FROM SMART HEAT METERS TO DIAGNOSTICS

DATA-DRIVEN METHODOLOGIES FOR BUILDING EFFICIENCY ASSESSMENT WITHIN DISTRICT HEATING

BY
DANIEL LEIRIA

AALBORG UNIVERSITY
DENMARK

Ph.D. Thesis 2024

Main Ph.D. supervisor:

Associate Professor Michal Zbigniew Pomianowski, Aalborg University

Co-supervisor(s):

Associate Professor Hicham Johra, Aalborg University

Associate Professor Anna Marszal-Pomianowska, Aalborg University

Curriculum vitae

Name: Daniel Leiria
Date of birth: 16.10.1996 (Faro, Portugal)
Email:

- AAU: dle@build.aau.dk
- Personal: daniel.h.leiria@gmail.com

LinkedIn: www.linkedin.com/in/danielhleiria/
Portfolio: www.datacamp.com/portfolio/danielhleiria
Research: www.scopus.com/authid/detail.uri?authorId=57211549827#

Professional Experience

2020 – 2024	Ph.D. fellow / Research assistant – Department of the Built Environment, Aalborg University (Denmark)
2017 – 2018	HVAC Project Design Engineer – Bioestratégia, Ltd. (Portugal)

Education

2018 – 2020	Master of Science: Indoor Environment and Energy Engineering – Department of the Built Environment, Aalborg University (Denmark)
2014 – 2017	Bachelor's degree: Mechanical Engineering, Thermal Energy Specialization – Universidade do Algarve (Portugal)

English Summary

The urgency to address environmental sustainability has intensified the focus on decarbonizing energy sources across various sectors, including district heating (DH) systems. As these systems are integral to urban infrastructure, providing heating solutions that are efficient and potentially environmentally friendly when configured and optimized correctly. However, although under transformation many existing DH networks still rely on carbon-intensive energy sources, which contributes significantly to large carbon emissions. Therefore, there is a pressing need to enhance the efficiency and operational effectiveness of these systems to replace their energy sources with renewable solutions. Innovations such as smart heat meters (SHM) offer potential breakthroughs in managing and optimizing DH systems, aiming to reduce their carbon footprint while improving overall grid performance.

This dissertation investigates the innovative use of SHM data in buildings connected to the DH grid while emphasizing advancements in energy efficiency and fault detection and diagnosis. It is structured across multiple key chapters, each responding to specific research questions outlined in Chapter 1.

Contributions:

The main contributions of this Ph.D. work are:

- Validated the effectiveness of SHM in assessing buildings within the DH grid, demonstrating that incorporating weather data can help operators better predict and manage their energy usage. Additionally, using information from EPC helps identify buildings needing efficiency improvements, while visualization tools assist in interpreting data to make informed decisions.

- Developed a methodology using machine learning models based on SHM data to accurately disaggregate and estimate energy demands for space heating and domestic hot water. This work also addressed the impacts of data rounding in utility data collection, proposing methods to mitigate these issues.

- Examined the influence of occupant behavior on energy usage, noting that SHM data alone are insufficient for capturing all behavioral

nuances. Incorporating detailed indoor condition data and occupant interviews provides a more comprehensive understanding, leading to improved energy management strategies.

- Investigated the efficacy of SHM in fault detection and diagnosis within DH systems. Demonstrated that continuous monitoring allows for the application of advanced analytical methods for automated fault diagnosis, hence improving operational efficiency and intervention costs.

Overall, this dissertation provides a foundational framework for utilizing SHM data in buildings with other sources of data to enhance the efficiency and sustainability of DH systems, highlighting the significant potential for future advancements through continued research and technology integration.

Keywords: District heating networks; Smart heat meters; Building energy efficiency; Fault detection and diagnosis; Machine learning; Heating systems performance.

Dansk Resumé

Det haster med at adressere miljømæssig bæredygtighedm hvilket har intensiveret fokuset på dekarbonisering af energikilder på tværs af forskellige sektorer, herunder fjernvarmesystemer (DH). Disse systemer er en integreret del af byinfrastrukturen og hvis de er optimeret korrekt, giver de varmeløsninger, der er effektive og potentielt miljøvenlige. Imidlertid er mange eksisterende fjernvarme netværk afhængige af fossile brændstoffer, hvilket bidrager væsentligt til høje udledninger af CO_2. Det skaber derfor et presserende behov for at erstatte deres energikilder med vedvarende løsninger, og derved øge effektiviteten samtidig med at CO_2 udledningen mindskes. Innovationer såsom smarte varmemålere (SHM) tilbyder potentielle gennembrud inden for overvågning, styring og optimering af fjernvarme systemer, med det formål at reducere deres CO_2 fodaftryk og samtidig forbedre den overordnede netydelse.

Denne afhandling undersøger den innovative brug af SHM-data i DH-systemer, mens den lægger vægt på fremskridt inden for øget energieffektivitet og fejldetektion og diagnose. Den er struktureret med flere kapitler, der hver svarer på sit specifikke forskningsspørgsmål, der er beskrevet i det første kapitel.

Bidrag:

Hovedbidragene fra dette ph.d. arbejde er:

- Valideret effektiviteten af smarte varmemålere i analyseringen af bygninger på fjernvarmenettet, og demonstreret at yderligere integration af vejrdata kan hjælpe operatører med bedre at forudsige og styre energiforsyningen. Derudover hjælper brugen af EPC oplysninger med at identificere bygninger, der har behov for effektivitetsforbedringer, og visualiseringsværktøjer hjælper med at fortolke data for at kunne træffe informerede beslutninger.

- Udviklede en metodologi ved hjælp af maskinlæringsmodeller baseret på data fra smarte varmemålere til nøjagtigt at adskille og estimere energibehovet til rumopvarmning og varmt brugsvand. Dette arbejde omhandlede også virkningerne af dataafrunding i dataindsamlingen af

forsyningsvirksomheder og foreslog metoder til at afhjælpe disse problemer.

- Undersøgt indflydelsen af beboernes adfærd på energiforbruget og observeret at data fra smarte varmemålere alene er utilstrækkelige til at indentificere alle adfærdsnuancer. Inkorporering af detaljerede indekvalitetsdata og beboerinterviews giver en mere omfattende forståelse, hvilket fører til forbedrede energistyringsstrategier.

- Undersøgt effektiviteten af smarte varmemålere i fejldetektion og diagnosticering inden for fjernvarmesystemer og demonstreret, at kontinuerlig overvågning giver mulighed for anvendelse af avancerede analysemetoder til automatiseret fejldiagnose, hvilket forbedrer driftseffektiviteten og omkostningerne ved løsning af problemerne.

Samlet set giver denne afhandling en grundlæggende rammeværktøj for at bruge data fra smarte varmemålere sammen med andre datakilder for at øge effektiviteten og bæredygtigheden af fjernvarmesystemer, hvilket viser det betydelige potentiale for fremtidige fremskridt gennem fortsat forskning og implementering af teknologien.

Nøgleord: Fjernvarmenet; Smart varmemålere; Energieffektivitet i bygninger; Fejlsøgning og diagnose; Maskinelæring; Varmesystemers ydeevne.

Acknowledgments

Completing this Ph.D. has been an immensely challenging and rewarding journey, one that I could not have embarked upon without any of you.

First and foremost, I express my deepest gratitude to all my supervisors (Michal Pomianowski, Hicham Johra, and Anna Marszal-Pomianowska), for their invaluable mentorship, patience, and unwavering support throughout this research. Your insights and direction were crucial in shaping both the research and my development as a Ph.D. And your dedication and patience are unparalleled, from the "quick drop-visit" at your office for doubts to late email answers – you have been another class of supervision. I also want to thank all my colleagues I had the pleasure to have in our division – the pieces of advice, exams, and celebrations (breakfasts on Friday!) have been great.

I am equally thankful to all our project partners in E-DYCE and PRELUDE who provided essential resources, insights, and collaboration opportunities that enriched this research.

My Ph.D. journey was also enriched by the camaraderie and support of my fellow Ph.D. colleagues (and a research assistant) – the ones at the *Corner Office*. Sharing this path with you has been one of the most rewarding aspects of my experience. Thank you for the stimulating discussions, for the late nights we spent in conferences, and for all the moments in between. A special thanks goes to my colleagues in Turin (BAEDA LAB). Your friendship, support, and supervision have made my time in Italy both productive and enjoyable.

Aos meus pais e familiares, pelo amor, apoio e por acreditarem nas minhas capacidades incondicionalmente. Estarei eternamente grato por todos os sacrifícios que vocês fizeram para garantir que eu pudesse chegar até aqui. E posso dizer, sem sombra de dúvida, que não seria o que sou hoje sem vocês.

Lastly, I must express my profound gratitude to my Giulia for her love and undeniable support. Today, we laugh once again, *Chiquitita*.

Tak, Thank you, Grazie, Obrigado.

"(...) Três vezes do leme as mãos ergueu,

Três vezes ao leme as reprendeu,

E disse no fim de tremer três vezes:

«Aqui ao leme sou mais do que eu:

Sou um Povo que quer o mar que é teu;

E mais que o mostrengo, que me a alma teme

E roda nas trevas do fim do mundo;

Manda a vontade, que me ata ao leme,

De El-Rei D. João Segundo!"

Fernando Pessoa, in Mensagem (1934)

Preface

The work presented in this thesis is part of a Ph.D. project funded by the European Union's Horizon 2020 research and innovation program under grant agreement No 893945 (E-DYCE). Also, this project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 958345 (PRELUDE). Daniel Leiria has carried out this work from January 2022 to May 2024.

This thesis is a composition of papers appended from 1 to 7, and they have been integrated directly into the main body of text to provide the reader with a smoother, uninterrupted reading experience. Unless otherwise stated, all illustrations are the author's own work.

Overview of papers

This paper-based thesis consists of the following collection of papers:

Paper 1: *"Using data from smart energy meters to gain knowledge about households connected to the district heating network: A Danish case"*
Daniel Leiria, Hicham Johra, Anna Marszal-Pomianowska, Michal Zbigniew Pomianowski, Per Kvols Heiselberg
Smart Energy 2021

Paper 2: *"A methodology to estimate space heating and domestic hot water energy demand profile in residential buildings from low-resolution heat meter data"*
Daniel Leiria, Hicham Johra, Anna Marszal-Pomianowska, Michal Zbigniew Pomianowski
Energy 2022

Paper 3: *"Validation of a new method to estimate energy use for space heating and hot water production from low-resolution heat meter data"*
Daniel Leiria, Hicham Johra, Evangelos Belias, Davide Quaggiotto, Angelo Zarrella, Anna Marszal-Pomianowska, Michal Zbigniew Pomianowski
BuildSim Nordic Conference 2022, Copenhagen, Denmark

Paper 4: *"Estimating residential space heating and domestic hot water from truncated smart heat data"*

Daniel Leiria, Markus Schaffer, Hicham Johra, Anna Marzsal-Pomianowska, Michal Zbigniew Pomianowski
CISBAT 2023, Lausanne, Switzerland

Paper 5: *"From showers to heaters: Evaluating the different factors that play a role in buildings' energy signature"*
Daniel Leiria, Hicham Johra, Yue Hu, Olena Kalyanova Larsen, Anna Marszal-Pomianowska, Martin Frandsen, Michal Zbigniew Pomianowski
Submitted in Energy and Buildings (2024)

Paper 6: *"Towards automated fault detection and diagnosis in district heating customers: generation and analysis of a labeled dataset with ground truth"*
Daniel Leiria, Kamilla Heimar Andersen, Simon Pommerencke Melgaard, Hicham Johra, Anna Marszal-Pomianowska, Marco Savino Piscitelli, Alfonso Capozzoli, Michal Zbigniew Pomianowski
IBPSA Building Simulation 2023, Shanghai, China

Paper 7: *"Is it returning too hot? Time series segmentation and feature clustering of end-user substation faults in district heating systems"*
Daniel Leiria, Hicham Johra, Justus Anoruo, Imants Praulins, Marco Savino Piscitelli, Alfonso Capozzoli, Anna Marszal-Pomianowska, Michal Zbigniew Pomianowski
Submitted in Applied Energy (2024)

Other peer-reviewed research papers

In addition to papers 1 to 7, the author of this dissertation has been involved in other peer-reviewed publications during the Ph.D. studies. These papers do not belong to the dissertation but are included in this list to show that the author has been involved in additional contributions to the field.

Paper 8: *"Increasing the accuracy of low-resolution commercial smart heat meter data and analysing its error"*
Markus Schaffer, **Daniel Leiria**, J. Eduardo Vera-Valdés, Anna Marszal-Pomianowska
EC3 Conference 2023, Crete, Greece

Paper 9: *"Who Produces the Peaks? Household Variation in Peak Energy Demand for Space Heating and Domestic Hot Water"*

Anders Rhiger Hansen, **Daniel Leiria**, Hicham Johra, Anna Marszal-Pomianowska
Energies 2022

Paper 10: *"Detailed operational building data for six office rooms in Denmark: Occupancy, indoor environment, heating, ventilation, lighting and room control monitoring with sub-hourly temporal resolution"*
Simon Pommerencke Melgaard, Hicham Johra, Victor Ørsoe Nyborg, Anna Marszal-Pomianowska, Rasmus Lund Jensen, Christos Kantas, Olena Kalyanova Larsen, Yue Hu, Kirstine Meyer Frandsen, Tine Steen Larsen, Kjeld Svidt, Kamilla Heimar Andersen, **Daniel Leiria**, Markus Schaffer, Martin Frandsen, Martin Veit, Lene Faber Ussing, Søren Munch Lindhard, Michal Zbigniew Pomianowski, Lasse Rohde, Anders R. Hansen & Per Kvols Heiselberg.
Data in Brief 2024

Paper 11: *"A Mixed-Method Approach to Understand Energy-Related Occupant Behavior and Everyday Practices in Multi-Story Residential Buildings"*
Kamilla H. Andersen, Anders R. Hansen, Anna Marszal-Pomianowska, Henrik N. Knudsen, **Daniel Leiria**, Per K. Heiselberg. Under review in the Energy & Buildings 2024.

Other scientific dissemination

In addition to the peer-reviewed articles, the author of this dissertation has also been involved in other activities during the Ph.D. studies. These publications do not belong to the dissertation but are included to show additional contributions to the field.

a) *"Digitalisation as a potential game changer to foster energy efficiency in the building stock"*
Markus Schaffer, Judith Fauth, Hicham Johra, Olena Kalyanova Larsen, **Daniel Leiria**, Anna Marszal-Pomianowska, Michal Zbigniew Pomianowski
REHVA Journal 2024

b) *"Working document: Summary of Existing FDD Frameworks for Building Systems"* –
Kamilla Heimar Andersen, Simon Pommerencke Melgaard, **Daniel Leiria**

DCE Technical Report No. 312 (2023)

c) *"Session summaries from: Study circle on The Use of Data in The Built Environment – Challenges and opportunities"*
Kamilla Heimar Andersen, **Daniel Leiria**, Simon Pommerencke Melgaard, Markus Schaffer, Anna Marszal-Pomianowska
DCE Latest News No. 60 (2023)

d) Several E-DYCE project reports

e) Several PRELUDE project reports

Table of contents

List of figures

Glossary

DH	*District heating* A system that distributes heat produced at a central location through a network of insulated pipes. This heat is used for residential and commercial purposes, including space heating and domestic hot water.
DHW	*Domestic hot water* Hot water supplied for domestic purposes, such as bathing/showering, cooking, and cleaning.
EPBD	*Energy Performance of Buildings Directive* A European Union directive aimed at improving the energy efficiency of buildings within the EU Member States.
EPC	*Energy performance certificates* Documents that provide information on the energy efficiency of a building (energy labels), including recommendations for improving energy efficiency and reducing energy costs.
EU	*European Union*
FDD	*Fault detection and diagnosis* A process or system for identifying and diagnosing faults in equipment, particularly in heating, ventilation, and air conditioning systems.
SH	*Space heating* The heating of indoor spaces to ensure comfort for occupants, typically achieved through systems like radiators, underfloor heating, etc.
SHM	*Smart heat meter(s)* Measurement devices that record the usage of heat in a more accurate manner for billing purposes, often providing real-time data to users and utility companies.

Chapter 1. Introduction

1.1 Research introduction

1.1.1 Current state of heating provision in Europe

Since the beginning of our species, humanity realized that to survive the elements, it needed to focus on having air, shelter, water, and food[1]. So, if shelter is as important as air and water, we as humanity should give the same type of concern as we give to the latter. Nevertheless, everyone can agree, that shelter in today's society has a different meaning, as it is the offices we work in, the hospitals we visit, the classrooms where we learn, and the homes we live in. All these buildings, regardless of their type and specific purpose, must be carefully designed and built to provide for us what we will always need – shelter.

Every building when designed, commissioned, or in operation, requires a lot of decisions and effort from different stakeholders. And these decisions affect each other, due to the fact that a building is a holistic system, where the interaction between the outdoor environment, building properties, systems, and occupants is always present [1]. And concerning this Ph.D. project, where we delve into the building's heating systems – they require the same level of focus regarding their interactions as they play a vital role in providing indoor comfort and domestic hot water for its occupants.

Taking a step back towards the bigger picture, due to this holistic nature, any decision and effort in a building project will affect the next decision and effort. Therefore, one can argue that no decision regarding a building is an *island*, e.g., what we decide regarding the construction materials affects the heating systems needed. And by taking this reasoning a step even further, the decisions we make today about our buildings will also affect them throughout their long lifespan.

[1] This is also known as the survival rule of thumb called the "rule of threes".

