construction, and Infrastructure.

BCG, 2020b. AI at Scale: The Next Frontier in Digital Transformation. *Digital, Technology, and Data.*

Béland, L.-P., Brodeur, A. & Wright, T., 2020. The short-term economic consequences of Covid-19: exposure to disease, remote work and government response.

Bourdeau, M. et al., 2019. Modeling and forecasting building energy consumption: A review of data-driven techniques. *Sustainable Cities and Society,* Volume 48.

Breiman, L., 1996. Bagging predictors. *Machine learning,* Volume 24, p. 123–140.

Buckman, A. H., Mayfield, M. & Beck, S. B. M., 2014. What is a smart building?. *Smart and Sustainable Built Environment.*

Bughin, J. & Van Zeebroeck, N., 2018. Artificial intelligence: Why a digital base is critical. *The McKinsey Quarterly.*

Burger, S. P. & Luke, M., 2017. Business models for distributed energy resources: A review and empirical analysis. *Energy Policy,* Volume 109, p. 230–248.

Cammack, R. et al., 2006. *relational database.* s.l.:Oxford University Press.

Chaudhuri, S., Dayal, U. & Ganti, V., 2001. Database technology for decision support systems. *Computer,* Volume 34, p. 48–55.

Chaudhuri, S., Dayal, U. & Narasayya, V., 2011. An overview of business intelligence technology. *Communications of the ACM,* Volume 54, p. 88–98.

Chen, H. et al., 2009. *The design and implementation of a smart building control system.* s.l., s.n., p. 255–262.

Dayal, U., Castellanos, M., Simitsis, A. & Wilkinson, K., 2009. *Data integration flows for business intelligence.* s.l., s.n., p. 1–11.

DiCiccio, T.J., Efron, B., 1996. Bootstrap Confidence Intervals. Statistical Science, Vol. 11, No. 3 (Aug. 1996), pp. 189-212. Institute of Mathematical Statistics. https://www.jstor.org/stable/2246110.

DiCicco-Bloom, B. & Crabtree, B. F., 2006. The qualitative research interview. *Medical education,* Volume 40, p. 314–321.

Dresch, A., Lacerda, D. P. & Antunes, J. A. V., 2015. General Aspects Related to Research in Management. In: *Design Science Research: A Method for Science and Technology Advancement.* Cham: Springer International Publishing, p. 1–10.

Drucker, H., 1997. *Improving regressors using boosting techniques.* s.l., s.n., p. 107–115.

Escrivá-Escrivá, G., Álvarez-Bel, C., Roldán-Blay, C. & Alcázar-Ortega, M., 2011. New artificial neural network prediction method for electrical consumption forecasting based on building end-uses. *Energy and Buildings,* Volume 43, p. 3112–3119.

EU Buildings Observatory, 2020. EU Buildings Database. *European Commission.*

Eurostat, 2020a. Energy statistics - supply, transformation and consumption. *Complete energy balances - annual data (nrg_110a).*

Eurostat, 2020b. Glossary: Final energy consumption. *Eurostat - Statistics Explained.*

Fabi, V. et al., 2011. *Description of occupant behaviour in building energy simulation: State-of-art and concepts for improvements.* s.l., s.n., pp. 2882-2889.

Fana, M., Pérez, S. T. & Fernández-Macías, E., 2020. Employment impact of Covid-19 crisis: from short term effects to long terms prospects. *Journal of Industrial and Business Economics,* p. 1–20.

Fan, C., Wang, J., Gang, W. & Li, S., 2019. Assessment of deep recurrent neural network-based strategies for short-term building energy predictions. *Applied Energy,* Volume 236, pp. 700-710.

Ferreira, H. C., Lampe, L., Newbury, J. & Swart, T. G., 2010. *Power line communications: theory and applications for narrowband and broadband communications over power lines.* s.l.:John Wiley & Sons.

Fingrid, 2020. Datahub. *https://palvelut.datahub.fi/en/datahub/general-info.*

Foucquier, A. et al., 2013. State of the art in building modelling and energy performances prediction: A review. *Renewable and Sustainable Energy Reviews,* Volume 23, p. 272–288.

Knight Frank, 2019. Global Outlook: Knight Frank's Chief Economist shares his outlook for real estate investment markets. *knightfrank.co.uk.*

Freund, Y., Schapire, R. E. & others, 1996. *Experiments with a new boosting algorithm.* s.l., s.n., p. 148–156.

Friedman, J. H., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics,* p. 1189–1232.

Gartner, 2019. Hype Cycle for Smart City Technologies and Solutions, 2019. *Gartner.*

Geisser, S., 1975. The predictive sample reuse method with applications. *Journal of the American statistical Association,* Volume 70, p. 320–328.

Géron, A., 2019. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition.. *O'Reilly Media, Inc.*

Giordano, A. et al., 2019. An Energy Community Implementation: The Unical Energy Cloud. *Electronics,* Volume 8, p. 1517.

Gutierrez, D. D., 2015. *Machine learning and data science: an introduction to statistical learning methods with R.* s.l.:Technics Publications.

Hashem, I. A. T. et al., 2015. The rise of "big data" on cloud computing: Review and open research issues. *Information systems,* Volume 47, p. 98–115.

Hernandez, L. et al., 2014. A survey on electric power demand forecasting: future trends in smart grids, microgrids and smart buildings. *IEEE Communications Surveys & Tutorials,* Volume 16, p. 1460–1495.

Hernandez, M. J., 2013. *Database design for mere mortals: a hands-on guide to relational database design.* s.l.:Pearson Education.

Hevner, A. R., March, S. T., Park, J. & Ram, S., 2004. Design Science in Information Systems Research. *MIS Quarterly,* Volume 28, p. 75–105.

Huebner, M., le Cessie, S., Schmidt, C. O. & Vach, W., 2018. A contemporary conceptual framework for initial data analysis. *Obs Stud,* Volume 4, p. 171–192.

Huebner, M., Vach, W. & [le Cessie], S., 2016. A systematic approach to initial data analysis is good research practice. *The Journal of Thoracic and Cardiovascular Surgery,* Volume 151, pp. 25-27.

Ikäheimo, J., Evens, C. & Kärkkäinen, S., 2010. DER Aggregator business: the Finnish case. *Technical Research Centre of Finland (VTT): Espoo, Finland.*

Jirdehi, M. A., Tabar, V. S., Ghassemzadeh, S. & Tohidi, S., 2020. Different aspects of microgrid management: A comprehensive review. *Journal of Energy Storage,* Volume 30.

Kastner, W., Neugschwandtner, G., Soucek, S. & Newman, H. M., 2005. Communication systems for building automation and control. *Proceedings of the IEEE,* Volume 93, pp. 1178-1203.

Ke, G. et al., 2017. *Lightgbm: A highly efficient gradient boosting decision tree.* s.l., s.n., p. 3146–3154.

Killian, M. & Kozek, M., 2016. Ten questions concerning model predictive control for energy efficient buildings. *Building and Environment,* Volume 105, p. 403–412.

Kiran, M. et al., 2015. *Lambda architecture for cost-effective batch and speed big data processing.* s.l., s.n., p. 2785–2792.

Koschwitz, D., Spinnräker, E., Frisch, J. & van Treeck, C., 2020. Long-term urban heating load predictions based on optimized retrofit orders: A cross-scenario analysis. *Energy and Buildings,* Volume 208, p. 109637.

Krishnadas, G. & Kiprakis, A., 2020. A machine learning pipeline for demand response capacity scheduling. *Energies,* Volume 13.

Kugley, S. et al., 2016. Searching for studies: A guide to information retrieval for Campbell. *Campbell Systematic Reviews.*

Kukkonen, V., 2017. SÄHKÖVERKKOLIIKETOIMINNAN UUDISTAMINEN HAJAUTETUN SÄHKÖNTUOTANNON MAHDOLLISTAMISEKSI. *Bachelor's thesis, Aalto University.*

L'heureux, A., Grolinger, K., Elyamany, H. F. & Capretz, M. A. M., 2017. Machine learning with big data: Challenges and approaches. *IEEE Access,* Volume 5, p. 7776–7797.

Lamprinos, I., Hatziargyriou, N. D., Kokos, I. & Dimeas, A. D., 2016. Making Demand Response a Reality in Europe: Policy, Regulations, and Deployment Status. *IEEE Communications Magazine,* Volume 54, pp. 108-113.

Lawrence, M. & Vrins, J., 2018. Energy Cloud 4.0. *Capturing Business Value through Disruptive Energy Platforms.*

Lawrence, T. M. et al., 2016. Ten questions concerning integrating smart buildings into the smart grid. *Building and Environment,* Volume 108, p. 273–283.

Litiu, A. et al., 2017. REHVA Guidebook No. 22 - Introduction to Building Automation, Controls and Technical Building Management.. *REHVA.*

Li, W., 2019. Application of Economical Building Management System for Singapore Commercial Building. *IEEE Transactions on Industrial Electronics,* Volume 67, p. 4235–4243.

Li, Z., Han, Y. & Xu, P., 2014. Methods for benchmarking building energy consumption against its past or intended performance: An overview. *Applied Energy,* Volume 124, p. 325–334.

Loh, W.-Y., 2011. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* Volume 1, p. 14–23.

Loh, W.-Y., 2014. Fifty years of classification and regression trees. *International Statistical Review,* Volume 82, p. 329–348.

Loh, W.-Y. & Shih, Y.-S., 1997. Split selection methods for classification trees. *Statistica sinica,* p. 815–840.

Lopes, J. A. P., Moreira, C. L. & Madureira, A. G., 2006. Defining control strategies for MicroGrids islanded operation. *IEEE Transactions on Power Systems,* Volume 21, pp. 916-924.

Lundberg, S. M. et al., 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence,* Volume 2, pp. 2522-5839.

Lu, X. et al., 2020. Fundamentals and business model for resource aggregator of demand response in electricity markets. *Energy,* p. 117885.

Maine, D., Keech, A. & Duchnowski, S., 2020. Memo to CEOs: Break Your Inertia on Sustainability. *Bain & Company.*

Marz, N. & Warren, J., 2015. *Big Data: Principles and best practices of scalable real-time data systems.* s.l.:New York; Manning Publications Co..

Matt, C., Hess, T. & Benlian, A., 2015. Digital transformation strategies. *Business & Information Systems Engineering,* Volume 57, p. 339–343.

Ma, Z., Billanes, J. D. & Jørgensen, B. N., 2017. Aggregation potentials for buildings—business models of demand response and virtual power plants. *Energies,* Volume 10, p. 1646.

McKinsey & Company, 2020. Pathways to a low-carbon Europe | Sustainability.

Mell, P., Grance, T. & others, 2011. The NIST definition of cloud computing.

Merriam-Webster, 2020. “Analysis.”. *Merriam-Webster.com Dictionary.*

Microsoft, 2020. LightGBMRelease 2.3.2. *LightGBM Documentation.*

Minoli, D., Sohraby, K. & Occhiogrosso, B., 2017. IoT considerations, requirements, and architectures for smart buildings—Energy optimization and next-generation building management systems. *IEEE Internet of Things Journal,* Volume 4, p. 269–283.

Mittal, S., Tam, W. T. & Ko, C., 2018. Internet of Things: The Pillar of Artificial Intelligence. *Report produced by Asian Insights Office: DBS Group.*

Mohammadi, F. G., Arabnia, H. R. & Amini, M. H., 2019. *On Parameter Tuning in Meta-Learning for Computer Vision.* s.l., s.n., p. 300–305.

Mohandes, S. R., Zhang, X. & Mahdiyar, A., 2019. A comprehensive review on the application of artificial neural networks in building energy analysis. *Neurocomputing,* Volume 340, pp. 55-75.

Morgan, J. N. & Sonquist, J. A., 1963. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association,* Volume 58, p. 415–434.

Müller, A. C., Guido, S. & others, 2016. *Introduction to machine learning with Python: a guide for data scientists.* s.l.:" O'Reilly Media, Inc.".

Munshi, A. A. & Mohamed, Y. A.-R. I., 2018. Data lake lambda architecture for smart grids big data analytics. *IEEE Access,* Volume 6, p. 40463–40471.

Musleh, A. S., Yao, G. & Muyeen, S. M., 2019. Blockchain applications in smart grid–review and frameworks. *IEEE Access,* Volume 7, p. 86746–86757.

Myers, M. D. & Newman, M., 2007. The qualitative interview in IS research: Examining the craft. *Information and organization,* Volume 17, p. 2–26.

Myles, A. J. et al., 2004. An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society,* Volume 18, p. 275–285.

Nakamoto, S. & others, 2008. Bitcoin: A peer-to-peer electronic cash system.

Nauclér, T. & Enkvist, P.-A., 2009. Pathways to a low-carbon economy: Version 2 of the global greenhouse gas abatement cost curve. *McKinsey & Company,* Volume 192.

Newaz, S. S. et al., 2014. *A web based energy cloud platform for campus smart grid for understanding energy consumption profile and predicting future energy demand.* s.l., s.n., p. 173–178.

next-kraftwerke.be, 2020. The role of the BRP. *https://www.next-kraftwerke.be/en/knowledge-hub/brps-and-portfolio-nominations/.*

Nielsen, C., 2015. The sustainability imperative: new insights on consumer expectations. *Nielsen Company New York.*

Official Statistics of Finland (OSF), 2017. Energy Accounts [e-publication]. *Helsinki: Statistics Finland.*

Official Journal of the European Union, 2010. Directive 2010/31/EU of the European Parliament and of the Council of 19 May 2010 on the energy performance of buildings (recast). *Official Journal of the European Union,* Volume 18, p. 2010.