However, our societies have not just one building alone, instead, they have several of them, where cities can easily reach thousands and sometimes millions of them. But how do our buildings impact our society, and specifically to this Ph.D. dissertation, do they have a large effect on our energy usage? In the European Union (EU) alone, according to 2020 data, it is estimated to have almost 112 million buildings [2], which are responsible for 40% of Europe’s total energy usage [3]. Of the EU buildings, 91% are residential [2]. In these residential buildings, around 80% of the energy is used for space heating (SH) and domestic hot water (DHW) [4]. By looking at these numbers, one can conclude that the EU's building sector greatly contributes to its overall energy usage. Additionally, for the EU to achieve its energy and sustainability targets due to its growing environmental awareness, it must focus in larger detail on its building stock and specifically on the heating systems.

1.1.2 Impact of major EPBD regulations in Europe

This growing awareness of climate change has triggered a shift towards sustainable and efficient heating solutions. Traditional heating methods, primarily reliant on fossil fuels, contribute significantly to greenhouse gas emissions. As a result, there is a pressing need to transition to low-carbon and renewable heating technologies [5]. European policies, such as the EU's Green Deal [6] and the Renewable Energy Directive [7], are driving the adoption of renewable energy sources like biomass, solar thermal, and geothermal heat. These initiatives aim to reduce the carbon footprint of heating systems and enhance their energy efficiency.

Another important driver as an energy policy is the legislation within the EU – Energy Performance of Buildings Directive (EPBD) [8,9], aimed at improving the energy efficiency of buildings. Introduced in 2002 and subsequently revised in 2010 and 2018, the EPBD is part of the EU’s broader strategy to fight climate change and enhance energy security. The directive sets stringent requirements for Member States to implement measures that reduce energy usage in both new and existing buildings. Its ultimate goal is to make the EU's building stock highly energy-efficient and decarbonized by 2050 [3].

The EPBD has been fundamental in promoting efficient heating systems and integrating renewable energy sources, significantly enhancing energy efficiency in buildings across the EU. It encourages the adoption of highly efficient heating and cooling systems, such as district heating networks, combined heat and power systems, and advanced heat pump technologies. By setting minimum energy performance requirements, the directive has led to a widespread implementation of these technologies, thus significantly reducing energy usage and greenhouse gas emissions. Additionally, the EPBD supports the integration of renewable energy into buildings, including solar thermal, geothermal, and biomass systems. It mandates that the energy performance of buildings accounts for the energy produced from on-site renewable sources, incentivizing the adoption of these technologies. This integration not only increases investment in renewable energy technologies but also promotes a broader acceptance of decentralized energy production models. Moreover, the directive requires Member States to establish long-term renovation strategies aimed at decarbonizing their building stocks. These strategies involve enhancing the energy performance of existing buildings through deep renovations, often incorporating upgrades to heating systems and the integration of renewable energy. This framework helps mobilize investments in building renovations, which leads to an increase in energy efficiency throughout the long lifespan of the buildings.

The vision of the EPBD extends as well to developing smart solutions that leverage the full potential of data collection and usage. This approach includes deploying intelligent building systems that optimize energy use in real-time, enhancing both operational efficiency and occupant comfort. By integrating smart technologies with traditional energy systems, the directive paves the way for a future where buildings are not only energy efficient but also intelligently responsive to the changing needs of their occupants and the environment.

1.1.3 District heating: One of the EPBD solutions

District heating (DH) systems are one of the solutions present in the EPBD for urban heating, characterized by the centralized generation of heat which is then distributed to multiple buildings within a locality via a network of insulated pipes [10]. These systems can utilize a variety of energy sources, including fossil fuels, biomass, geothermal, and waste heat from industrial processes. The centralized heat generation occurs at a district heating plant, and the produced heat is transported as hot water or steam to residential, commercial, and industrial consumers. This method contrasts with decentralized heating systems where each building or unit has its own local heating source.

Concerning this heating solution, it can achieve high energy efficiencies by utilizing combined heat and power (CHP) plants. As these plants simultaneously can generate electricity and useful heat from the same energy source, maximizing the energy extracted from its source and reducing wastage. This dual-generation process typically achieves efficiencies of 80-90%, significantly higher than conventional power plants which only generate electricity with efficiencies around 30-40% [11]. By leveraging renewable energy sources and waste heat, district heating systems contribute to lower carbon emissions. For instance, the integration of biomass and geothermal energy, along with the utilization of waste heat from industrial operations, significantly reduces the reliance on fossil fuels [12]. Furthermore, DH systems benefit from economies of scale due to their large-scale operation. Centralized production and distribution of heat allow for cost savings in fuel purchase, plant operation, and maintenance [13].

One of the significant challenges of DH systems is the transmission heat loss through the pipe networks. Even with well-insulated pipes, some heat is inevitably lost as the hot water or steam travels from the central plant to the end-users. This heat loss can reduce the overall efficiency of the system, particularly over long distances. Furthermore, the establishment of a DH system requires substantial upfront investment. Costs associated with constructing the central heating plant, laying down the extensive network of insulated pipes, and connecting individual buildings to the network can be prohibitively high. These initial expenses often necessitate significant public or private financing, making the widespread adoption of DH challenging, especially in areas without existing infrastructure. DH systems are characterized by their substantial and often rigid infrastructure. Once installed, the network of pipes and central plant facilities are not easily modified or relocated. This inflexibility can pose challenges in rapidly evolving urban environments where new buildings are constructed, and old ones are demolished or repurposed. Additionally, as energy landscapes change and new technologies emerge, adapting existing DH systems to incorporate these innovations can be complex and costly.

1.1.4 The role of Denmark in pioneering district heating

Denmark has established itself as a global leader in the implementation and innovation of DH systems. Since the 1970s, following the oil crises, Denmark has strategically invested in DH as a means to enhance energy security and lower greenhouse gas emissions [14]. The Danish government has implemented comprehensive policies and regulations to support the development and expansion of DH networks, making Denmark a model for other countries seeking to develop sustainable heating solutions.

Today, approximately 64% of Danish households are connected to the DH grid, making it one of the highest penetration rates globally [15]. This widespread adoption is largely due to strong governmental support, which includes financial incentives, legislative measures, and public awareness campaigns [16].

The Danish DH sector is also undergoing a significant transformation driven by the rise of digital technologies and analytics and the EPBD motion towards digitalization. This revolution is characterized by the integration of advanced data collection, processing, and analysis techniques into DH systems, enabling more efficient and effective management of heat distribution. Digital technologies such as big data analytics and machine learning are being increasingly adopted to optimize DH operations, improve energy efficiency, and reduce carbon emissions [17]. A cornerstone of this data revolution in the DH sector is the deployment in large scale of smart heat meters (SHM). These devices are installed at various points within the DH network, including at customer premises, substations, and distribution lines. The SHM collects real-time data on a wide range of parameters such as temperature, flow rates, and energy usage enabling utilities to monitor the performance of the DH system with unprecedented granularity and accuracy [18]. From the SHM resulting data, it is envisioned to provide valuable insights into customer heat usage patterns, allowing for more precise billing and improved customer service. Additionally, these sensors deployed across the network can help in detecting anomalies and inefficiencies, such as heat losses, leakages, or equipment malfunctions, facilitating prompt corrective actions. Nevertheless, a question arises – is it really possible to use the data devices to have a deeper under-the-hood understanding of the buildings connected to the DH grid? Several peer-reviewed publications and Ph.D. dissertations were written on this topic, nonetheless, more work is necessary if we want to attain the full application of data in the Danish DH networks.

1.2 Research outline

Facing this data revolution in the DH sector and acknowledging that this energy-providing system has a large potential to improve the Danish (and European) heating sector, this dissertation was founded. Explicitly this Ph.D. thesis attempts to address the following research questions (RQ):

RQ1. How can weather data, energy performance certificates, and visualization tools be used to aid in decision-making for energy management within district heating?

RQ2. Can machine learning models based on smart heat meter data accurately predict the energy demand for space heating and domestic hot water?

RQ3. Can smart heat meter data, supplemented with information about indoor environmental conditions, enhance our understanding of the energy use and heating practices of building occupants?

RQ4. How effective are smart heat meters at identifying specific types of faults within the district heating system customers?

1.3 Thesis outline

This dissertation explores the use of SHM data to enhance the understanding of the performance of buildings connected to the DH grid. Throughout this work, it was developed and applied various methods to integrate SHM data with additional sources of information, such as indoor sensors and energy label reports. Thus, the thesis is structured as follows:

Chapter 1: Introduction

This chapter introduces the thesis, outlines the research objectives, and sets the context for the subsequent investigation.

Chapter 2: Assessing smart heat meters applicability

This chapter evaluates the usefulness of SHMs in analyzing buildings connected to the DH grid in order to answer RQ1. It involves an initial assessment, integrating energy label reports, and the development of a tool to synthesize collected data. This work had the outcome of **Paper 1** [19].

Chapter 3: Hourly heating share estimation using smart heat meters

This chapter details the application of SHMs to estimate hourly heating distributions for space heating and domestic hot water production in order to answer RQ2. It addresses limitations in existing methodologies and adapts the new approach to diverse datasets, including buildings outside of Denmark and modified SHM data. From this chapter, it was published in **Papers 2, 3,** and **4** [20-22].

Chapter 4: Integrating smart heat meters and indoor sensors data

By investigating the integration of SHM data with indoor sensor readings, this chapter delves into occupant behaviors and their interactions with the buildings and their heating systems. The analysis aims to determine the impact of these interactions on energy performance, thus answering RQ3. This investigation led to the submission of **Paper 5** [23].

Chapter 5: Fault detection and diagnosis with district heating customers

This chapter introduces several analytic methodologies developed for fault detection and diagnosis in buildings connected to the DH grid. It integrates SHM data with technical reports on identified faults from inspections of targeted buildings to create a reliable ground truth dataset for modeling fault diagnosis algorithms. This investigation resulted in the publication and submission of **Papers 6 and 7** [24,25] to answer RQ4.

Chapter 6: Conclusions

The final chapter summarizes the research findings, highlighting key insights and implications. It also discusses potential areas for further investigation, as suggested by the author, to advance the field and address remaining challenges.

The overall structure of this dissertation can be seen in Figure 1:

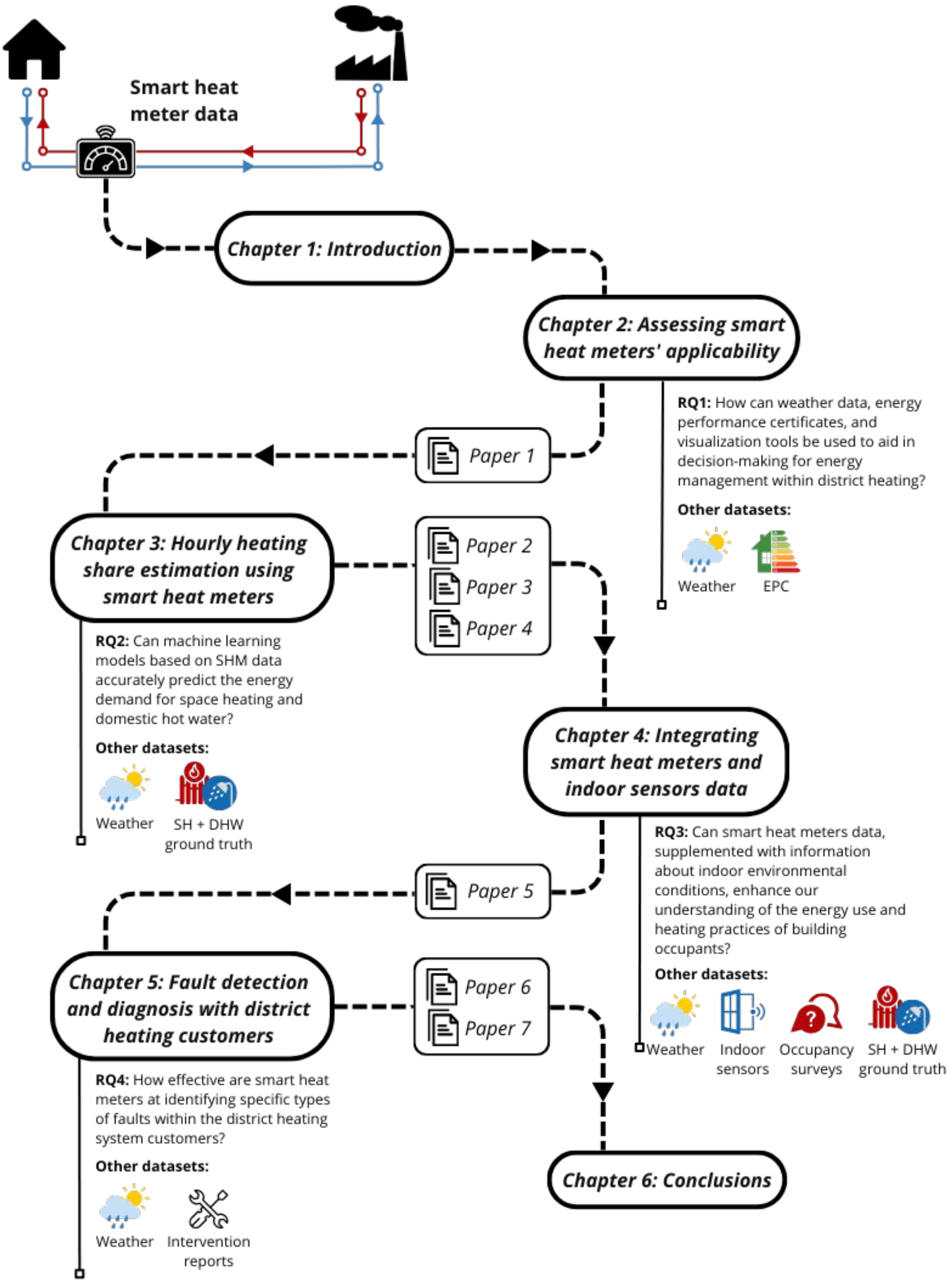


Figure 1: Dissertation structure.

Chapter 2. Assessing smart heat meters' applicability

This chapter presents the smart heat meters (SHM) dataset, detailing the variables it contains. An algorithm is proposed to preprocess the raw dataset from the metering devices for further analysis. Additionally, a simple methodology was developed to extract the thermal characteristics of the buildings where the measurements were taken, and these were compared with information from the energy performance certificates (EPC). A visualization tool developed in Shiny [26] is also briefly introduced, designed to enhance the understanding of the dataset for various stakeholders, such as building owners and energy utility managers. Finally, this chapter discusses how the outcomes of this research have inspired further studies on the topic.

2.1 The role of smart heat meters in the district heating

The widespread installation of SHM in Denmark exemplifies the significant impact these devices can have on understanding and improving DH network performance. By doing so, the billing process becomes much more straightforward but also enables a more in-depth analysis of the building stock connected to the grid [15]. Several articles point to the obtained advantages from access to SHM data, like the following:

- Improved energy efficiency: Smart meters provide real-time data on energy usage, enabling better monitoring and management of the heat supply. This allows for optimizing the overall efficiency of the heating distribution by adjusting the supply based on actual demand rather than estimations [27].

- Enhanced operational performance: With continuous monitoring, operators can quickly identify and address issues such as leaks or major inefficiencies in the heat distribution network. Thus, reducing downtime and extending the lifespan of the infrastructure [28].

- Cost savings for consumers: By providing detailed information on heating usage, SHM allows consumers to better understand their usage patterns and make informed decisions to reduce waste and lower their heating bills [29].

- Data-driven decision-making: The aggregation of data from multiple buildings provides valuable insights that can influence future planning and investments. For instance, trends in energy usage can guide the development of new energy-saving solutions or infrastructure renovations both on the distribution and in the end-users levels (buildings) [30].

- Environmental impact: Optimizing heat distribution and reducing energy waste directly contribute to lowering the carbon footprint of district heating systems [31].

- Transparency: Regular and accurate billing based on actual heat measurements, rather than estimates, improves transparency and trust between consumers and utility providers [32].

To reach all the advantages of SHM listed above, a larger understanding of the data is mandatory. Therefore, in the following section, it is presented an initial framework for the application of DH data to understand the buildings connected to the network. This framework employs a data curation process, energy analysis based on the weather data and buildings' EPCs, and a simple visualization tool for grid assessment.

2.2 Framework of SHM data for energy performance assessment

Paper 1

"Using data from smart energy meters to gain knowledge about households connected to the district heating network: A Danish case"

Daniel Leiria, Hicham Johra, Anna Marszal-Pomianowska, Michal Zbigniew Pomianowski, Per Kvols Heiselberg, Smart Energy, 2021, https://doi.org/10.1016/j.segy.2021.100035

Smart Energy 3 (2021) 100035

Contents lists available at ScienceDirect

Smart Energy

journal homepage: www.journals.elsevier.com/smart-energy

Using data from smart energy meters to gain knowledge about households connected to the district heating network: A Danish case

Daniel Leiria*, Hicham Johra, Anna Marszal-Pomianowska, Michal Zbigniew Pomianowski, Per Kvols Heiselberg

Aalborg University, Department of the Built Environment, Thomas Manns Vej 23, 9220, Aalborg Øst, Denmark

ARTICLE INFO

Article history:
Received 12 December 2020
Received in revised form
21 May 2021
Accepted 9 June 2021
Available online 12 June 2021

Keywords:
District heating
Smart energy meters
Big data mining
Linear regression analysis
Energy performance certificates
Building characterization

ABSTRACT

In Europe, one of the most sustainable solutions to supply heat to buildings is district heating. It has good acceptance in the Northern countries, a low-carbon footprint, and can easily integrate intermittent renewable energy sources when coupled to the electrical grid. Even though district heating is seen as a vital element for a sustainable future, it requires extensive planning and long-term investments. To increase the understanding of the district heating network performance and the demand-side dynamics of the connected buildings, several countries, including Denmark, have installed smart heat meters in different cities. In that context, this paper presents several methodologies to analyze the datasets from the smart heat meters installed in a small Danish town. The first method is concerning data curation to remove the anomalies and missing data points. The second method analyses measured variables (heat consumption, outdoor temperature, wind speed, and global radiation) to acquire new knowledge on the building characteristics. These results were compared with the values given by the energy performance certificates of a smaller sample of 41 households. Finally, to communicate and visualize the analysis outputs in a user-friendly way, an interactive web interface tool has been created.