Power Authority of the State of New York, 2013. Build Smart NY. *EXECUTIVE ORDER 88 GUIDELINES, NEW YORK STATE GOVERNMENT BUILDINGS.*

Palensky, P. & Dietrich, D., 2011. Demand side management: Demand response, intelligent energy systems, and smart loads. *IEEE Transactions on Industrial Informatics,* Volume 7, pp. 381-388.

Pan, Y. & Zhang, L., 2020. Data-driven estimation of building energy consumption with multi-source heterogeneous data. *Applied Energy,* Volume 268.

Papadopoulos, S., Azar, E., Woon, W.-L. & Kontokosta, C. E., 2018. Evaluation of tree-based ensemble learning algorithms for building energy performance estimation. *Journal of Building Performance Simulation,* Volume 11, pp. 322-332.

Peffers, K., Tuunanen, T., Rothenberger, M. A. & Chatterjee, S., 2007. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems,* Volume 24, pp. 45-77.

Porter, M. E., 1985. *Competitive advantage: Creating and sustaining competitive advantage.* s.l.:New York: Free Press.

Prouzeau, A. et al., 2018. *Visual Analytics for Energy Monitoring in the Context of Building Management.* s.l., s.n., p. 1–9.

PwC, 2020. Emerging Trends in Real Estate®: Europe 2020.

Qiu, J. et al., 2016. A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing,* Volume 2016, p. 67.

Ransbotham, S. et al., 2018. Artificial intelligence in business gets real. *MIT Sloan Management Review and Boston Consulting Group.*

Ravulavaru, A., 2018. *1.1 Internet of Things.* s.l.:Packt Publishing.

Reunanen, J., 2003. Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research,* Volume 3, p. 1371–1382.

Roldán-Blay, C. et al., 2013. Upgrade of an artificial neural network prediction method for electrical consumption forecasting using an hourly temperature curve model. *Energy and Buildings,* Volume 60, p. 38–46.

Rubel, H., Baker, T., Zenneck, J. & Aubert, G., 2017. Finding the Sweet Spot in Distributed Energy. *BCG Perspectives.*

Russo, M. & Wang, G., 2019. The Incumbent's Advantage in the Internet of Things. *BCG Henderson Institute.*

Sánchez-Jiménez, M., 2006. European SmartGrids Technology Platform. Vision and Strategy for Europe's Electricity Networks of the Future.. *European Commission: Directorate-General for Research.*

Sequeira, H., Carreira, P., Goldschmidt, T. & Vorst, P., 2014. *Energy cloud: Real-time cloud-native energy management system to monitor and analyze energy consumption in multiple industrial sites.* s.l., s.n., p. 529–534.

Seuring, S. & Gold, S., 2012. Conducting content-analysis based literature reviews in supply chain management. *Supply Chain Management: An International Journal.*

Sioshansi, F. P., 2013. *Energy efficiency: towards the end of demand growth.* s.l.:Academic Press.

Smitha Rao, M. S., Pallavi, M. & Geetha, N., 2019. *Conceptual Machine Learning Framework for Initial Data Analysis.* Singapore, Springer Singapore, p. 51–59.

Strasser, T. et al., 2014. A review of architectures and concepts for intelligence in future electric energy systems. *IEEE Transactions on Industrial Electronics,* Volume 62, p. 2424–2438.

Su, W. & Huang, A., 2018. *The Energy Internet: An Open Energy Platform to Transform Legacy Power Systems into Open Innovation and Global Economic Engines.* s.l.:Woodhead Publishing.

Taieb, S. B. & Hyndman, R. J., 2014. A gradient boosting approach to the Kaggle load forecasting competition. *International Journal of Forecasting,* Volume 30, pp. 382-394.

Taylor, M., Watts, J. & Bartlett, J., 2019. Climate crisis: 6 million people join latest wave of global protests. *The Guardian,* Volume 28.

Team, J., 2020. Jupyter Notebook 7.0.0.dev0 documentation. *Jupyter-notebook.readthedocs.io.*

The Economist, 2017. The world's most valuable resource is no longer oil, but data. *The Economist: New York, NY, USA.*

Thomas, S. & Rosenow, J., 2020. Drivers of increasing energy consumption in Europe and policy implications. *Energy Policy,* Volume 137, p. 111108.

Thornhill, A., Saunders, M. & Lewis, P., 2009. *Research methods for business students.* s.l.:Prentice Hall: London.

Tilastokeskus, 2020. Final energy consumption by sector. *Statistics Finland's PxWeb databases.*

Toller, S. et al., 2011. Energy use and environmental impacts of the Swedish building and real estate management sector. *Journal of Industrial Ecology,* Volume 15, p. 394–404.

Tommerup, H. & Svendsen, S., 2006. Energy savings in Danish residential building stock. *Energy and buildings,* Volume 38, p. 618–626.

Tomsic, J. L., 2000. *relational data base.* s.l.:SAE International.

Torpp, Ø. & Rød, K., 2017. Green Leadership: A Practical Guide to Winning in the Green Economy.

Touzani, S., Granderson, J. & Fernandes, S., 2018. Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings,* Volume 158, pp. 1533-1543.

United Nations, 2018. United Nations, Department of Economic and Social Affairs, Population Division (2018). *World Urbanization Prospects: The 2018 Revision, Online Edition.*

Vanhove, T. et al., 2016. Managing the synchronization in the lambda architecture for optimized big data analysis. *IEICE Transactions on Communications,* Volume 99, p. 297–306.

Varma, S. & Simon, R., 2006. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics,* Volume 7, p. 91.

Verhoeven, R., 2009. Pathways to World-Class energy efficiency in Belgium. *McKinsey & Company-2009, Belgium.*

Wang, R., Lu, S. & Li, Q., 2019. Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consumption for residential buildings. *Sustainable Cities and Society,* Volume 49.

Wang, Z. & Srinivasan, R. S., 2017. A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. *Renewable and Sustainable Energy Reviews,* Volume 75, pp. 796-808.

Warner, M., 2013. *Understanding the intelligence cycle.* s.l.:Routledge.

Wetzstein, S., 2017. The global urban housing affordability crisis. *Urban Studies,* Volume 54, p. 3159–3177.

Wille-Haussmann, B., Erge, T. & Wittwer, C., 2010. Decentralised optimisation of cogeneration in virtual power plants. *Solar Energy,* Volume 84, pp. 604-611.

Woetzel, J. R., 2014. A blueprint for addressing the global affordable housing challenge.

World Economic Forum, 2016. *Shaping the Future of Construction: a Breakthrough in Mindset and Technology.* s.l.:WEF Cologny, Switzerland.