1. Introduction

Denmark has very ambitious goals in terms of the sustainability of its energy sector. In 2030, the share of renewable energy sources (RES) shall be 55% in final energy usage, 100% in electricity use and 90% in the district heating (DH) sector [1]. The transition to a low-carbon future is built on two pillars: lowering energy demand and increasing intermittent RES usage, such as wind and solar power.

Buildings account for roughly 40% of all energy use in the European Union [2] and have, therefore, a prominent role in this transition towards sustainability. In addition to a general decrease of their energy demand, buildings should be operated in a smart way, i.e., modulate their energy demand according to the availability of local RES.

District Heating systems also play a vital role in shifting to a low carbon society as they provide most of the energy used for space heating and domestic hot water in Danish buildings (>60%). In other countries like Iceland, Poland, Lithuania, Estonia, Sweden, Finland, and Northern China, more than half of their building stock is connected to the DH grid [3]. In addition, DH systems can integrate a wide range of RES [4] and excess heat from local industrial processes. However, this requires a change in the DH network's operation by lowering supply and return temperatures and intelligent control at the building/customer level with heat demand-side management.

The digitalization of energy use in buildings brings new opportunities. However, smart technologies in residential buildings are expected to be adopted by 27% of EU households by 2025 [5], leaving 73% of building stock as conventional buildings equipped solely with energy meters lacking IoT (Internet of Things) enabled sensors and devices. Therefore, the lack of available technological infrastructure is a significant barrier that will prevent a substantial EU building stock share from being involved in the proactive and integral part of the evolving energy system. Moreover, current methods for assessing and optimizing building performance and evaluating smart readiness are based on simplified calculations [6,7]. Therefore, the difference between predicted and actual energy use of a building (the performance gap) can be up to a factor of 2.5 [8,9].

Currently, the real-time consumption data from DH smart heat

* Corresponding author.
E-mail address: dle@build.aau.dk (D. Leiria).

https://doi.org/10.1016/j.segy.2021.100035

meters is primarily used for billing the customers. In Denmark, it will be obligatory to collect dynamic heating data by using smart meters for every building connected to the DH grid from 2027 [10]. Therefore, this paper aims to present how to exploit these energy meters' large potential and how they can deliver new information on building connected to the DH grid: faulty operation of the ventilation/infiltration system, poor performance of the building envelope, etc. This information is crucial for CO_2-saving actions in energy-optimized buildings that smartly interact with the local DH system and facilitate its green transition.

These data will also provide a better knowledge of the actual building's energy use, which enables an accurate estimation and planning of the demand-side management for heating grids. By increasing the accuracy, it pushes forward the DH systems into the 4th Generation District Heating (4GDH). The 4GDH systems are the new evolutionary step of the DH networks, where the fluid-supply temperature is lowered (50–60 °C) to decrease the overall heat losses, reduce the distribution pipe diameters, lower the fluid flow rates and increase, consequently, the system's efficiency [11]. Even though the 4GDH systems bring several advantages, the system's implementation must be precise to assure the user's comfort. For this reason, it is argued that smart energy data will benefit the transition to the 4GDH systems because all data collected and its analysis will provide a better understanding of the grid, allowing a more reliable design and implementation of the 4GDH networks.

This paper's contributions are a methodology used to perform the cleansing and preparation of the DH dataset to be used for further analysis. Moreover, the treated dataset can be applied to more precisely assess the building's transmission losses, the ventilation and infiltration dependency on the energy consumption, and the impact of the solar radiation has on the building to decrease the heating demand. This study also proposes incorporating the methodology results to compare with the Energy Performance Certificates (EPC) calculations to evaluate their similarity and detect the key variables that might contribute to the energy performance gap.

After reviewing the current publications about data analysis of smart meters in DH systems in section 2, the current study's methodology is detailed in section 3. Section 4 presents and discusses the results from the application of the methods developed in this study. The article closes with conclusions and suggestions for future work.

2. State of the art

In a recent follow-up study of heat load profiles using smart heat data [12], Calikus et al. [13] presented an automated method to analyze heat load profiles for non-residential buildings. Since the latter have clear occupation schedules and heating management systems, their heat demand profiles are easier to predict compared to residential houses, in which occupants' behavior and heating practices are more diverse [14]. With the application of smart heat meters in the residential sector, the recent research work has focused on identifying typical customer segments according to heat consumption [15–17]. Johra et al. [18] have shown that clustering of buildings according to metered parameters, i.e., the temperature of return water to DH network (T_{return}) and the temperature difference (ΔT) between the supply and return fluid, can help to identify buildings with efficient heating systems.

Gianniou et al. [19] presented a simple methodology with uniform and steady-state assumptions about the occupants in all houses, i.e., heating practices, occupancy schedule, and single-zone modeling of the building. The smart heat data are used to derive the temperature setpoint and U-value of the building envelope. However, the number of occupants and their heating habits [14] and the number and use of appliances [20,21] are different and dynamic. Thus the proposed methodology needs to adjust for such dynamic boundary conditions.

Recently, studies have also shown that real-time heat data could facilitate the field of urban building energy modeling (UBEM) in the calibration of archetype building energy models [22], for modeling of demand-respond [23], and load forecasting [24]. The smart electricity meter data have proven to provide great environmental, social and economic benefits [25]. However, the research work on smart heat data is still in its infancy, and we have yet to discover the knowledge gain captured in real-time heat data to speed up the green transition of building stock and energy systems. Therefore, there is a great need for research on identifying novel methodologies to convert the promising potential of smart meters to transform conventional buildings into energy-optimized and smart buildings.

3. Methodology

The dataset provided by the DH company Aalborg Forsyning [26] consists of heat meter recordings from 1665 buildings (mostly single-family dwellings) located in a small town in the North of Jutland, Denmark. All the smart energy meters installed in the buildings measure the cumulative energy and fluid usage, the hourly-averaged supply and return fluid temperatures and the instantaneous measurements of fluid flowrate, supply and return temperatures. The devices also store the faults detected during operation. The space heating and heating for domestic hot water production are accounted together. The measurement time period spans from the October 1, 2018 until the October 7, 2019. The measurements are recorded every hour. Because of the scope of the research project supporting the current study, only the single-family residential buildings connected to the DH network were studied.

3.1. Data pre-processing

In this study, the methodology to pre-process the DH dataset is similar to the one described by Johra et al. [18]. This algorithm performs the cleansing and treatment of the data for further analysis. The flowchart in Fig. 1 details the different steps of the data treatment.

The first part of the algorithm organizes the dataset, performs resampling, and quantifies the number of anomalies detected by the smart energy meters and the possible data outliers. The second part of the algorithm determines the number of missing data points in the DH dataset as well as the gap sizes of the consecutive missing values for each building. Buildings with data gaps larger than nine consecutive hours or buildings with more than 10% of data missing are excluded from the rest of the analysis. Different imputation techniques are then applied to generate the remaining missing data points. In this case study, the most suitable imputation for cumulative values was found to be linear interpolation. For instantaneous values, it is the exponential weighted moving average with a window-size of 8 data points before and after the missing value. In the last part of the algorithm, a well-known statistical technique of outlier detection is applied. It considers values higher or lower than 1.5 of the variable's interquartile range as an outlier. This technique evaluates the maximum cumulative energy and fluid-volume consumption parameters and the maximum mean calculated power to identify the non-residential buildings. The latter can then be excluded from further analysis to fit the scope of the current study. As an output of this pre-processing algorithm, when applied to the dataset of 1665 buildings, only 969 buildings (58.2% of the original dataset) fulfill all the requirements to be used for further analysis.

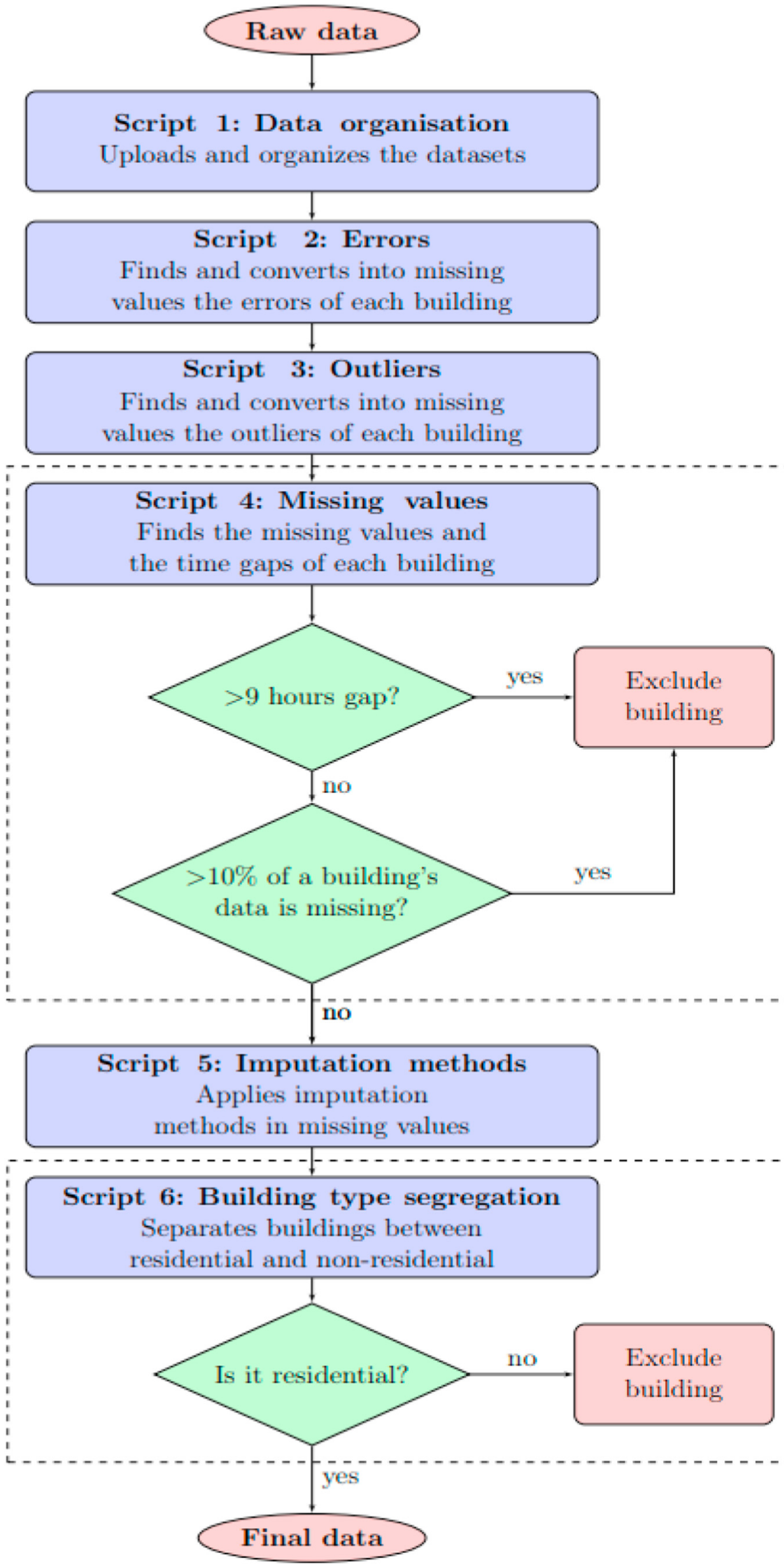


Fig. 1. Pre-processing algorithm.

The weather data (outdoor temperature, wind speed and solar radiation) from the local weather station is also integrated into the dataset.

In addition, information extracted from the EPC of each building is integrated into the dataset. This EPC reports the building energy use and production from which an energy label is issued. Label A represents the best energy performance, while the letter G is the worst energy performance grade that can be obtained [27]. Because EPC is only mandatory for buildings that are for sale or rent, EPC information could not be extracted for all buildings of the case study. For the available EPC reports, the extracted parameters are the construction year, the year of major renovation (if any), the heated surface area (m^2), the total specific heat losses from the opaque and glazed envelope (W/K), the solar exposure of the glazed elements, the volumetric flow rates of natural and mechanical ventilation for Winter and Summer seasons (litres/s.m^2), the description of the ventilation and the space heating systems, the estimated energy usage (kWh/m^2 year) and its associated energy label. From the dataset of 969 buildings, a subset of 41 buildings was selected as they present good quality EPC information.

3.2. District heating variables and coefficients

The recorded variables from the smart energy meters data are as follows: the cumulative energy usage (E_{cum}), cumulative fluid use (V_{cum}), the product between the cumulative water use and the hourly-averaged supply and return temperatures ($V_{cum}\overline{T}_s$ and $V_{cum}\overline{T}_r$). From these measurements, other variables are calculated, as described hereafter.

3.2.1. Hourly-averaged temperature and temperature difference

The hourly-averaged temperature recorded by the sensors is calculated back from (1):

$$\overline{T}_x = \frac{V_{cum}\overline{T}_x}{V_{cum}} \tag{1}$$

Where the variable $\overline{T}_x$ can be either the supply or the return fluid temperature. With the estimated hourly-averaged temperatures, the temperature difference is calculated as:

$$\overline{\Delta T} = \overline{T}_s - \overline{T}_r \tag{2}$$

3.2.2. Hourly energy usage and fluid volume use

The smart energy meters measure the building energy usage and fluid use by summing it up with the previous measurements. The current hourly values of energy usage and fluid flow are thus obtained by subtracting the previous data point from the current one:

$$E_i = E_{cum,\,i} - E_{cum,\,i-1} \tag{3}$$

$$V_i = V_{cum,i} - V_{cum,i-1} \tag{4}$$

3.2.3. Building thermal characteristics: heat transmittance, ventilation/infiltration and solar gains

The heating demand of a building is calculated from the steady-state energy balance between the heat losses through the envelope, ventilation and infiltration, and the heat gains from solar radiation and other internal loads (occupants, equipment, etc.):

$$E_{demand} = E_{trans} + E_{vent} - E_{solar} - E_{int} \tag{5}$$

The sum of heat losses by transmission, ventilation and infiltration can be expressed as a function of the temperature difference between the indoor and outdoor environments:

$$E_{trans} + E_{vent} = (UA + nc_p\rho)(T_{int} - T_{out}) \tag{6}$$

Where the U-value is the overall thermal transmittance of the building (W/m^2 K), A is the overall envelope area (m^2), n is the volumetric flow rate of the ventilation and infiltration (m^3/s), C_p is the specific heat capacity of the air at a constant-pressure system (J/kg K), and ρ is the air density (kg/m^3). By using the values calculated in equation (3), a correlation between heating demand and outdoor temperature is deduced for each building. Fig. 2 shows a scatter point plot from one of the buildings in the dataset where the heating season points (Autumn, Winter and Spring) are marked in red and the no-heating season (Summer) in blue.

However, one can see that there is no clear correlation between energy consumption and outdoor temperature beyond the high levels of heat demanded when the outdoor temperature is low. This unclear relationship between the variables might be due to the low resolution of the energy meter and the DHW energy share mixed with the space heating needs. To tackle that, the hourly energy data points were summed up for each day, and the outdoor conditions were averaged over the same period of time. By using this methodology, similar to the one used in [19], a more explicit relationship between the outdoor temperature and the energy demand can be observed (see Fig. 3).

The shape of the cloud of points in Fig. 3 is commonly called "hockey-stick" by utility companies. It represents the piecewise linear correlation between the heat demand and the outdoor temperature during the heating season and the constant energy values during the season without space heating needs and only domestic hot water production.

The linear relationship $y = mx + b$ between these two variables during the heating season can be linked to the steady-state energy balance of the building by combining equations (5) and (6):

$$E_{demand} = E(T_{out}) = -(UA + nc_p\rho)T_{out} + (UA + nc_p\rho)T_{int} - E_{solar} - E_{int} \tag{7}$$

Where the m- (slope of the line) is the term $-(UA + nc_p\rho)$ and the b- is the term $(UA + nc_p\rho)T_{int} - E_{solar} - E_{int}$:

Heating season linear regression:

$$E(T_{out}) = m_{Heating} \times T_{out} + E(T_{out} = 0) \tag{8}$$

No heating season linear regression:

$$E(T_{out}) = E\left(T_{out} = T_{No\ Influence}\right) \tag{9}$$

As seen, the m- is a descriptive value of the building characteristics, which is dependent on the ventilation/infiltration levels and the transmission losses. The value $T_{No\ Influence}$ is an outdoor temperature threshold under which the heating demand is no longer dependent on the outside temperatures. This may represent the outside temperature when the heating system is turned off. In order to quantify the parameters of (7), it is necessary to isolate the outdoor factors. The outdoor cause that influences the solar gains is solar radiation, and the infiltration/ventilation losses are mainly dependent on the wind speed. Therefore, solar radiation and wind speed are used as filtering conditions to isolate data points that are more conditioned on one energy component than the others. In order to reduce the impact of the internal gains and the potential

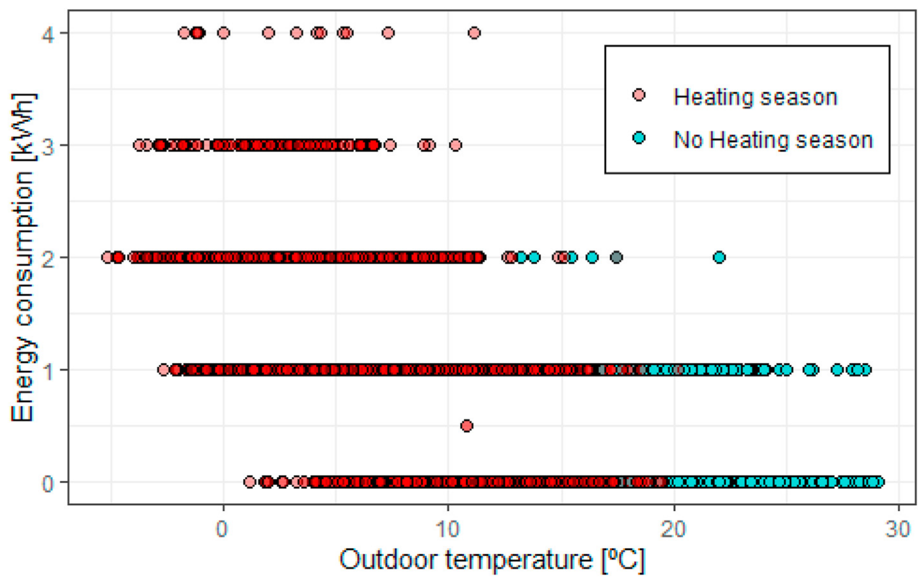


Fig. 2. Scatterplot between energy usage and outdoor temperature for a particular building.