Ympäristöministeriö, 1978. Rakennusten energiatalous, Määräykset ja ohjeet. *Suomen Rakentamismääräyskokoelma D3.*

Ympäristöministeriö, 1985. Lämmöneristys, Määräykset 1985. *Suomen Rakentamismääräyskokoelma C3.*

Ympäristöministeriö, 2003. Rakennuksen lämmöneristys, Määräykset 2003. *Suomen Rakentamismääräyskokoelma C3.*

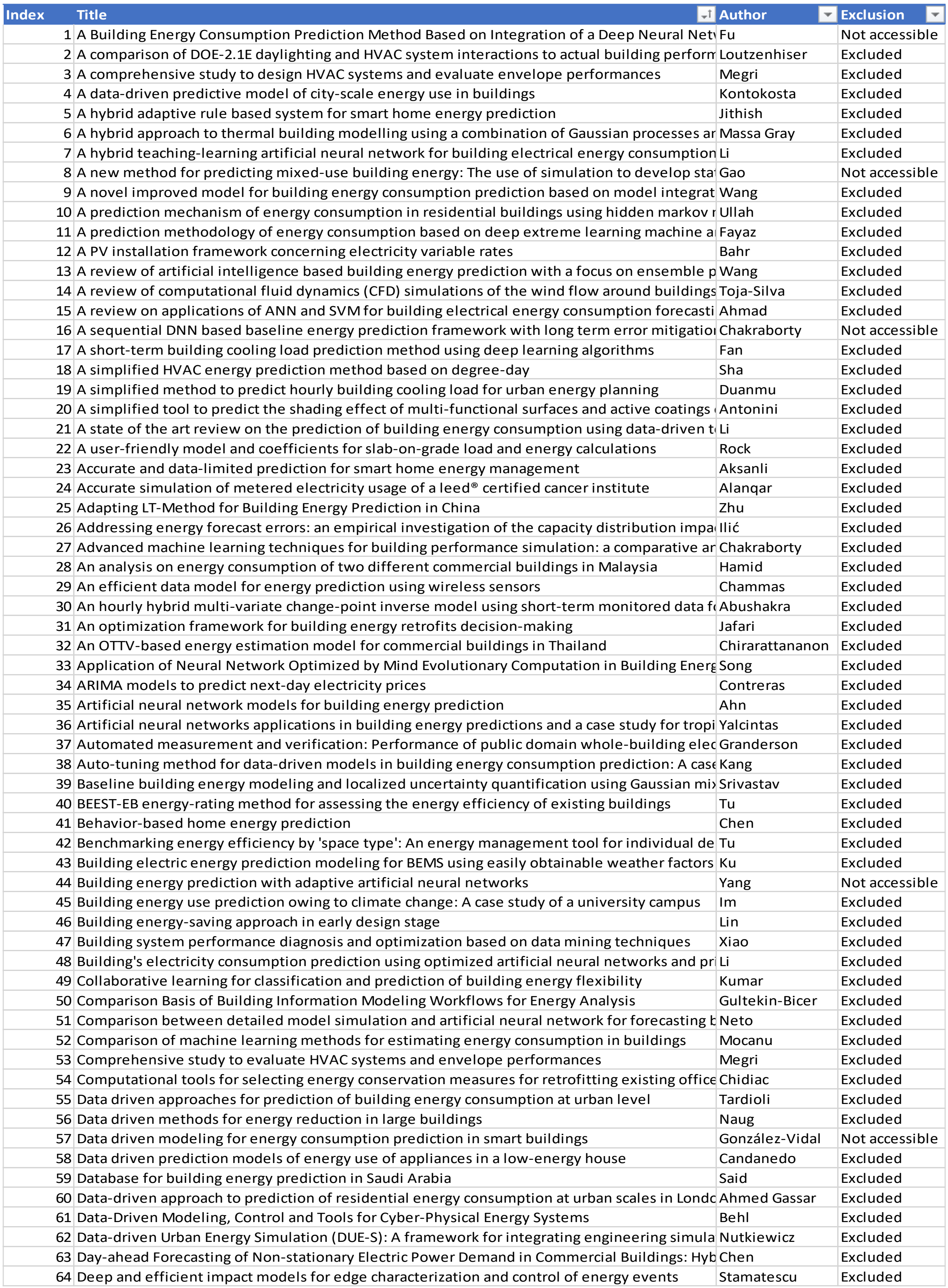
Ympäristöministeriö, 2007. Rakennuksen lämmöneristys, Määräykset 2007. *Suomen Rakentamismääräyskokoelma C3.*

Ympäristöministeriö, 2010. Rakennusten lämmöneristys, Määräykset 2010. *Suomen Rakentamismääräyskokoelma C3.*

Ympäristöministeriö, 2020. https://www.ym.fi/fi-FI/Maankaytto_ja_rakentaminen/Lainsaadanto_ja_ohjeet/Rakennuksen_energiatehokkuutta_koskeva_lainsaadanto

Zheng, Z. et al., 2017. *An overview of blockchain technology: Architecture, consensus, and future trends.* s.l., s.n., p. 557–564.