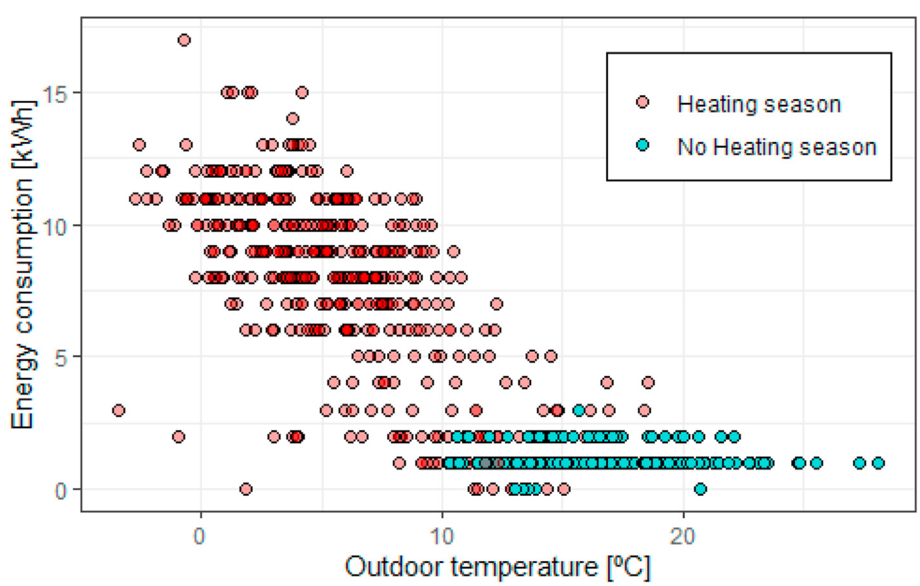


Fig. 3. Scatterplot between energy usage and outdoor temperature of a particular building with daily aggregated values.

natural ventilation from the opening of windows, the data points in the time intervals 11:00–16:00 and 23:00–04:00 were selected, as people are usually not at home or are sleeping. Fig. 4 presents the distribution of different weather parameters. One can see that the outdoor temperature and the wind speed have a quasi-normal distribution with a mean value of 6.4 °C and 3.6 m/s, respectively. In terms of solar radiation, the mean value is 103.5 W/m^2, with most of the values below 100 W/m^2, which is expected for Denmark.

The filtering conditions to be applied in each building of the dataset were selected by taking into account the weather variable's distributions (see Table 1).

The temperature/transmission component is isolated in the dataset when considering zero solar radiation (night period) and the wind velocity lower than 2 m/s. The same reasoning is followed for the ventilation/infiltration factor, but considering the data points where the wind speed is higher than 3 m/s. The solar gain component is filtered when the radiation is higher than 30 W/m^2 (daytime), and the wind velocity is lower than 2 m/s. The same methodology of linear regression is then reapplied on these new filtered subsets to obtain the parameters that are dependent on a specific energy component for the heating season:

Transmission losses condition:

$$E(T_{out}) = m_1(UA) \times T_{out} + b_1(UA) \tag{10}$$

Solar gains condition:

$$E(T_{out}) = m_2(UA) \times T_{out} + b_2(E_{solar}) \tag{11}$$

Ventilation and infiltration losses condition:

$$E(T_{out}) = m_3(n) \times T_{out} + b_3(n) \tag{12}$$

Equation (10) is the linear regression made from the data points

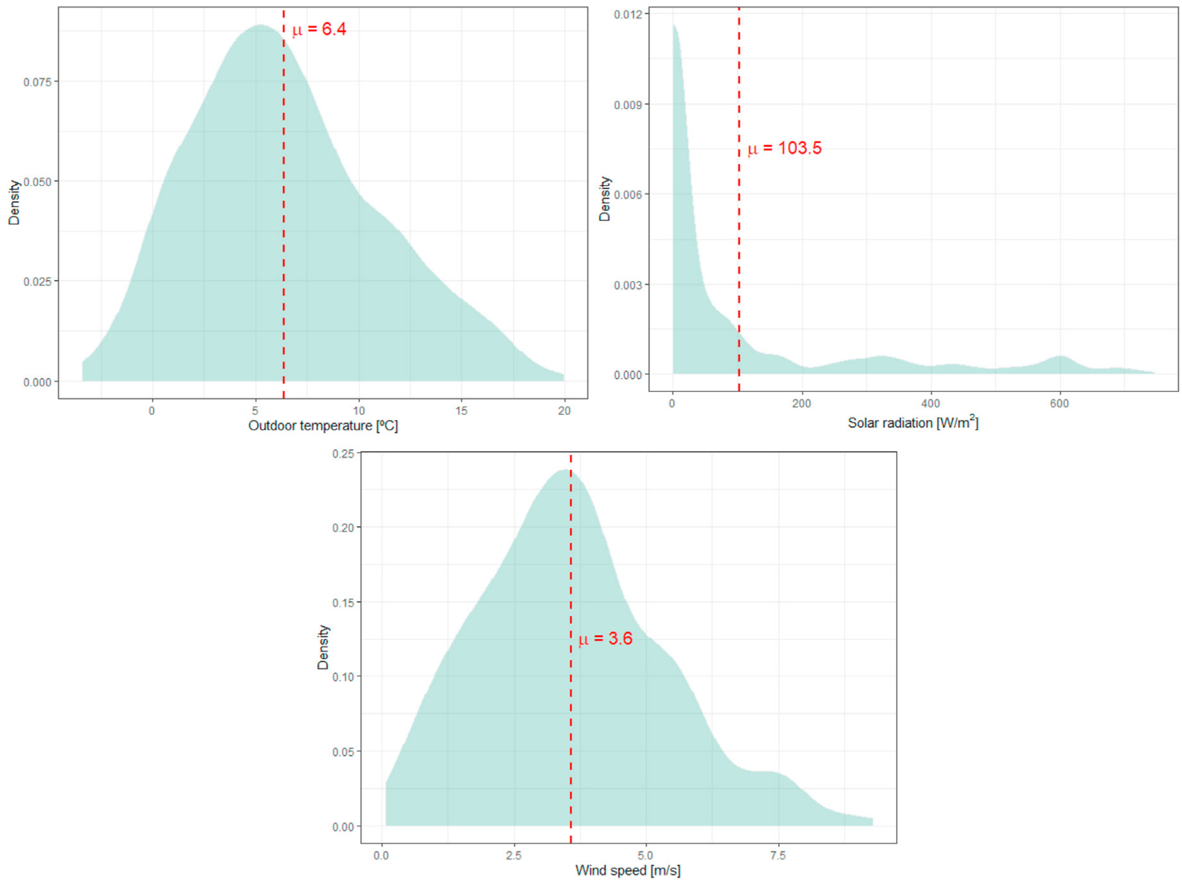


Fig. 4. Outdoor variables distribution.

Table 1
Subset filtering conditions.

Isolated energy component	Solar radiation ($\boldsymbol{R_d}$) [W/m^2]	Wind speed ($\boldsymbol{v_{wind}}$) [m/s]
Transmission losses (Outdoor temperature)	0	≤2
Solar gains (Solar radiation)	>30	≤2
Ventilation and infiltration losses (Wind speed)	0	>3

subset that is highly correlated with the outdoor temperature. Therefore the value m_1 is much more dependent on the constant *UA* than the air rate from the ventilation/infiltration component. In equation (11), b_2 is much more correlated with the solar gains than with the other energy parameters. In equation (12), by following the same reasoning, the slope m_3 characterizes the ventilation and infiltration impact on the building's heating demand. It is still important to highlight that these values are still influenced by the outdoor temperature, the building's air change rate, indoor temperature setpoint and internal gains. Regarding the values extracted from the equations, it is concluded that the higher the absolute value of m_1, the higher the heat transmission losses will be through the building's envelope. The b_2 has the solar gain as negative term; therefore, the lower this coefficient is, the higher the solar gains in the building will be. This happens in buildings with large glazing areas, low use of solar shading, or most glazing surfaces facing south. The coefficient m_3 is much more dependent on the energy losses due to wind speed. For this value, the higher its absolute value is, the higher the impact of wind on a building. This is seen in buildings with high infiltration/ventilation rates due to their ventilation system, or where the windows are opened frequently with the heating system in operation or due to high air leakage.

3.3. Comparison of the district heating coefficients with the EPC results

After extracting all DH variables from the dataset and calculating the linear regression coefficients, the derived building characteristics are compared with the building information reported in the EPC. These inputs are the information collected by the EPC certifier during the assessment and are considered reference values to evaluate the accuracy of the coefficients. The inputs are valuable to the utility companies when analyzing the building heating data because they are more detailed than what is available in the Danish building and housing register.

At this research stage, all EPC information is extracted manually to be analyzed. Therefore, a subset of 41 buildings was used. Among the subset, five buildings had significant renovations. All the buildings have a similar heated surface area, with a mean and standard deviation of 137.1 ± 32.1 m^2. The main ventilation system is natural ventilation. Few of them also have mechanical ventilation systems installed. Apart from two cases with indirect connection, the space heating system of the buildings is directly connected to the thermal grid without any intermediate heat exchanger. The domestic hot water is produced in a heat exchanger, often located

Table 2
Number of buildings with the same EPC label.

EPC label	Number of buildings
A2010	2
B	16
C	8
D	10
E	3
F	1
G	1
Total	41

in the utility room. Each building is assigned an energy label, which represents its energy performance level. One can see in Table 2 the number of buildings for each energy label.

The first variable that will be compared is the difference between the total measured energy usage by the smart meters and the annual estimated energy by the EPCs (kWh/m^2 year):

$$\Delta E = E_{EPC} - E_{DH} \tag{13}$$

The EPC annual energy, E_{EPC}, is the predicted household's energy usage concerning space heating, DHW and the electricity consumption by the buildings' installed systems. Because it is only being studied single-family houses, the majority of the E_{EPC}-value is the sum between space heating and DHW, as it is in the smart meter's measurements.

In equation (13), if the value ΔE is negative, then the EPC underestimated the building's energy usage and vice-versa. If the EPCs are in good agreement with the smart meter recordings, it might be argued that the former can be used by the utility companies when designing the expansion of their network.

The main parameters that are extracted and calculated from the linear regressions are m_1, b_2 and m_3. The estimate of these values can be compared against the input parameters from the EPC. For the case of m_1, its dependency is with the transmission losses of a building. From the EPC, the value used to test m_1 is the total specific heat loss from the opaque and glazed elements on the building. The parameter b_2 quantifies the dependency of solar radiation on the overall energy usage. Therefore the coefficient was compared with the heat gain share from the heat balance of all windows in a building. The heat gain share in the energy balance is dependent on the window's area, orientation, inclination and total solar energy transmittance (g_w), which are all described in the EPC. The ventilation and infiltration losses (m_3) are compared with the ventilation heat transfer coefficient provided by the TABULA WebTool [28].

4. Results

4.1. Comparison between district heating smart meter results and building EPCs

4.1.1. Smart energy meters measurements and EPC energy predictions

In Fig. 5, one can see the relationship between the EPC estimated energy consumption and the energy usage measured by the smart energy meters. The data points in the red upper zone indicate an

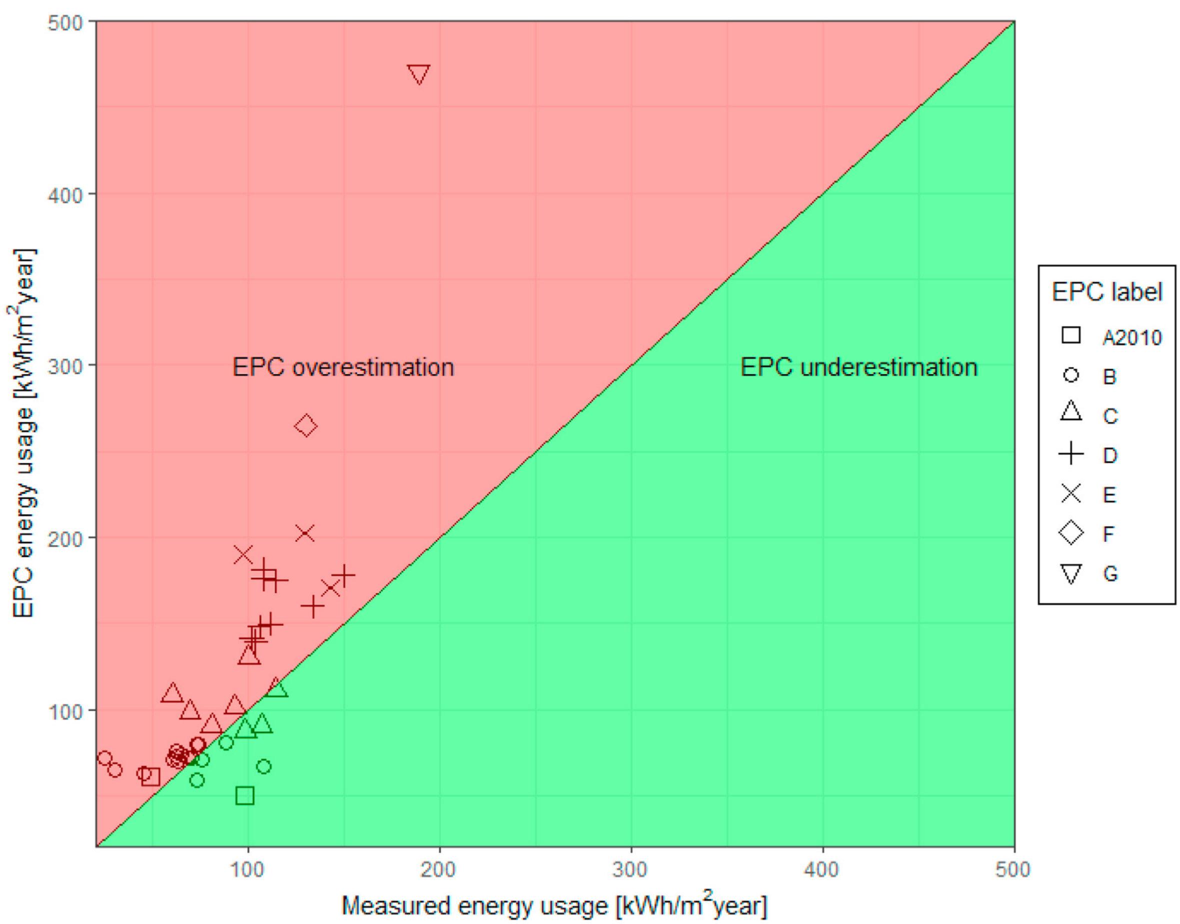


Fig. 5. Relationship between EPC energy prediction and the measured building's consumption.

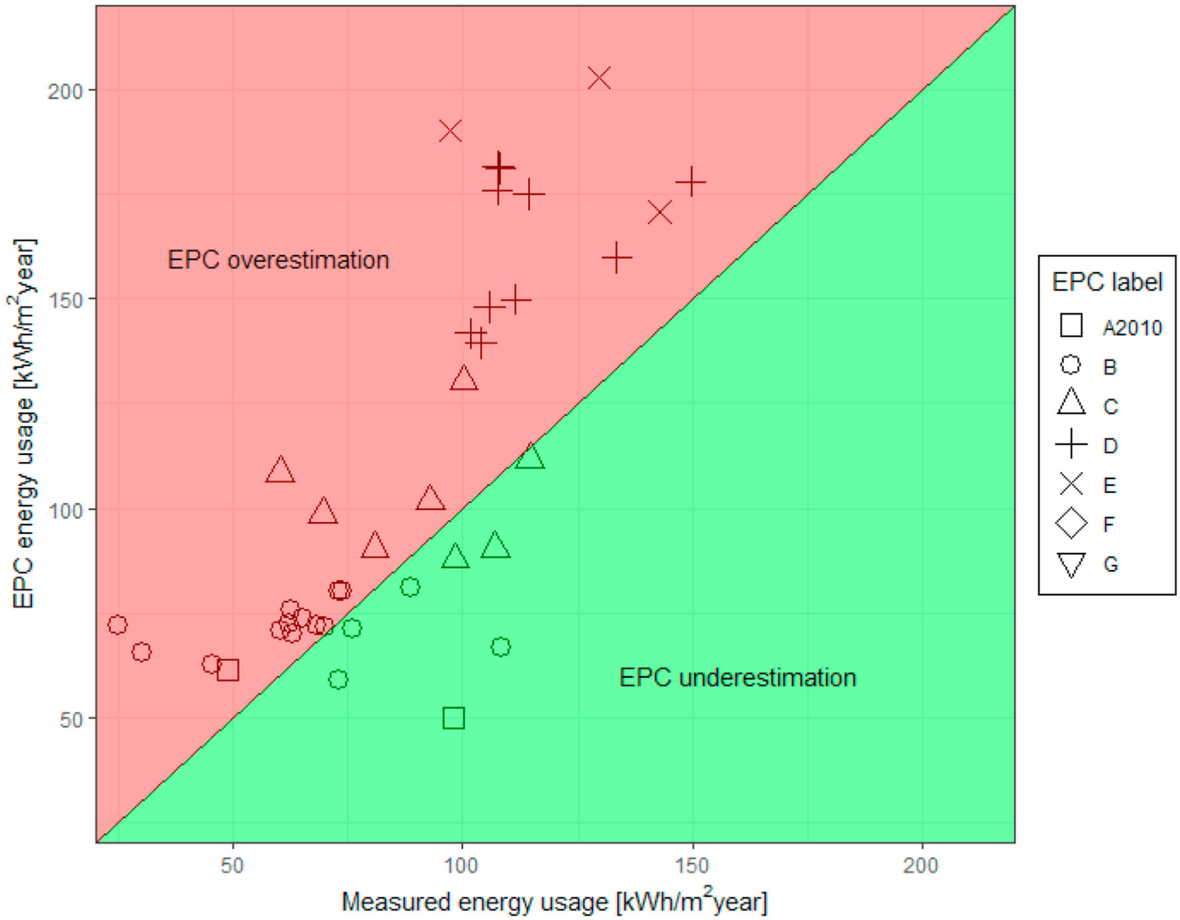


Fig. 6. Relationship between EPC energy prediction and the measured building's consumption (Zoom).

overestimation of the energy use by the EPC. Conversely, the green zone's data points indicate an underestimation of the energy use by the EPC.