Appendix A. Systematic literature review: excluded articles

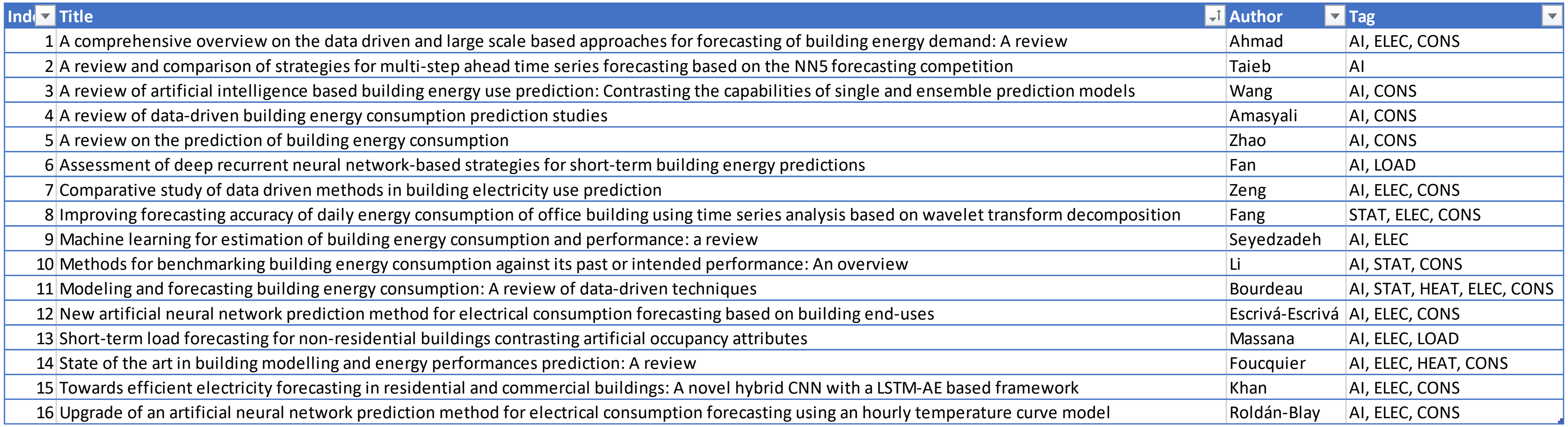


Index	Title	Author	Exclusion
1	A Building Energy Consumption Prediction Method Based on Integration of a Deep Neural Net	Fu	Not accessible
2	A comparison of DOE-2.1E daylighting and HVAC system interactions to actual building perform	Loutzenhiser	Excluded
3	A comprehensive study to design HVAC systems and evaluate envelope performances	Megri	Excluded
4	A data-driven predictive model of city-scale energy use in buildings	Kontokosta	Excluded
5	A hybrid adaptive rule based system for smart home energy prediction	Jithish	Excluded
6	A hybrid approach to thermal building modelling using a combination of Gaussian processes ar	Massa Gray	Excluded
7	A hybrid teaching-learning artificial neural network for building electrical energy consumption	Li	Excluded
8	A new method for predicting mixed-use building energy: The use of simulation to develop stat	Gao	Not accessible
9	A novel improved model for building energy consumption prediction based on model integrat	Wang	Excluded
10	A prediction mechanism of energy consumption in residential buildings using hidden markov	Ullah	Excluded
11	A prediction methodology of energy consumption based on deep extreme learning machine a	Fayaz	Excluded
12	A PV installation framework concerning electricity variable rates	Bahr	Excluded
13	A review of artificial intelligence based building energy prediction with a focus on ensemble p	Wang	Excluded
14	A review of computational fluid dynamics (CFD) simulations of the wind flow around buildings	Toja-Silva	Excluded
15	A review on applications of ANN and SVM for building electrical energy consumption forecasti	Ahmad	Excluded
16	A sequential DNN based baseline energy prediction framework with long term error mitigatio	Chakraborty	Not accessible
17	A short-term building cooling load prediction method using deep learning algorithms	Fan	Excluded
18	A simplified HVAC energy prediction method based on degree-day	Sha	Excluded
19	A simplified method to predict hourly building cooling load for urban energy planning	Duanmu	Excluded
20	A simplified tool to predict the shading effect of multi-functional surfaces and active coatings	Antonini	Excluded
21	A state of the art review on the prediction of building energy consumption using data-driven t	Li	Excluded
22	A user-friendly model and coefficients for slab-on-grade load and energy calculations	Rock	Excluded
23	Accurate and data-limited prediction for smart home energy management	Aksanli	Excluded
24	Accurate simulation of metered electricity usage of a leed® certified cancer institute	Alanqar	Excluded
25	Adapting LT-Method for Building Energy Prediction in China	Zhu	Excluded
26	Addressing energy forecast errors: an empirical investigation of the capacity distribution impa	Ilić	Excluded
27	Advanced machine learning techniques for building performance simulation: a comparative ar	Chakraborty	Excluded
28	An analysis on energy consumption of two different commercial buildings in Malaysia	Hamid	Excluded
29	An efficient data model for energy prediction using wireless sensors	Chammas	Excluded
30	An hourly hybrid multi-variate change-point inverse model using short-term monitored data f	Abushakra	Excluded
31	An optimization framework for building energy retrofits decision-making	Jafari	Excluded
32	An OTTV-based energy estimation model for commercial buildings in Thailand	Chirarattananon	Excluded
33	Application of Neural Network Optimized by Mind Evolutionary Computation in Building Energ	Song	Excluded
34	ARIMA models to predict next-day electricity prices	Contreras	Excluded
35	Artificial neural network models for building energy prediction	Ahn	Excluded
36	Artificial neural networks applications in building energy predictions and a case study for tropi	Yalcintas	Excluded
37	Automated measurement and verification: Performance of public domain whole-building elec	Granderson	Excluded
38	Auto-tuning method for data-driven models in building energy consumption prediction: A case	Kang	Excluded
39	Baseline building energy modeling and localized uncertainty quantification using Gaussian mix	Srivastav	Excluded
40	BEEST-EB energy-rating method for assessing the energy efficiency of existing buildings	Tu	Excluded
41	Behavior-based home energy prediction	Chen	Excluded
42	Benchmarking energy efficiency by 'space type': An energy management tool for individual de	Tu	Excluded
43	Building electric energy prediction modeling for BEMS using easily obtainable weather factors	Ku	Excluded
44	Building energy prediction with adaptive artificial neural networks	Yang	Not accessible
45	Building energy use prediction owing to climate change: A case study of a university campus	Im	Excluded
46	Building energy-saving approach in early design stage	Lin	Excluded
47	Building system performance diagnosis and optimization based on data mining techniques	Xiao	Excluded
48	Building's electricity consumption prediction using optimized artificial neural networks and pri	Li	Excluded
49	Collaborative learning for classification and prediction of building energy flexibility	Kumar	Excluded
50	Comparison Basis of Building Information Modeling Workflows for Energy Analysis	Gultekin-Bicer	Excluded
51	Comparison between detailed model simulation and artificial neural network for forecasting b	Neto	Excluded
52	Comparison of machine learning methods for estimating energy consumption in buildings	Mocanu	Excluded
53	Comprehensive study to evaluate HVAC systems and envelope performances	Megri	Excluded
54	Computational tools for selecting energy conservation measures for retrofitting existing office	Chidiac	Excluded
55	Data driven approaches for prediction of building energy consumption at urban level	Tardioli	Excluded
56	Data driven methods for energy reduction in large buildings	Naug	Excluded
57	Data driven modeling for energy consumption prediction in smart buildings	González-Vidal	Not accessible
58	Data driven prediction models of energy use of appliances in a low-energy house	Candanedo	Excluded
59	Database for building energy prediction in Saudi Arabia	Said	Excluded
60	Data-driven approach to prediction of residential energy consumption at urban scales in Londo	Ahmed Gassar	Excluded
61	Data-Driven Modeling, Control and Tools for Cyber-Physical Energy Systems	Behl	Excluded
62	Data-driven Urban Energy Simulation (DUE-S): A framework for integrating engineering simula	Nutkiewicz	Excluded
63	Day-ahead Forecasting of Non-stationary Electric Power Demand in Commercial Buildings: Hyb	Chen	Excluded
64	Deep and efficient impact models for edge characterization and control of energy events	Stamatescu	Excluded