Fig. 6 shows the same plot as in Fig. 5 but focuses on the lower range of energy use.

The figure above shows a significant mismatch between the EPC estimation and the actual heating need of many buildings throughout all energy label categories. Building energy performance from the EPC can thus be a problem if used by the DH utility companies when planning the extension of their thermal grid. Estimating the heat demand from the EPC of existing buildings connected to the DH network could lead to a large oversizing of the latter. Unfortunately, the size of the analyzed sample is too small to make definite conclusions.

4.1.2. Smart energy meters measurements and EPC building thermal characteristics

Regarding the building's thermal characteristics, a methodology was developed to quantify the envelope and ventilation/infiltration heat losses and heat gains from solar radiation. These characteristics can be assessed in a more accurate manner than what is stated in EPCs. It is thus possible to identify the main reason behind the high energy demand of certain buildings. Building system faults could thus be detected. A renovation scheme could also be suggested to the households.

In Fig. 7, the correlation between the specific transmission heat losses from the EPC and the m_1 coefficient is presented.

The specific transmission losses from the EPC are the product between the total area of the opaque and glazed envelope elements and their thermal transmittance. There is a good correlation between the EPC values and the m_1 coefficients obtained by linear regression on the smart meter data. This implies that this methodology is suitable to understand the building's transmission heat losses from the smart energy meters. Hence, it can be used as an indicator to the utility companies to identify the buildings where their space heating usage is highly dependent on the envelope heat losses.

Regarding the solar gains, the b_2-coefficient was calculated. This value is connected with the impact that solar radiation has on the building's heat demand. And the lower this variable is, the higher the solar gains are in a building. In Fig. 8, three of the EPC label categories with more buildings in the subset are represented, with each building's b_2-coefficient and their associated overall solar exposure.

In the plot, each point represents a building, and it is expected that the buildings with more prominent south solar exposure have lower b_2 (green data points). However, this relation is not observed. In label B, there are some buildings with low b_2 coefficients exposed mainly to the north. As for the label C and D buildings, some south solar-exposed buildings have large b_2-values. The reason behind it might be that the building's linear regressions are not adequate and that other unknowns significantly impact the building's energy performance.

Concerning the ventilation and infiltration losses, they are estimated by using the m_3 coefficient. To evaluate this variable's relevance, the correlation between the m_3 value and the expected heat transfer coefficient by ventilation in the buildings was estimated. Moreover, the assessors have sometimes described the

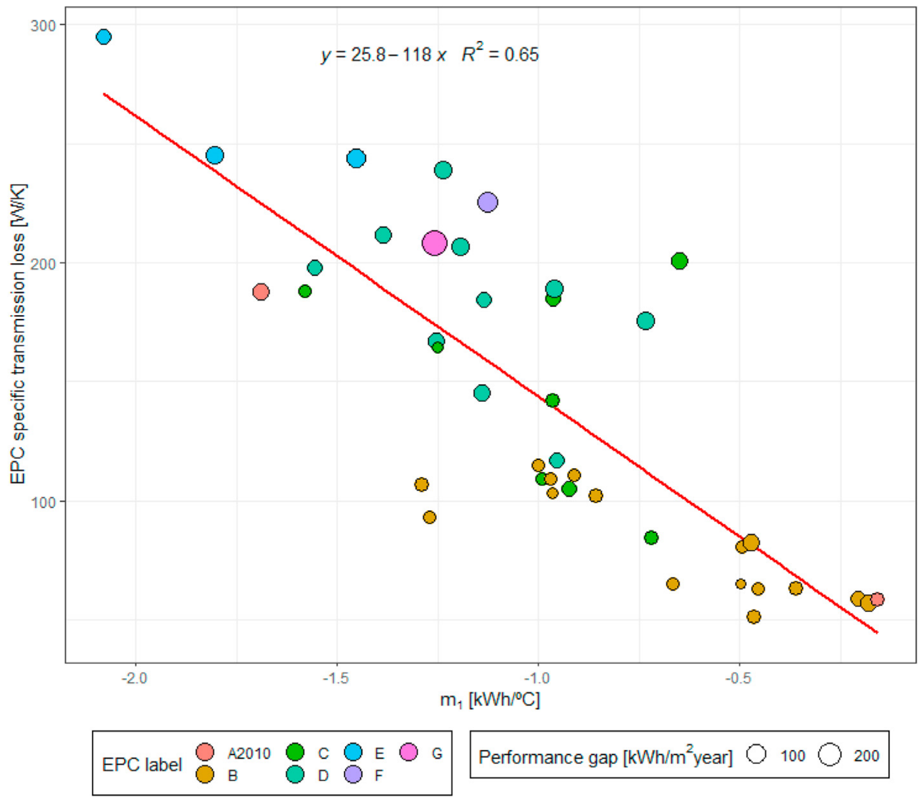


Fig. 7. Correlation between EPC specific transmission losses and the m_1-coefficient.

condition of the building joints (construction joints, joints at windows and door openings) to assess the building's leakage. Therefore the condition of the building joints was also compared to the m_3 coefficient (see Fig. 9).

In the scatterplot, one can observe a reasonably good correlation between what are the expected losses and the m_3 values. Therefore this method might be adequate to identify buildings with high ventilation and infiltration losses. Another point worth mentioning is regarding the building (Label A2010) with the lowest m_3 coefficient (-2.19 kWh/°C). Even though it is newly built and with a good EPC label, it presents high ventilation and infiltration losses. In the EPC assessment, this is one of the few buildings with mechanical ventilation, which might be the reason for high ventilation/infiltration losses and, therefore, a low m_3 coefficient. This also shows that the utility companies can also detect the households with large heat demand due to ventilation and advice their customers to take actions to reduce their consumption. Concerning the joints condition, buildings with low m_3 coefficients are expected to be less air tight. In the boxplot, a large portion of the building did not have a description (Not defined). The "Reasonable" conditions have low m_3 values, which is expected. The description "Not good" is only two buildings where its m_3 should be much lower than it is. Therefore, by observing the plot, the sample is too small to draw meaningful conclusions regarding the buildings' envelope airtightness.

Even though this methodology is quite promising due to its simplicity, it might not perform very well, in some cases, when compared with the EPC values. Several reasons might explain the poor performance of this method. The energy consumption measured by the devices is for space heating and DHW production. Therefore, it is hard to isolate the energy required only for space heating, which depends on the outdoor conditions. Also, the method used in this study tried to evaluate the different energy components on the building's heat balance by filtering the data points that have adequate outdoor conditions to eliminate certain terms of the heat balance. However, this filtering might not be perfect or not minimize enough certain gain terms in the energy balance. Additional unknowns might also have a significant impact that cannot be captured by the energy meters, EPC or weather station, e.g., people behavior, natural ventilation through windows openings, internal gains. Furthermore, in the filtering conditions, another problem that might cause the methodology not to work correctly is the studied location's meteorological characteristics. This study was performed in Aalborg, which is known for being a windy city with low sun exposure; therefore, the data points will not be equally distributed, creating unrealistic linear regressions without any physical meaning, as seen in Fig. 10.

Because there are not that many daylight hours during the heating season in Aalborg, it might be the main reason for the b_2 coefficient to contradict the EPC results. So, this coefficient might be more accurate for countries where the daylight hours are much higher than Denmark. The methodology accuracy was also tested when compared with the values from the EPCs. The EPCs are highly dependent on the assessor's knowledge and inputs as well as the building standard values provided in the national building standards. However, the standards and EPC inputs may not be the actual building's values, contradicting the calculated coefficients consequently. Another reason worthy of mentioning is the small EPC sample size (only 41 buildings).

4.2. Interactive web-based interface for data visualization - shiny

The methodology presented above is of great interest for households, urban planners, municipalities and utility companies

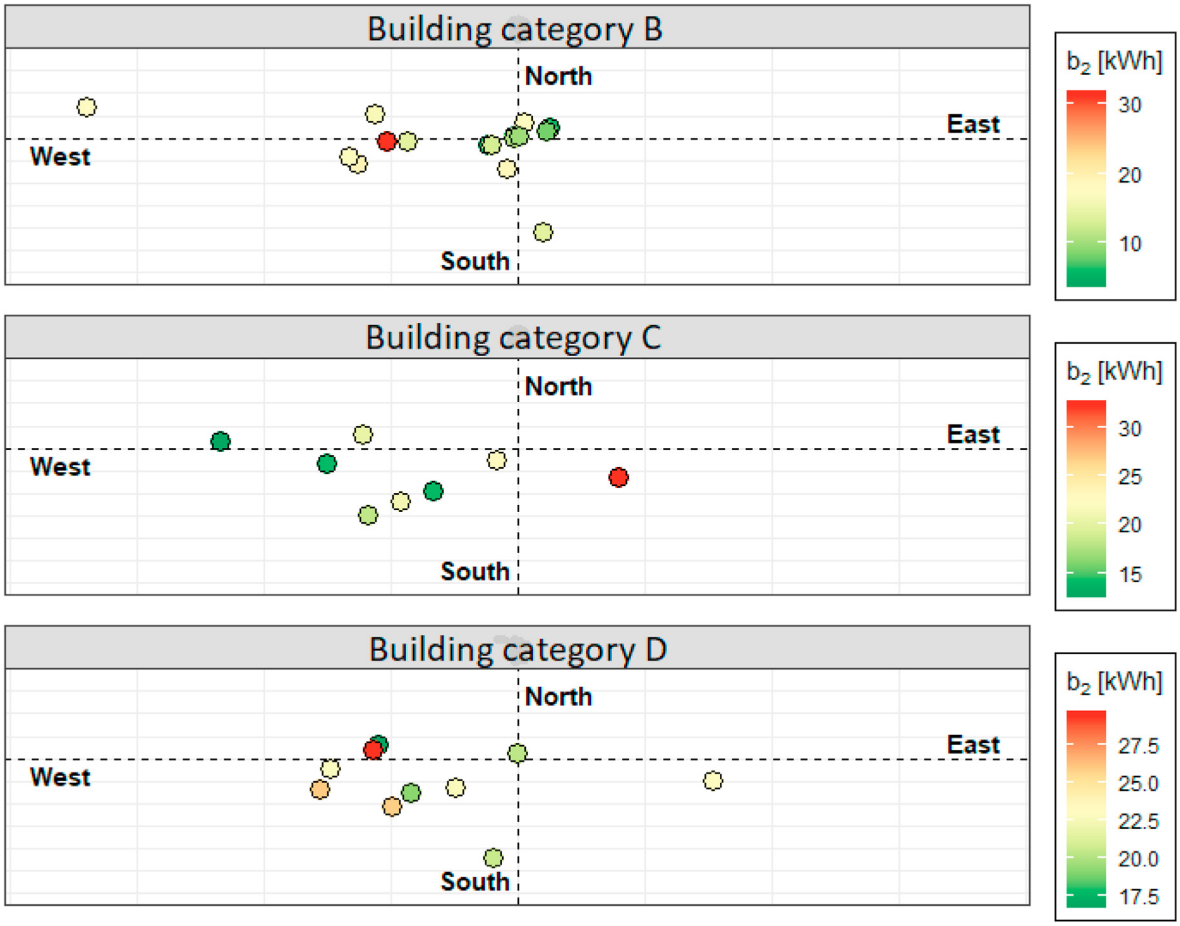


Fig. 8. Relation between EPC overall solar exposure and the b_2 coefficient.

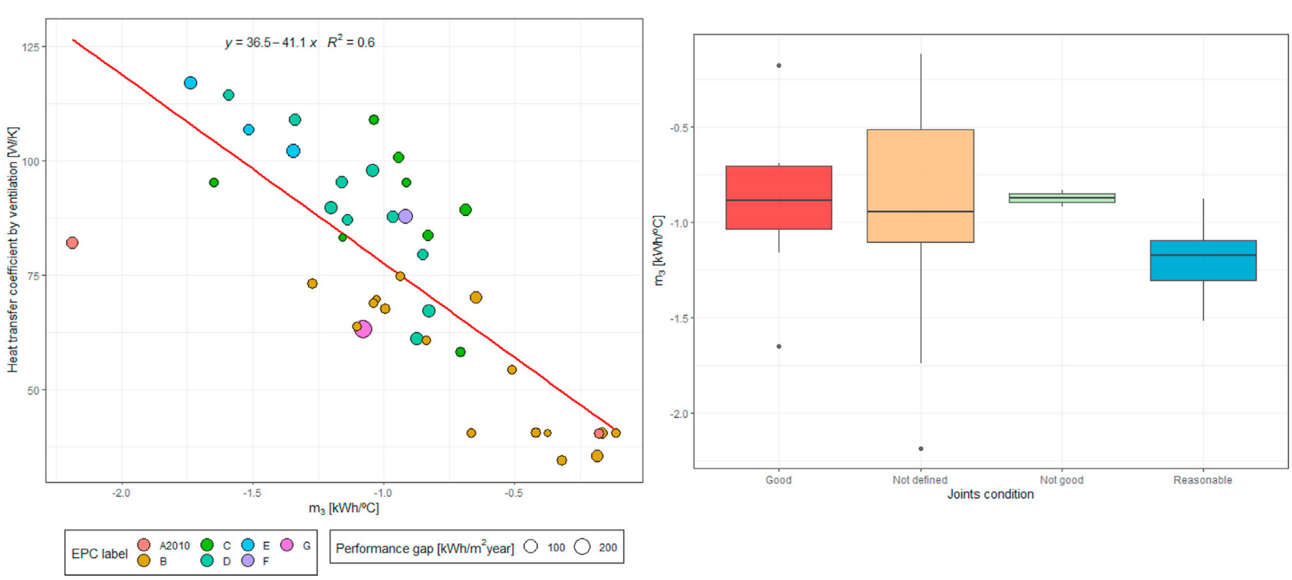


Fig. 9. Correlation between EPC specific ventilation loss and the m_3 coefficient. Relation between the building's joints condition and the m_3 coefficient.

managing energy distribution grids like district heating networks. However, sharing so much information with building professionals in a concise yet meaningful and flexible way can be challenging. Given this, it was chosen to bundle all the data analysis results of this study into a web-based interactive map. Maps are intuitive tools to rapidly grasp an overview of the state and characteristics of a given cluster of buildings or an entire city.

The "Shiny" package [29] is a free library for the R programming environment that enables the simple development of web-based graphical interfaces to display data plots and navigable satellite

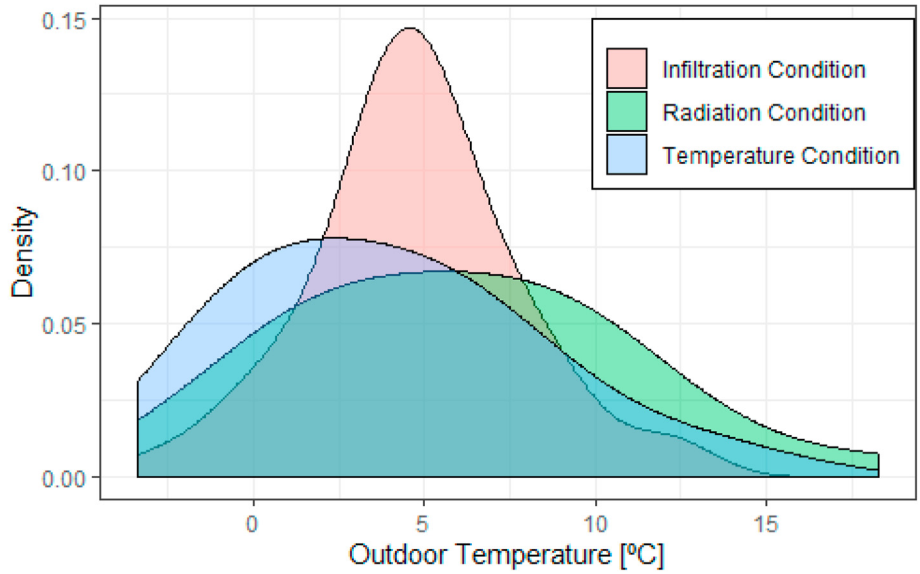


Fig. 10. Filtered data points distribution.

maps with super-imposed graphics, symbols, and data points (bubble maps). A dedicated data visualization interface has thus been created with the "Shiny" package to present the processed smart energy meter data to a larger audience.

In Fig. 11, one can see an overview of the web-based interface. The user can navigate on the map and display overlaying colored points (corresponding to a color scale) for the following building characteristics: district heating fluid supply temperature, the temperature difference between supply and return fluid of district heating, yearly volume of fluid passing through the substation of the building, yearly heating demand per m^2, clustering categories, amount of erroneous data from the smart meter, and amount of missing data from the smart meter. The user can filter the visible building data points on the map by selecting the filtering range corresponding to the aforementioned parameters.