65	Deep learning-based feature engineering methods for improved building energy prediction	Fan	Excluded
66	Deep-learning neural-network architectures and methods: Using component-based models in	Singaravel	Excluded
67	Detection of low-dimensional chaos in buildings energy consumption time series	Karatasou	Excluded
68	Developing a Hybrid Model of Prediction and Classification Algorithms for Building Energy Con	Banihashemi	Excluded
69	Development and improvement of occupant behavior models towards realistic building perfor	Li	Excluded
70	Development and validation of regression models to predict monthly heating demand for resi	Catalina	Excluded
71	Development of a consecutive occupancy estimation framework for improving the energy dem	Kim	Excluded
72	Development of a generalized neural network model to detect faults in building energy perfor	Breekweg	Not accessible
73	Development of a mobile application for building energy prediction using performance predic	Kim	Excluded
74	Development of a moisture buffer value model (MBM) for indoor moisture prediction	Zu	Excluded
75	Development of a surrogate model by extracting top characteristic feature vectors for building	Sangireddy	Excluded
76	Development of an energy prediction tool for commercial buildings using case-based reasonin	Monfet	Excluded
77	Development of prediction models for next-day building energy consumption and peak power	Fan	Excluded
78	Different occupant modeling approaches for building energy prediction	Ahn	Excluded
79	Domestic building energy prediction in design stage utilizing large-scale consumption data fro	Kim	Excluded
80	DR-Advisor: A data-driven demand response recommender system	Behl	Excluded
81	E2-diagnoser: A system for monitoring, forecasting and diagnosing energy usage	Ploennigs	Excluded
82	Effect of length of measurement period on accuracy of predicted annual heating energy consu	Cho	Excluded
83	Effects of building energy optimisation on the predictive accuracy of external temperature in f	Shiel	Excluded
84	Electric Load Prediction Baselines for Airport Buildings: A Case Study	Kang	Excluded
85	Energy analysis of a building using artificial neural network: A review	Kumar	Excluded
86	Energy consumption models for residential buildings: A case study	Ferrarini	Excluded
87	Energy consumption prediction from usage data for decision support on investments: The EnPI	Neves-Silva	Excluded
88	Energy demand prediction for the implementation of an energy tariff emulator to trigger dema	Noyé	Excluded
89	Energy demand prediction through novel random neural network predictor for large non-dome	Ahmad	Excluded
90	Energy efficiency analysis carried out by installing district heating on a university campus. A cas	Marina Domingo	Excluded
91	Energy forecasting for event venues: Big data and prediction accuracy	Grolinger	Excluded
92	Energy planning and forecasting approaches for supporting physical improvement strategies in	Chalal	Excluded
93	Energy prediction model for smart home using heating degree days	Park	Not accessible
94	Energy prediction of electric floor radiation systems using a new integrated modeling approach	Megri	Excluded
95	Energy Prediction versus Energy Performance of Green Buildings in Malaysia. Comparison of Pr	Zaid	Excluded
96	Energy slices: benchmarking with time slicing	Grolinger	Excluded
97	Evaluation and monitring of energy consumption patterns using statistical modeling and simul	Khalid	Excluded
98	Evaluation of anchor bolt effects on the thermal performance of building insulation materials	Ji	Excluded
99	Fast bidirectional building performance optimization at the early design stage	Li	Excluded
100	Feature selection for support vector regression in the application of building energy prediction	Zhao	Excluded
101	Forecasting building energy consumption based on hybrid PSO-ANN prediction model	Hu	Excluded
102	Forecasting building energy consumption using neural networks and hybrid neuro-fuzzy system	Li	Excluded
103	Forecasting district-scale energy dynamics through integrating building network and long short	Wang	Excluded
104	Forecasting energy consumption from smart home sensor network by deep learning	Dey	Excluded
105	Forecasting energy consumption of multi-family residential buildings using support vector reg	Jain	Excluded
106	Impact of urban microclimate on summertime building cooling demand: A parametric analysis	Toparlar	Excluded
107	Implementing artificial neural networks in energy building applications - A review	Georgiou	Excluded
108	Indoor Air-Temperature Forecast for Energy-Efficient Management in Smart Buildings	Aliberti	Not accessible
109	Information exchange scenarios between machine learning energy prediction model and BIM a	Singh	Excluded
110	Information requirements for multi-level-of-development BIM using sensitivity analysis for en	Singh	Excluded
111	Lecture and non-lecture week baseline energy model development and energy prediction in M	Mustapa	Excluded
112	Legislating building energy performance: Putting EU policy into practice	Raslan	Excluded
113	Leveraging weather forecasts in renewable energy systems	Sharma	Excluded
114	Linear and classification learner models for building energy predictions and predicting saving e	Eaton	Not accessible
115	Microclimate mitigation for enhancing energy and environmental performance of Near Zero En	Cardinali	Excluded
116	Minimising the deviation between predicted and actual building performance via use of neura	Hammad	Excluded
117	Model predictive control strategies for buildings with mixed-mode cooling	Hu	Excluded
118	Modeling and predicting building's energy use with artificial neural networks: Methods and res	Karatasou	Excluded
119	Modeling and predictive control of buildings with distributed energy generation and storage	Li	Excluded
120	Modelling as an accurate indicator of exemplary building performance - Three australian case s	Taylor	Excluded
121	Modelling the occupant behaviour impact on buildings energy prediction	Virote	Excluded
122	Module emendation research in the building energy simulation program of EnergyPlus based o	Xia	Excluded
123	Morphology Parameters Quantitative Research of Multi-storey Office Building Design in Harbin	Sun	Excluded
124	Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consump	Wang	Excluded
125	Multiple power-based building energy management system for efficient management of build	Yoon	Excluded
126	Numerical and experimental results of a novel and generic methodology for energy performan	Lazrak	Excluded
127	Occupant behaviour and robustness of building design	Buso	Excluded
128	On-line building energy prediction using adaptive artificial neural networks	Yang	Excluded
129	Optimised building energy and indoor microclimatic predictions using knowledge-based syste	Ganguly	Excluded