It is also possible to include all the other building characteristics calculated during the data processing or extracted from the national building register. However, for clarity, these have not been included here.

In the interface, the user can select a specific point on the map and display a summary of the building characteristics (yearly values) together with the address, year of construction and energy label. The user can also open the data time series of the selected case and browse through it by selecting a specific period of time. Furthermore, display correlation plots between different measurement parameters in a selected building pop-up window.

Finally, a parallel plot (see Fig. 12) can be generated to give an overview of all analysis results and building parameters of the different households. Each line passing through the different parameter columns is a unique building case. The user can narrow

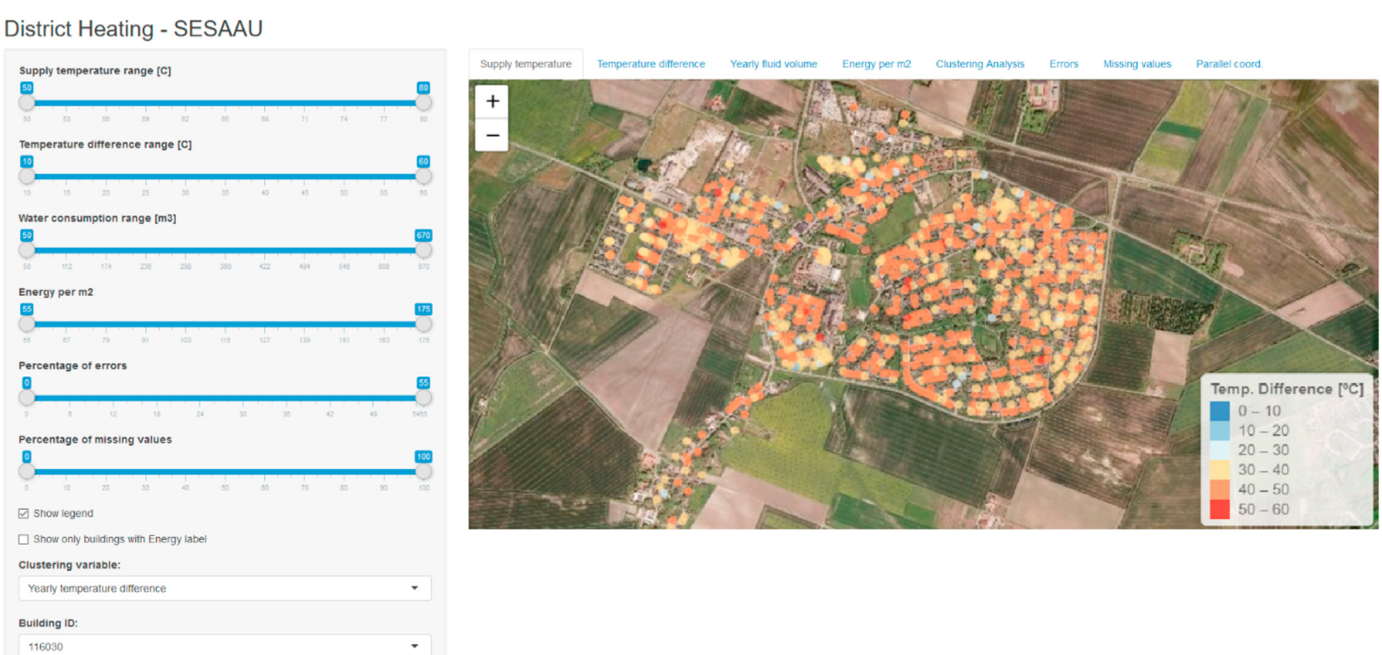


Fig. 11. Overview of the graphical user interface to visualize this study's processed data: district heating temperature difference to buildings in a specific region.

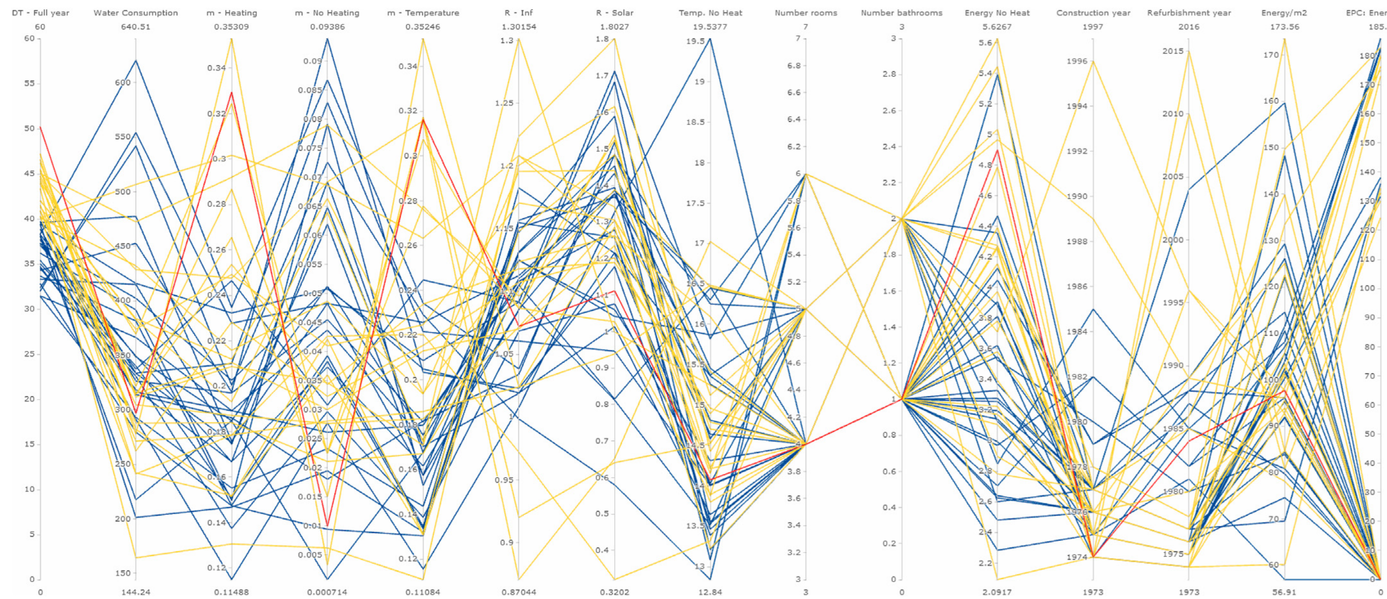


Fig. 12. Parallel plot of the building characteristics for some of the buildings in the case study.

down each parameter's range to identify and compare the other characteristics of the remaining building cases. The parameter and characteristic columns can be manually swapped around by the user.

5. Discussion

The systematic implementation of smart meters in the district heating network opened up for the utilities to learn more about their customers. The potential of this action is significant and should be beneficial for both the production and demand-side (i.e., utilities and customers). The digitalization of the demand side is also foreseen as the key component of the district heating transition towards 4th generation systems (i.e., 4GDH).

The presented methodology shows how utilities can identify customers with high energy use and determine the reasons for their performance. DH utilities can thus provide their customers with tailored-suited energy-efficiency actions, thereby lowering the network energy demand more efficiently, allowing a successful transition towards 4GDH.

Another pillar of the 4GDH concept is the high share of fluctuating renewable energy sources (e.g., wind and sun) on the production side. The methodology provides utilities with knowledge on which users are expected to use more or less heating depending on weather conditions like wind speed and solar radiation. Commonly the energy use at the demand side is correlated only with the outdoor temperature. Neither solar radiation nor wind conditions are taken into consideration when sizing the production mix. With a higher share of intermitted RES in 4GDH, the knowledge on the expected energy use must be better foreseen.

The expansion of hourly data from smart heat meters delivers great potential to learn more about buildings. In Figs. 11 and 12, one can see the easiness of identifying the building characteristics, the data outliers, and the corresponding clients. For example, a district heating company can map the clients with a high percentage of missing data or erroneous data from their smart meters and send a technician to verify and repair the latter. The utilities would also be very interested in identifying clients with a very low fluid-temperature difference because they lessen the district heating system's energy efficiency.

6. Conclusion

In this study, a simple methodology was used to treat and analyze the data recorded by smart energy meters installed in 1665 buildings connected to a district heating network. Regarding the different variables measured by the devices, all of them can be used to evaluate the building's energy performance concerning the different building characteristics, systems and user's behavior. To evaluate the validity of the methodology, its results were compared with the information from energy performance certificates of a smaller sample (41 buildings).

The methodology developed in this study aimed to assess several household characteristics from the smart meter data analysis: the actual heating season in the building and the influence of the outdoor temperature, wind speed, and solar radiation on the energy usage. When compared to EPC information (considered as reference), the simple linear regression method gives positive results for the outdoor temperature and wind speed influence. However, it is not conclusive for the sensibility of solar radiation. For the Danish case, the outdoor temperature and the wind speed influence can be assessed by the utility companies to understand the source of significant heat losses in the buildings connected to the grid. Even in the present research, it was seen that a highly efficient building had significant ventilation and infiltration losses, and therefore a large performance gap, most likely due to their ventilation system operation. To assess the solar gain, it is expected that the methodology might work in countries with more considerable daylight hours.

In the paper, it was also compared the smart energy meters measurements with the EPC estimations. The small sample used shows that the difference between energy estimations and measurements increases for buildings labeled as low-energy efficient. In terms of District Heating systems, the EPC inputs, even though used to test the accuracy of the linear regression method, cannot be used as guidelines for design and planning the creation or expansion of the DH networks.

In this study, a simple data visualization interface created with the “Shiny” package (R) is presented as a starting point for creating a real tool that can be used for the utility and consulting companies to analyze the energy meter data and detect possible problems occurring at their customers.

In the present paper, several parameters increased the uncertainty associated with this simple methodology. It is required to adjust this methodology with more DH data points, more buildings with EPC information, and higher resolution DH data with indoor measurements that will clarify which values are more accurate, the linear regression coefficients or the EPC inputs. As suggestions for further work, it is idealized to use other linear regression algorithms that are less susceptible to the different outdoor conditions distribution and outliers and apply grey-box models with system identification techniques.

Author contributions

Daniel Leiria: Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Validation; Visualization; Writing - original draft; Writing - review & editing. **Hicham Johra:** Investigation; Methodology; Project administration; Resources; Supervision; Validation; Visualization; Writing - original draft; Writing - review & editing. **Anna Marszal-Pomianowska:** Methodology; Project administration; Resources; Supervision; Validation; Writing - original draft; Writing - review & editing. **Michal Zbigniew Pomianowski:** Project administration; Supervision; Writing - review & editing. **Per Kvols Heiselberg:** Funding acquisition; Supervision; Writing - review & editing.

All authors have read and agreed to the published version of the manuscript.

Funding

This work was part of the E-DYCE project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 983945. This work was also supported by the InterHUB project funded by Aalborg University's strategic funding for interdisciplinary research.

Declaration of competing interest

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

Acknowledgments

The authors would like to thank Aalborg Forsyning that contributed by providing the data from smart energy meters. And to Flávio do Vale Wichman for contributing with the data from the building's EPCs.

References

[1] Energi- Forsynings- og Klimaministeriet. “Denmark's integrated national energy and climate plan under the regulation OF the EUROPEAN parliament and OF the council on the governance of the energy union and climate action. 2019.

[2] Energy performance of buildings directive. https://ec.europa.eu/energy/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en. [Accessed 25 March 2021]. accessed.

[3] Top district heating countries – euroheat & power 2015 survey analysis | euroheat & power. https://www.euroheat.org/news/district-energy-in-the-news/top-district-heating-countries-euroheat-power-2015-survey-analysis/. [Accessed 25 March 2021]. accessed.

[4] Lund H, Möller B, Mathiesen BV, Dyrelund A. The role of district heating in future renewable energy systems. Energy Mar. 2010;35(3):1381–90. https://doi.org/10.1016/j.energy.2009.11.023.

[5] Statista.com. Smart Home penetration rate per segment in Europe_ 2023 _ Statista. https://www-statista-com.zorac.aub.aau.dk/forecasts/887721/smart-home-penetration-rate-per-segment-in-europe. [Accessed 25 March 2021]. accessed.

[6] Smart readiness indicator for buildings | smart readiness indicator for buildings. https://smartreadinessindicator.eu/. [Accessed 25 March 2021]. accessed.

[7] Bsi and Cen, “Bs En 15603. Energy performance of buildings. Overall energy use and definition of energy ratings. Accessed European Standard 2008:1–45 [Online]. Available: https://standards.iteh.ai/catalog/standards/cen/7a0df579-c84c-4223-a270-23785e7e3f9f/en-15603-2008. [Accessed 19 May 2021].

[8] Molin A, Rohdin P, Moshfegh B. Investigation of energy performance of newly built low-energy buildings in Sweden. Energy Build Oct. 2011;43(10): 2822–31. https://doi.org/10.1016/j.enbuild.2011.06.041.

[9] Gram-Hanssen K, Georg S, Christiansen ET, Heiselberg PK. How building regulations ignore the use of buildings, what that means for energy consumption and what to do about it. Jun. 2017. p. 2095–104. Accessed: Mar. 25, 2021. [Online]. Available: https://vbn.aau.dk/en/publications/how-building-regulations-ignore-the-use-of-buildings-what-that-me.

[10] Council of the European Union. “Directive (EU) 2018/2002 of the European parliament and of the council of 11 december 2018 amending directive 2012/27/EU on energy efficiency,” *off.* J. Eur. Union 2018;61(210–328). Accessed: Mar. 25, 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L_.2018.328.01.0210.01.ENG&toc=OJ%3AL%3A2018%3A328%3ATOC.

[11] Lund H, et al. 4th Generation District Heating (4GDH). Integrating smart thermal grids into future sustainable energy systems. Energy Apr. 15, 2014;68:1–11. https://doi.org/10.1016/j.energy.2014.02.089. Elsevier Ltd.

[12] Gadd H, Werner S. Heat load patterns in district heating substations. Appl Energy Aug. 2013;108:176–83. https://doi.org/10.1016/j.apenergy.2013.02.062.

[13] Calikus E, Nowaczyk S, Sant'Anna A, Gadd H, Werner S. A data-driven approach for discovering heat load patterns in district heating. Appl Energy Oct. 2019;252:113409. https://doi.org/10.1016/j.apenergy.2019.113409.

[14] Gram-Hanssen K. Residential heat comfort practices: understanding users. Build Res Inf Mar. 2010;38(2):175–86. https://doi.org/10.1080/09613210903541527.

[15] Gianniou P, Liu X, Heller A, Nielsen PS, Rode C. Clustering-based analysis for residential district heating data. Energy Convers Manag Jun. 2018;165: 840–50. https://doi.org/10.1016/j.enconman.2018.03.015.

[16] Tureczek AM, Nielsen PS, Madsen H, Brun A. Clustering district heat exchange stations using smart meter consumption data. Energy Build Jan. 2019;182: 144–58. https://doi.org/10.1016/j.enbuild.2018.10.009.

[17] Le Ray G, Pinson P. Online adaptive clustering algorithm for load profiling. Sustain. Energy, Grids Networks Mar. 2019;17:100181. https://doi.org/10.1016/j.segan.2018.100181.

[18] Johra H, Leiria D, Heiselberg P, Marszal-Pomianowska A, Tvedebrink T. Treatment and analysis of smart energy meter data from a cluster of buildings connected to district heating: a Danish case. In: In *E3S Web of conferences*, 172; Jun. 2020. p. 12004. https://doi.org/10.1051/e3sconf/202017212004.

[19] Gianniou P, Reinhart C, Hsu D, Heller A, Rode C. Estimation of temperature setpoints and heat transfer coefficients among residential buildings in Denmark based on smart meter data. Build Environ Jul. 2018;139:125–33. https://doi.org/10.1016/j.buildenv.2018.05.016.

[20] Widén J, Lundh M, Vassileva I, Dahlquist E, Ellegård K, Wäckelgård E. Constructing load profiles for household electricity and hot water from time-use data-Modelling approach and validation. Energy Build Jul. 2009;41(7): 753–68. https://doi.org/10.1016/j.enbuild.2009.02.013.

[21] Marszal-Pomianowska A, Heiselberg P, Kalyanova Larsen O. Household electricity demand profiles - a high-resolution load model to facilitate modelling of energy flexible buildings. Energy May 2016;103:487–501. https://doi.org/10.1016/j.energy.2016.02.159.

[22] M. H. Kristensen, R. E. Hedegaard, and S. Petersen, “Hierarchical calibration of archetypes for urban building energy modeling,” Energy Build, vol. 175, pp. 219–234, Sep. 2018, doi: 10.1016/j.enbuild.2018.07.030.

[23] Hedegaard RE, Kristensen MH, Pedersen TH, Brun A, Petersen S. Bottom-up modelling methodology for urban-scale analysis of residential space heating demand response. Appl Energy May 2019;242:181–204. https://doi.org/10.1016/j.apenergy.2019.03.063.

[24] M. H. Kristensen, R. E. Hedegaard, and S. Petersen, “Long-term forecasting of hourly district heating loads in urban areas using hierarchical archetype modeling,” Energy, vol. 201, p. 117687, Jun. 2020, doi: 10.1016/j.energy.2020.117687.

[25] Alahakoon D, Yu X. Smart electricity meter data intelligence for future energy systems: a survey. IEEE Trans. Ind. Informatics Feb. 2016;12(1):425–36. https://doi.org/10.1109/TII.2015.2414355.

[26] Aalborg forsyning. https://www.aalborgforsyning.dk/. [Accessed 25 March 2021]. accessed.

[27] Energimærkning boliger | energistyrelsen. https://sparenergi.dk/forbruger/boligen/energimaerkning-boliger. [Accessed 25 March 2021]. accessed.

[28] Wohnen Institut, GmbH Umwelt. Tabula WebTool. https://webtool.building-typology.eu/#bm. [Accessed 25 March 2021]. accessed.

[29] RStudio. Shiny from RStudio. 2017. https://shiny.rstudio.com/. [Accessed 25 March 2021]. accessed.

2.3 Influence of this research on subsequent studies

Based on this published research work [19], other researchers developed subsequent publications. As one can observe in **Paper 1**, there are three main outcomes, the DH data curation proposal, analysis between the measured energy and estimated energy by the EPCs based on the weather data, and the application of a developed visualization tool for easier understanding of the DH data.

Regarding the data curation proposal, Schaffer, Tvedebrink, and Marszal-Pomianowska (2022) investigated with a larger sample of a DH dataset several imputation methods applied to solve the issue of missing data [18]. Also worth mentioning is Søndergaard, Shaker, and Jørgensen (2024) who worked similar dataset in Odense, Denmark where they cited **Paper 1** regarding its data preprocessing algorithm [33].

Another reason for several citations of **Paper 1** is due to its main analysis using the weather and EPC data. Worth mentioning is Hansen *et al.* (2022) where they investigated further a similar dataset and the household characteristics (employment, number of people, total income, etc.) that might be responsible for the energy peaks at the consumer level [34]. Furthermore, Schaffer, Vera-Valdés, and Marszal-Pomianowska (2024) stepped further into the topic of building thermal characteristics using a co-clustering methodology to group similar energy patterns based on the weather and later investigated if similar clusters are due to alike EPC reports [35].

2.4 Further discussion

The method described in **Paper 1** provides a strong correlation between expected and actual energy losses, being particularly useful for pinpointing buildings that suffer from poor insulation or significant losses due to high ventilation and infiltration. This allows utility companies, EPC certifiers, building owners (and other stakeholders) to identify households with increased heat demand attributable to inadequate ventilation or insulation, enabling them to recommend targeted refurbishments to reduce heat usage.

Despite its utility, the method exhibits certain limitations that may affect its performance, especially when compared to EPC values. Challenges include the impact of variables such as occupant behavior, natural ventilation, and internal gains that are not fully accounted for, potentially skewing results. The meteorological characteristics specific to the study location in Aalborg, such as prevalent windiness and limited sunlight, further complicate data point distribution and the accuracy of the regression models.

Issues with EPC comparisons also arise, as EPC values are dependent on the knowledge of the assessor and national building standards and regulations, which might not accurately reflect the actual conditions of the buildings. Moreover, separating the energy used for space heating from the total heating demand recorded by SHMs is challenging because SHMs measure both space heating (SH) and domestic hot water (DHW) together.

This dissertation progresses to explore how data from SHM can be integrated with other data sources to address the last problem mentioned above, regarding not knowing the share of SH and DHW usage, as the SHM records them together. Chapter 3 specifically aims to tackle the challenge by developing a methodology to estimate the energy shares individually using weather data and machine learning.

Chapter 3. Hourly heating share estimation using smart heat meters

This chapter presents the second outcome of the Ph.D. project on using the smart heat meters' (SHM) data from residential buildings to attempt to estimate their hourly space heating (SH) and domestic hot water (DHW) energy shares.

In this chapter, it is proposed a novel methodology that applies machine learning to estimate the SH and DHW energy demand of residential buildings based on their SHM measurements. The present method was developed using a Danish measurements dataset, however, it was further validated with smaller datasets from Italy and Switzerland. Also, it investigated the impact that truncated (values rounded down to the nearest integer) measurements have on this methodology and how is it possible to offset such issues. This chapter also briefly discusses how the outcomes of this research have inspired further studies on the topic.

3.1 The need for heating demand disaggregation

The concept of energy demand disaggregation is not new for electrical-powered appliances, as a large investigative effort was made on this topic as it plays a crucial role in optimizing energy usage [36]. However, for heated water systems, the topic is much less researched, but it is arguably as important for enhancing energy efficiency in households with hydronic systems, e.g., radiators.

In **Paper 2** (section 3.2), a methodology is proposed to disaggregate the hourly SH and DHW energy demand from the total measurements of the SHM. This methodology relied only on weather data and it was proposed on the basis that the typical temporal resolution observed in SHM is hourly, which becomes a hurdle when estimating the DHW share [20].

Further in the investigation, it was observed that a usual preprocessing step performed by utility companies in Denmark is the truncation (rounding down) of the SHM measurements. Therefore, **Paper 3** investigates this issue in the same dataset as **Paper 2**, however with its former values truncated instead. Furthermore, **Paper 3** applied the method to smaller datasets from

Switzerland and Italy which are different building types than **Paper 2** for validation purposes [21].

From **Paper 3**, it was concluded that the truncation process undermines significantly the disaggregation methodology. Therefore, a methodology to "de-round" the measurements before applying the disaggregation is proposed by Schaffer *et al.* (2023) [37] and applied in **Paper 4**, where it is investigated if the new de-rounded values improve the disaggregation method [22].

3.2 SH and DHW disaggregation methodology

<u>Paper 2</u>

"A methodology to estimate space heating and domestic hot water energy demand profile in residential buildings from low-resolution heat meter data"

Daniel Leiria, Hicham Johra, Anna Marszal-Pomianowska, Michal Zbigniew Pomianowski, *Energy*, 2023, https://doi.org/10.1016/j.energy.2022.125705.

Energy 263 (2023) 125705

Contents lists available at ScienceDirect

Energy

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

A methodology to estimate space heating and domestic hot water energy demand profile in residential buildings from low-resolution heat meter data

Daniel Leiria *, Hicham Johra, Anna Marszal-Pomianowska, Michal Zbigniew Pomianowski

Aalborg University, Department of the Built Environment, Thomas Manns Vej 23, 9220, Aalborg Øst, Denmark

A R T I C L E I N F O

Keywords:
Smart energy meter
Data disaggregation
Building energy usage
Load profiles
Time series analysis
District heating

A B S T R A C T

This article presents a new methodology to disaggregate the energy demand for space heating (SH) and domestic hot water (DHW) production from single hourly smart heat meters installed in Denmark. The new approach is idealized to be easily applied to several building typologies without the necessity of in-depth knowledge regarding the dwellings and their occupants. This paper introduces, tests, and compares several algorithms to separate and estimate the SH and DHW demand. To validate the presented methodology, a dataset of 28 Danish apartments with detailed energy monitoring (separated SH and DHW usage) is used. The comparison shows that the best method to identify energy demand data points corresponding to DHW production events is the so-called "maximum peaks" approach. Furthermore, the best algorithm to estimate the SH and DHW separately is a combination of two methods: the Kalman filter and the Support Vector Regression (SVR). This new methodology outperforms the current Danish compliances typically used to estimate the annual DHW usage in residential buildings.

Credit author statement

***Daniel* Leiria:** Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Data curation; Writing - original draft; Writing - review & editing; Visualization. ***Hicham* Johra:** Conceptualization; Methodology; Resources; Writing - review & editing; Supervision. ***Anna* Marszal-Pomianowska:** Conceptualization; Resources; Writing - review & editing; Supervision. ***Michal* Zbigniew Pomianowski:** Conceptualization; Resources; Writing - review & editing; Supervision; Project administration; Funding acquisition. All authors have read and agreed to the published version of the manuscript.

1. Introduction

With the growing global concern regarding climate changes and the sustainability of our technologies, the different sectors of our society are challenged and urged to take a sharp turn to alleviate their impact on the environment. This is especially the case for energy production, distribution, and usage activities. Among them, the building sector has a major role in this sustainability transition. According to [1], the European Union (EU) building sector has an estimated share of 40% of the total energy end-usage, where 79% of it is for space heating (SH) and domestic hot water (DHW) production alone [2]. Specifically, in Denmark, 81.8% of the annual energy is used for heating (SH and DHW) in a typical house, while the other appliances (electrical consumers, lighting, etc.) have an annual share of 18.2% [3]. Regarding the Danish heating demand, 64% of the housing stock is connected to the district heating (DH) network. Furthermore, around 50% of the building stock in Iceland, Lithuania, Estonia, Sweden, Finland, Russia, Poland, and Northern China have their energy demand for space heating, cooling, and domestic hot water provided by district heating and cooling (DHC) networks [4]. The DHC systems and their potential for cost-effective, flexible, and sustainable heating and cooling supply are considered a strategic component of the roadmap toward a low-carbon future and gas-free neighborhoods in Europe, the USA, Canada, and Asia [5,6].

Research in the field of DH system improvement and integration of renewable energy sources leads to new DH concepts or configurations called "generations". Currently, the newly-installed and refurbished DH networks are transitioning from the 3rd to the 4th generation [7]. The 4th generation of district heating (4GDH) systems is mainly characterized by low-temperature heat-carrier fluid supply (40–70°C). The articles [7–10] outline several uprising advantages of implementing the 4GDH systems. Some of these advantages are the increase of energy efficiency in the network distribution due to the lowering of heat losses, a higher output capacity from different low-temperature sources

* Corresponding author.
E-mail address: dle@build.aau.dk (D. Leiria).

https://doi.org/10.1016/j.energy.2022.125705
Received 12 April 2022; Received in revised form 28 August 2022; Accepted 8 October 2022
Available online 12 October 2022

Nomenclature

Acronyms

4GDH	4th generation district heating
CPT	Change point temperature
DHW	Domestic hot water
DH	District heating
DHC	District heating and cooling
EU	European Union
HVAC	Heating, ventilation, and air conditioning
NA	Not available – missing data point
nZEB	Nearly zero-energy building
SH	Space heating
SMA	Simple moving average
SVR	Support vector regression
NMBE	Normalized mean bias error
CVRMSE	Coefficient of variation of the root mean square error

Symbols and variables

A	Heated area
C	Cost (SVR parameter)
$C_{p,water}$	Specific heat capacity of water
$E_{DHW,\ estim}$	Estimated domestic hot water energy usage
E_{DHW}	Measured domestic hot water energy usage
$E_{DHW,\ compl}$	Yearly estimated domestic hot water energy usage by the Danish compliance calculation
$E_{SH,estim}$	Estimated space heating energy usage
$E_{SH,max}$	Maximum measured space heating in the dataset
E_{SH}	Measured space heating energy usage
$E_{SH,min}$	Minimum measured space heating in the dataset
$\overline{E_{SH}}$	Mean measured space heating in the dataset
E_{Total}	Measured total heat demand for space heating and domestic hot water production
T_{cold}	Cold water supply temperature from Danish standards
T_{DHW}	Domestic hot water supply temperature from Danish standards
α	Moving average weight
γ	Gamma (SVR parameter)
ρ_{water}	Water density

integrated with DH systems, a smaller risk of pipe leakages caused by thermal stress, a better relation with the new building requirements regarding thermal usage, to name a few. Nevertheless, the 4GDH transition also faces particular challenges in decreasing the supply temperature needed for the building's SH and DHW demands. The challenges are the proper coordination in integrating the multiple low-temperature heat and waste-heat sources (renewable and recycled), the coupling to other energy grids (e.g., electricity, gas), the smart monitoring and control of such thermal grids and all its sub-components (including accurate prediction of production and demand, and demand-side management of the heat end users), the cost-effectiveness and the achievement of high reliability of heat supply at all time within a given (and often unflexible) legislative framework, and the operation of oversized or faulty systems on the building side.

It is clear from the barriers stated above that it is necessary to understand the DH network in detail. Therefore [7,11], outline the importance of smart meter data in the future of district heating. This metering initiative makes it possible: to efficiently manage the energy production, distribution grid, and the end-consumers; to optimize the DH system and its interconnection with other energy sources; to detect and fix the different faults occurring in the system; and to provide more information to the end-users regarding their energy usage, instigating them to change their consumption behavior.

As a front-runner, Denmark has made a great effort to install smart heat meters in buildings connected to the heating grid, and from 2027 it will be obligatory to collect dynamic heating data by using smart meters for every building connected to the DH grid [12]. These meters have up to 1-hour resolution measurements, and their collected data is easily accessible by utility companies. This metering initiative aims to obtain a detailed insight into the heat load patterns in each building and, when coupled with other sources of information, to unravel the reasons behind them. Even though this initiative is a significant step toward reaching the energetic goals set by Denmark [13,14] and the EU [15], it has a major drawback with respect to its data collection. In most buildings, only one smart energy meter is being installed per household. Each meter thus collects the total heat usage without distinguishing the energy used for SH or DHW production. Regarding SH, it depends on the outdoor conditions, building characteristics, occupants' preferences, and installed space heating systems [16]. In contrast, DHW production is correlated with people's consumption habits and the installed hot water production system. Because these two types of energy usage are associated with different variables, it is essential to estimate them separately to have a deeper insight into the building itself and its occupants [17].

Another aspect to consider on the importance of knowing these energy shares is regarding refurbishment initiatives. In [18], the authors argue that global building regulations have stricter SH efficiency rules while overlooking DHW consumption. Therefore, these new buildings, also known as low-energy buildings, have a much higher DHW share due to the continuous decrease of SH usage over the years and the higher levels of comfort concerning heating practices demanded by the residents.

Thus, a better assessment of the thermal appliances can be achieved by disaggregating the energy used in buildings. This contributes to a more detailed understanding and control on the user side and promotes better decision-making strategies regarding heat production and distribution.

1.1. Literature review

As mentioned above, most installed smart energy meters only measure the building's total heat usage. These total measurements often equal the sum of SH and DHW in a household. Even though this is already a great source of information, a clear distinction between SH and DHW production must be made. To tackle this problem, several research studies have developed different methods to estimate both utilities from total heating measurements. The present research focuses on instantaneous DHW production systems without thermal storage tanks, due to being a typical installation in Danish households, and all apartments in the dataset had this type of system. Hence most of the reviewed articles are regarding disaggregating methods applied in these systems.

One of the first studies to explore this problem is [19], which presents a statistical time-series approach to estimate the SH from the total heat usage measurements. The method assumes that the space heating demand varies smoother due to small outdoor temperature changes than the DHW usage, which, conversely, is more sporadic with higher peaks due to the very short time length of the different hot water draw-off events. This method estimates the SH by applying a kernel smoother to the total data points, where all measurements above a defined smoothed threshold are due to DHW usage. This method seems promising, and the authors formulated several kernel functions to increase the estimation accuracy. Nevertheless, it still lacks validation with separated space heating and DHW usage measurements, which the authors did not have at the time. Another drawback of this method is the necessity of high-resolution data (10-min measurements) to detect the sporadic peaks from DHW usage. Unfortunately, most of the installed smart meters in Denmark do not provide these high-frequency measurements.

Differently in [20], a simpler methodology is proposed to disaggregate the smart meters data by considering that the total measurements are equal to the DHW usage during summer, i.e., no SH demand. Based on this assumption, their approach does not estimate the different household heating utilities during the whole year but estimates the household average DHW load profile. If defined correctly, this type of profile provides valuable information concerning the customers' DHW habits. Regarding the method's accuracy, it is shown that it performs better for newly-built households with a large DHW usage share. However, the authors also concluded that several houses use space heating during summer, invalidating their initial assumption and significantly decreasing the profile accuracy. Similarly, in [21], a method is proposed to decompose SH and DHW usage in total measurements. The proposed method is called *hybrid summer signature.* It is based on discovering the DHW profiles when the total heating is equal to the DHW usage (no SH demand), taking into account the outdoor temperature. When the DHW profiles are discovered, the space heating demand equals the subtraction of the total values and the DHW daily profiles. The method was validated with several Norwegian buildings (apartments and hotels) and compared with other existing methods.

In [22], another approach is proposed to separate the different measurements in a Norwegian hotel. Two methods were presented and compared. Both approaches began by estimating the SH demand through its linear dependency on the outdoor temperature. The main difference between the methods is that the first calculates the DHW needs by subtracting the estimated SH from the total measured heat demand. And the second method, before calculating the DHW usage, the SH (already calculated by its outdoor temperature dependency) is adjusted by applying a singular spectrum analysis algorithm. The second methodology had the highest accuracy in predicting both heating utilities. With a different approach [23], estimates the SH and DHW usage weekly profiles using grey-box models. Their study concluded that the calculated values were slightly overestimated compared to the actual measurements. However, the method is accurate, and the authors argue that the models can be improved to increase even further its accuracy. The methodology developed in [24] is also worth mentioning. A pattern recognition algorithm was applied to disaggregate SH from other appliances in two households in the UK. Nevertheless, the household's heating source is a natural gas boiler instead of DH to provide thermal energy to SH, DHW, and cooking utilities (e.g., oven).

1.2. Contributions

Some of the methods developed to disaggregate the heating measurements are present in the section above. However, they have some drawbacks that this methodology attempts to solve. Firstly, this novel method aims to separate these energy shares using 1-h resolution measurements, which was proven by [19,23] to be extremely difficult and susceptible to inaccurate estimations. Another problem that the present methodology seeks to address is its non-dependence on other sources of information. Some of the reviewed methods require more information regarding the building (e.g., thermal envelope properties) and people (e.g., consumption habits) to proceed with SH and DHW estimation. This information is usually difficult to retrieve. Therefore, the proposed technique requires only the hourly total recorded heating values from the heat meters and the associated local weather data (outdoor temperature and global radiation). Lastly, the methodology algorithms were made simple and easy to implement and do not require any grey-box models' calibration.

Moreover, the contributions of this paper are:

1. The development of a new methodology to disaggregate SH and DHW from 1-h resolution total heating measurements. Besides, the method's algorithm is created to be easily implemented and only requires weather data as input.
2. The validation of the present methodology with a dataset of separated measurements of the different heating appliances from 28 Danish single-family apartments. All the apartments have an instantaneous DHW production system without a storage tank.
3. The comparison between DHW demand estimated through our disaggregation method and the current Danish annual DHW compliances. In order to assess the method's performance compared with the current calculation used in Denmark for the energy labeling in buildings.