130	Optimized Power Control Methodology Using Genetic Algorithm	Ali	Excluded
131	Performance predictions of ground source heat pump system based on random forest and bacl	Lu	Excluded
132	Photovoltaics energy prediction under complex conditions for a predictive energy managemer	Schmelas	Excluded
133	Planning and Monitoring of Building Energy Demands under Uncertainties by Using IoT Data	Chang	Excluded
134	Poster abstract: Big Data beats engineering in residential energy performance assessment—a (	Fridgen	Excluded
135	Power prediction through energy consumption pattern recognition for smart buildings	Jin	Excluded
136	Predict electric power demand with extended goal graph and heterogeneous mixture modelir	Kushiro	Excluded
137	Predictability of occupant presence and performance gap in building energy simulation	Ahn	Excluded
138	Predicting energy usage using historical data and linear models	Safa	Not accessible
139	Predicting future hourly residential electrical consumption: A machine learning case study	Edwards	Excluded
140	Predicting impact of cooling set-point change on demand reduction in real-time	Lingamallu	Excluded
141	Predicting unusual energy consumption events from smart home sensor network by data strea	Fong	Excluded
142	Prediction methods and precise electricity energy prediction of school facility	Ryu	Not accessible
143	Prediction model based on an artificial neural network for user-based building energy consum	Lee	Excluded
144	Prediction model of Cooling Load considering time-lag for preemptive action in buildings	Lim	Excluded
145	Prediction of building energy consumption by using artificial neural networks	Ekici	Excluded
146	Prediction of building energy needs in early stage of design by using ANFIS	Bektas Ekici	Excluded
147	Prediction of thermal energy inside smart homes using IoT and classifier ensemble techniques	Xu	Excluded
148	Prediction system based on domotic weather sensors for the energy production of solar powe	Benítez	Excluded
149	Prediction-learning algorithm for efficient energy consumption in smart buildings based on pa	Malik	Excluded
150	Predictive model of energy consumption for office building by using improved GWO-BP	Tian	Excluded
151	Pseudo dynamic transitional modeling of building heating energy demand using artificial neur	Paudel	Excluded
152	Quantitative study on environment and energy information for land use planning scenarios in	Yeo	Excluded
153	Random Forest based hourly building energy prediction	Wang	Excluded
154	Re-examination of external temperature as a predictor of energy usage in buildings	Shiel	Excluded
155	Regression and artificial neural network models with data classifications for building energy pr	Nassif	Not accessible
156	Reliability of building embodied energy modelling: An analysis of 30 Melbourne case studies	Langston	Excluded
157	Review on stochastic modeling methods for building stock energy prediction	Lim	Excluded
158	Robustness of building design with respect to energy related occupant behaviour	Fabi	Excluded
159	Selection of climatic variables and time scales for future weather preparation in building heati	Chen	Excluded
160	Self-activating uncertainty analysis for BIM-based building energy performance simulations	Kim	Excluded
161	Self-Adaptive Genetic Algorithm for Modeling Energy Consumption in a Passive House	Stegaru	Excluded
162	Short-term smart learning electrical load prediction algorithm for home energy management s	El-Baz	Excluded
163	Simplified energy prediction method accounting for part-load performance of chiller	Kim	Excluded
164	Simulation, implementation and monitoring of heat pump load shifting using a predictive cont	Allison	Excluded
165	Solar energy harvesting wireless sensor network nodes: A survey	Sharma	Excluded
166	Statistical decision assistance for determining energy-efficient options in building design unde	Singh	Excluded
167	Statistical investigations of transfer learning-based methodology for short-term building ener	Fan	Excluded
168	Stochastic models for building energy prediction based on occupant behavior assessment	Virote	Excluded
169	Supervised based machine learning models for short, medium and long-term energy predictio	Ahmad	Excluded
170	System identification and model-predictive control of office buildings with integrated photov	Li	Excluded
171	Temporal and spatial predictability of occupants' presences in a library building	Ahn	Excluded
172	Testing the accuracy of synthetically-generated weather data for driving building energy simul	Degelman	Excluded
173	The early design model for prediction of energy and cost performance of building design optio	Yohanis	Excluded
174	The importance of integrally simulating the building, HVAC and control systems, and occupant	Kramer	Excluded
175	The uncertainty of manual shade control on west-facing facades and its influence on energy pe	Yao	Excluded
176	Towards a web-based energy consumption forecasting platform	Taborda	Excluded
177	Transfer Learning for Leisure Centre Energy Consumption Prediction	Banda	Excluded
178	Transfer learning with seasonal and trend adjustment for cross-building energy forecasting	Ribeiro	Excluded
179	Trees vs Neurons: Comparison between random forest and ANN for high-resolution predictior	Ahmad	Excluded
180	Tuning machine learning models for prediction of building energy loads	Seyedzadeh	Excluded
181	Uncertainty of building energy performance at spatio-temporal scales: A comparison of aggreg	Yao	Excluded
182	Unsupervised energy prediction in a Smart Grid context using reinforcement cross-building tra	Mocanu	Excluded
183	Urban energy flux: Spatiotemporal fluctuations of building energy consumption and human m	Mohammadi	Excluded
184	Urban energy use modeling methods and tools: A review and an outlook	Abbasabadi	Excluded
185	Use of dynamic occupant behavior models in the building design and code compliance process	Gilani	Excluded
186	Using deep learning approaches with variable selection process to predict the energy perform	Hwang	Excluded
187	Using Support Vector Machine to Predict Next Day Electricity Load of Public Buildings with Sub-	Fu	Excluded
188	Vector field-based support vector regression for building energy consumption prediction	Zhong	Excluded
189	WISE: web of object architecture on IoT environment for smart home and building energy man	Yu	Excluded

Appendix B. Systematic literature review: Included articles with tags



Ind	Title	Author	Tag
1	A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review	Ahmad	AI, ELEC, CONS
2	A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition	Taieb	AI
3	A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models	Wang	AI, CONS
4	A review of data-driven building energy consumption prediction studies	Amasyali	AI, CONS
5	A review on the prediction of building energy consumption	Zhao	AI, CONS
6	Assessment of deep recurrent neural network-based strategies for short-term building energy predictions	Fan	AI, LOAD
7	Comparative study of data driven methods in building electricity use prediction	Zeng	AI, ELEC, CONS
8	Improving forecasting accuracy of daily energy consumption of office building using time series analysis based on wavelet transform decomposition	Fang	STAT, ELEC, CONS
9	Machine learning for estimation of building energy consumption and performance: a review	Seyedzadeh	AI, ELEC
10	Methods for benchmarking building energy consumption against its past or intended performance: An overview	Li	AI, STAT, CONS
11	Modeling and forecasting building energy consumption: A review of data-driven techniques	Bourdeau	AI, STAT, HEAT, ELEC, CONS
12	New artificial neural network prediction method for electrical consumption forecasting based on building end-uses	Escrivá-Escrivá	AI, ELEC, CONS
13	Short-term load forecasting for non-residential buildings contrasting artificial occupancy attributes	Massana	AI, ELEC, LOAD
14	State of the art in building modelling and energy performances prediction: A review	Foucquier	AI, ELEC, HEAT, CONS
15	Towards efficient electricity forecasting in residential and commercial buildings: A novel hybrid CNN with a LSTM-AE based framework	Khan	AI, ELEC, CONS
16	Upgrade of an artificial neural network prediction method for electrical consumption forecasting using an hourly temperature curve model	Roldán-Blay	AI, ELEC, CONS

Appendix C. Model validation: interview questions

Technical

- **Recording:** Is it okay if I record the meeting?
- **Publishing:** Is it okay to publish your comments in the work with your name?

Warm-up

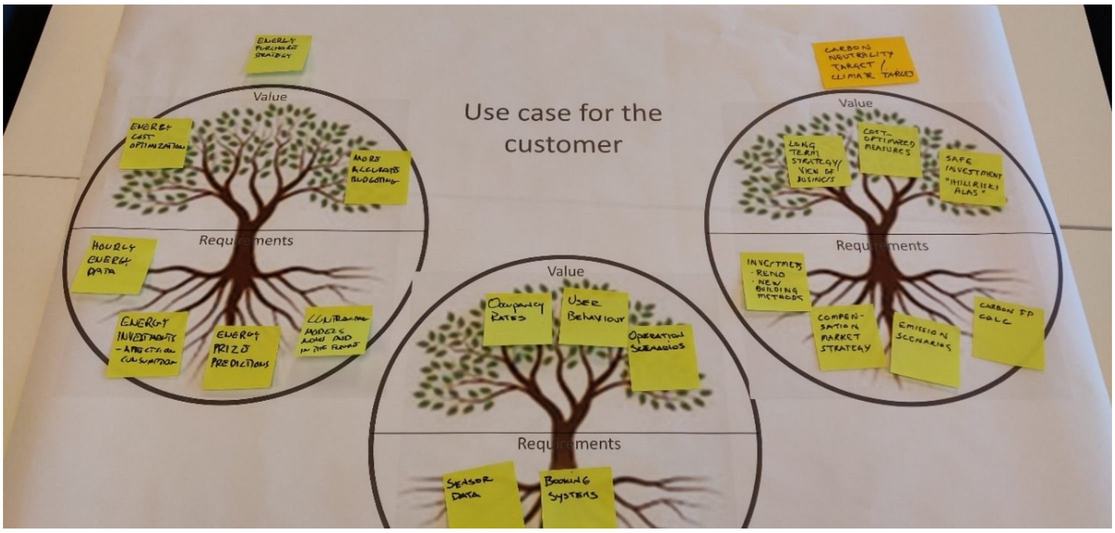
- **Basic information:** Name, occupation?
- **Experience:** In number of years, how would you describe your experience in:
    - Statistical modelling?
    - System architecture/data engineering?
    - Buildings-related?
- **Clarity:** Do you have any questions about the concept?
- **Sentiment:** Other initial thoughts about the concept?

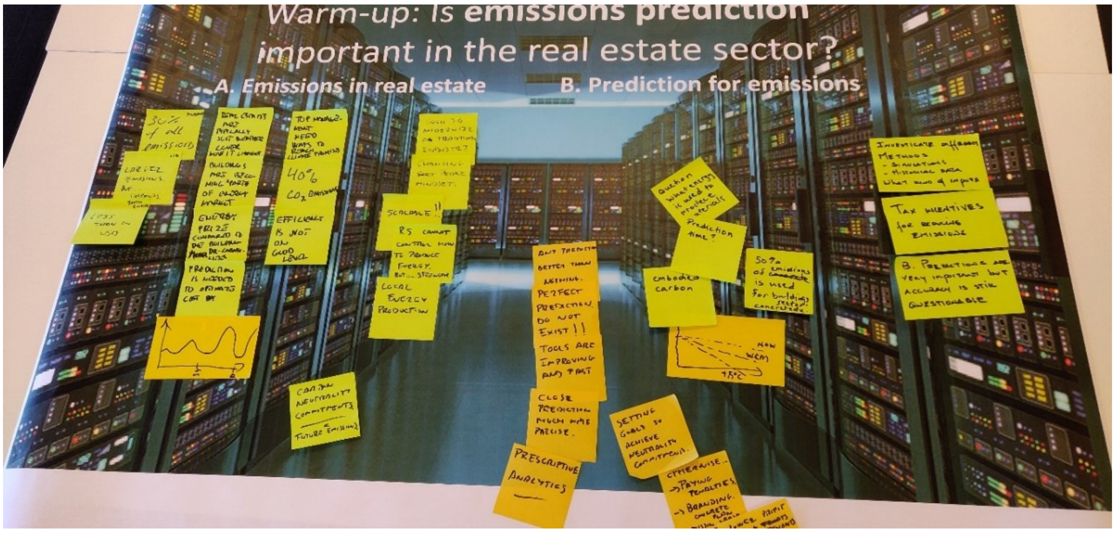
Interview questions

- **Model characteristics and architectural choices**
    - Validity: *Do you regard the choice of model to fit the task?*
    - Weaknesses: *What weaknesses/concerns can you identify regarding the choice of model?*
        - *How could these be combated?*
        - *Would you have chosen a different method?*
            - *Why?*
- **Model deployment & performance in the building setting**
    - Performance: *Do you think the model is successful in its environment given the performance evaluation done?*
    - Instructions of implementation: *What do you see as the main threats regarding the correct implementation of the model?*
    - Upkeep & monitoring: *Do you think the model would need fine tuning or monitoring?*
- **System/pipeline architecture**
    - *What are your thoughts on the overall system architecture?*
    - *What kind of an architecture would you advise?*
- **Importability of the model**
    - *Do you see any other potential application for a similar GBM model?*
    - *How easy do you think it would be to modify the model to another use case?*

Appendix D. Workshop documentation

Date: 28.02.2020


Use case for the customer
Value
Requirements
ENERGY PURCHASE STRATEGY
ENERGY COST OPTIMIZATION
MORE ACCURATE BUDGETING
HOURLY ENERGY DATA
ENERGY INVESTMENTS - AFFECTS ON CONSUMPTION
ENERGY PRIZE PREDICTIONS
CONTROLLING MODELS NOW AND IN THE FUTURE
Occupancy Rates
User Behaviour
Operation Scenarios
SENSOR DATA
BOOKING SYSTEMS
CARBON NEUTRALITY TARGET / CLIMATE TARGET
LONG TERM STRATEGY/ VIEW OF BUSINESS
COST-OPTIMIZED MEASURES
SAFE INVESTMENT "HIILIRISKI ALAS"
INVESTMENTS - RENO - NEW BUILDING METHODS
COMPENSATION MARKET STRATEGY
EMISSION SCENARIOS
CARBON FP CALC


Warm-up: Is emissions prediction important in the real estate sector?
A. Emissions in real estate
B. Prediction for emissions
30% of all emissions
LARGER EMISSIONS. BUT...
LESS THAN IN USA
BUILDINGS ARE BECOMING "PARTS" OF ENERGY MARKET
ENERGY PRIZE COMPARED TO THE BUILDING
PREDICTION IS NEEDED TO OPTIMIZE COST
TOP MANAGEMENT NEED WAYS TO REACH CLIMATE PROMISES
40% CO_2 EMISSIONS
EFFICIENCY IS NOT ON GOOD LEVEL
GOOD TO MODERNIZE THE TRADITIONAL INDUSTRY
CHANGING REAL ESTATE MINDSET
SCALABLE!!
RS CANNOT CONTROL HOW TO PRODUCE ENERGY. BUT... STRENGTH
LOCAL ENERGY PRODUCTION
CARBON NEUTRALITY COMMITMENTS
FUTURE EMISSIONS
ANY PREDICTION BETTER THAN NOTHING. PERFECT PREDICTION DO NOT EXIST!!
TOOLS ARE IMPROVING AND FAST
CLOSE PREDICTION MUCH MORE PRECISE.
PRESCRIPTIVE ANALYTICS
SETTING GOALS TO ACHIEVE NEUTRALITY COMMITMENTS.
OTHERWISE... → PAYING PENALTIES → BRANDING
LOWER PROFIT
Question what energy is used to produce materials
Prediction time?
embodied carbon
50% of emissions is used for buildings
NOW
WEM
1.5°C
INVESTIGATE different METHODS - SIMULATIONS - HISTORICAL DATA What kind of inputs
TAX INCENTIVES for REDUCING EMISSIONS
B. PREDICTIONS ARE VERY important but ACCURACY IS STILL QUESTIONABLE