1.3. Outline

Following this section, the developed methodology is described. The results from the method's validation are presented and discussed in section 3. The article closes with the main conclusions and suggestions for further work in sections 4 and 5. All the algorithms developed in this work are coded with the software *Rstudio* [25].

2. Methodology

2.1. Research roadmap

The method assumes that the SH system continuously operates during the heating season. At the same time, the DHW usage is expected to be produced sporadically throughout the day. Thus, during a day (which has 24 recorded data points – hourly resolution), only a few of these points will consist of collective SH and DHW production, whereas the other measurements will be SH usage alone. Every measurement identified with DHW production is converted to a missing point (NA point). Hence, each NA value is constituted by two energy shares, one for SH and another for DHW usage. Conversely, the non-NA values only have the SH share. Because the non-NA points in the dataset are the ones with SH usage alone, they are used to estimate the SH component of the NA points. The DHW usage in each NA point is calculated a posteriori through the difference between the total heat measurement from the smart meter and the estimated SH.

Based on these assumptions, several approaches have been developed to find the best procedure to separate and estimate the utilities' heating usage. In Fig. 1, one can see the research roadmap with the different studied approaches.

After the datasets are retrieved and pre-processed in step 1, different approaches to separate the data points are investigated in step 2. The energy separation stage identifies and labels all hours when the dwellers use DHW. In step 3, the points labeled as "not having DHW" (SH only) will be used to estimate the SH share of the points labeled with DHW and SH usage happening simultaneously. In step 4, the estimated values are compared to the actual separated measurements and the Danish DHW compliance calculations to test the methods' accuracy.

2.2. Dataset description and pre-processing

The dataset used in this study for validation is constituted of 28 apartments. All apartments are located in a social housing complex in Aalborg, Denmark. The complex was gradually renovated to the nearly Zero Energy Building (nZEB) standard from 2012 to 2020. The apartments included in this block were modernized in 2015. The concrete sandwich elements in the façade were replaced with insulated wooden cassettes with different façade cladding (i.e., brick wall, wood, or zinc). The roof construction was supported with new insolation. The heating, ventilation, and air conditioning (HVAC) installations were replaced with new ones. The interior of the apartments was fully renovated, and the new space heating installation includes radiators in all rooms and kitchens and underfloor heating in the bathrooms and hallways. The heat for SH and DHW is produced at the building block level and

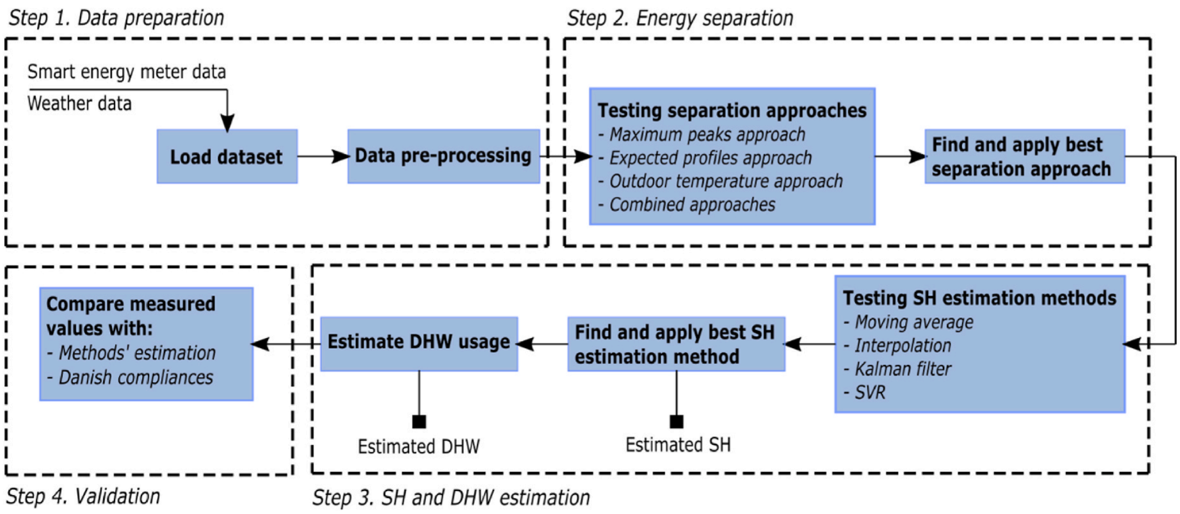


Fig. 1. Research roadmap.

distributed to each apartment. Apartments are equipped with individual SH and total heat demand meters (measuring SH and DHW without other appliances, e.g., electricity). The DHW is calculated through the difference between measurements from the meters. The floor area of the apartments is between 97 and 112 m^2.

The local weather data is extracted from the Danish Meteorologic Institute (DMI) website. The outdoor temperature and the global radiation were the only variables extracted with 1-h time resolution. The selected weather station is Tylstrup, as it is the closest station from Aalborg available in the DMI database.

In this work, the data pre-processing consisted in detecting the number of missing and negative measurements and removing them. In the 28 apartments dataset (187 123 data points), with approximately nine months of monitoring for each dwelling, there are 46 661 missing hours (~25% of the dataset). The apartment with the lowest missing data has approximately 3% missing data. On the other hand, some apartments have up to 43% of missing data. Regarding negative energy usage measurements (erroneous values), there are few dwellings with those. In total, these values only represent 0.013% of the dataset. Therefore, all missing measurements and erroneous values were removed from the dataset before its analysis.

2.3. Energy separation

In Denmark, the SH system generally operates continuously during the heating season, while the DHW is only produced sporadically throughout the day. Thus, only a few hours of the day correspond to the majority of the DHW usage, whereas the other data points are SH usage alone. To estimate the SH and DHW usage, it is thus necessary to identify which hourly measurement data points correspond to DHW and SH use from those only comprising SH demand. To that matter, five new approaches to identify these points are developed and investigated in this paper. All these methods are tested against ground truth data from the 28-apartment dataset in section 3.

2.3.1. Maximum peaks approach

This method starts from the premise that the outdoor temperature has small fluctuations during the day, contributing to smooth SH demand variations throughout its continuous daily operation. Considering this assumption, all meters' significant peaks in the measured heat can be accounted for DHW usage. Therefore, the "maximum peaks" algorithm detects all daily highest data points (E_{Total}) and considers them as comprising DHW production and SH ($E_{Total} = E_{SH} + E_{DHW}$). If a data point is not one of the maximum values, it is considered only SH usage ($E_{Total} = E_{SH}$). For each day, the method assumes the seven-highest measurements as DHW production, while the other 17 hourly data points are considered SH alone. It is also assumed a daily sleeping period from 1:00–4:00 h. Therefore, only SH operates during this period, and the high values are due to the low outdoor temperatures. In Fig. 2, one can see the algorithm's data flow diagram (a) and the representation of the method during a day for a single household (b).

After detecting all data points with DHW usage, they are converted into NA-values, and the household's dataset is updated with only SH measurements and the NA-values.

2.3.2. Expected profiles approach

This new method follows the same reasoning as the one used behind the "Maximum peaks approach". However, it is based on the hypothesis that weekdays have a certain regularity (i.e., routine) regarding the hot water usage pattern, as opposed to weekends. In this study, weekdays are considered from Monday to Thursday, while weekends are considered from Friday to Sunday. The reason for this division is that it is expected that a larger variation of hot domestic water usage occurs on Fridays afternoon and evening. This reasoning is also corroborated in [23], where Fridays were considered a different profile from the other weekdays. Therefore, from Monday to Thursday, the daily profile was separated into three groups – morning (5:00–11:00 h), afternoon (12:00–16:00 h), and evening (17:00–00:00 h). The highest value is found and considered as being "SH + DHW" in each time range. During the morning and evening periods, the adjacent hours (−1 and +1 h) of the peak heating usage are also identified as "SH + DHW". Concerning periods spanning from Fridays to Sundays, the "Maximum peaks approach" is used to detect the "SH + DHW" points because it is not likely to follow a routine. In Fig. 3, one can see the algorithm's data flow diagram (a) and the representation of the method during a day for a single household (b).

After detecting all data points with DHW usage, they are converted into NA-values, and the household's dataset is updated with only SH measurements and the NA-values.

2.3.3. Outdoor temperature approach

It is known that building SH needs have a strong negative linear correlation with outdoor temperature during the heating season [26]. However, if this trend with the outdoor temperature is not observable with the total energy measurements, it is due to significant DHW production events.

As illustrated in Fig. 4, this method starts by subsetting each household's dataset with only the total measured values from 1:00–4:00

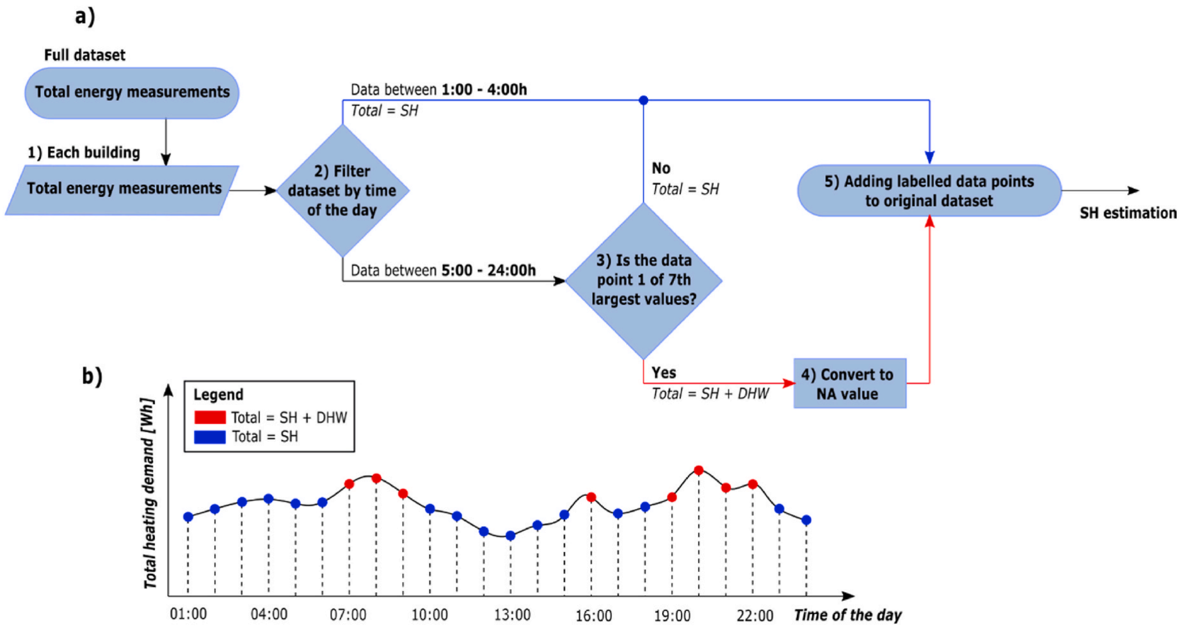


Fig. 2. a) Data flow diagram: Maximum peaks approach; b) Method representation.

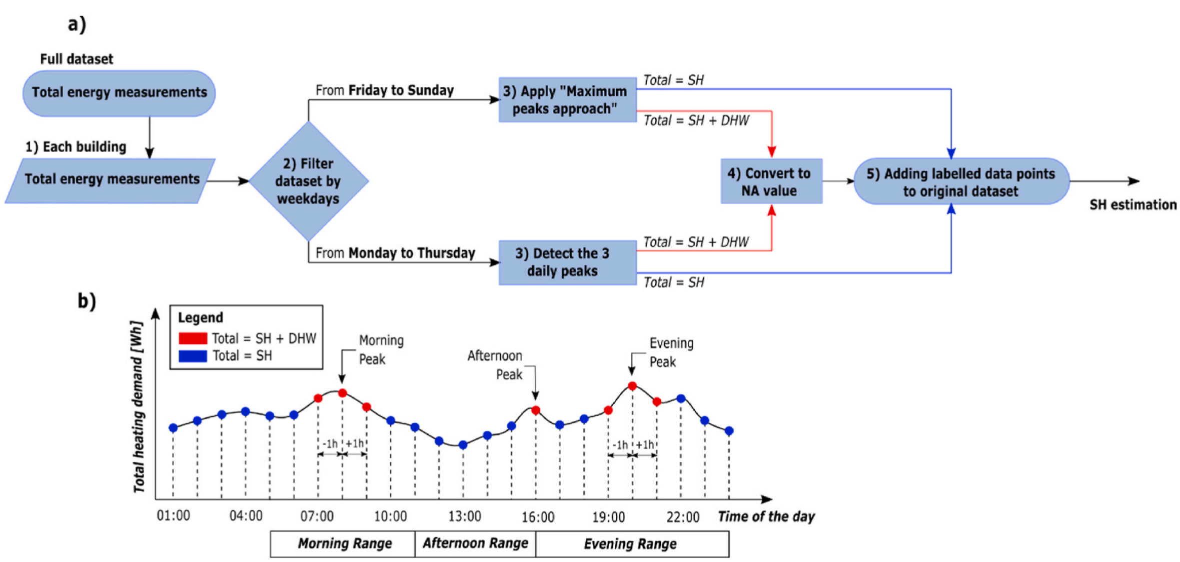


Fig. 3. a) Data flow diagram: Expected profiles approach; b) Method representation.

h (step 2). This time-conditioned subset is used because it is assumed that during this period, people are asleep. Thus, all total measurements must be SH usage only. With the subset, it is generated a piecewise linear regression [27] that estimates the SH demand as a function of the outdoor temperature for both the heating and no-heating seasons (commonly known as the “heat demand signature curve”). The junction between the negative linear trend of the heating season and the linear constant (horizontal) trend of the no-heating season forms the “change point temperature” (CPT) [28] (step 3).

In step 4, a prediction interval is developed for the two seasons. For the heating season, the interval’s tolerance is iteratively defined as 0.90, and for the no-heating period, a narrower tolerance of 0.60. By establishing the prediction intervals, the building’s dataset is divided into data points that are positioned above the intervals (step 5). If the measurement is below the interval, it follows the SH trend and therefore is SH usage ($E_{Total} = E_{SH}$). If the value is outside the interval, the total energy equals SH and DHW simultaneously ($E_{Total} = E_{SH} + E_{DHW}$). All points with DHW usage are converted into NA-values in step 6. The last step is updating the building’s dataset with only SH measurements and the NA-values. In Fig. 4, one can see the approach’s representation.

2.3.4. Combined approaches

The combined method merges the separation techniques described in subsections 2.3.1 and 2.3.3. Two different combined approaches were developed. “Combined method 1” only categorizes a data point as “SH + DHW” if both approaches, “Maximum peaks” and “Outdoor temperature”, together label the same point as “SH + DHW”. This method’s data flow diagram can be seen in Fig. 5a. The “Combined method 2” categorizes a data point according to its measured total heating usage. If the total energy of the datapoint is lower than 250 Wh or higher than 3,000 Wh, then the “Outdoor temperature approach” is used. If not, the “Maximum peaks approach” is used. These threshold values are established due to a preliminary investigation of the performance of the “Outdoor temperature” approach for one of the apartment’s data. This

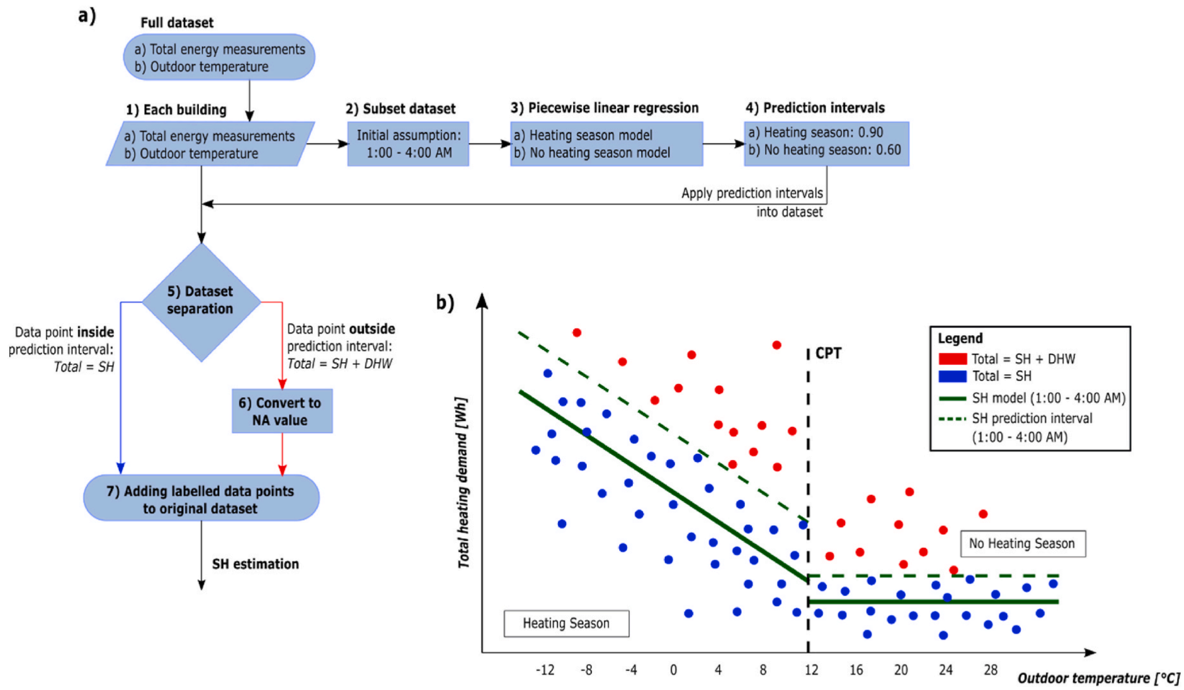


Fig. 4. a) Data flow diagram: Outdoor temperature approach; b) Method representation.

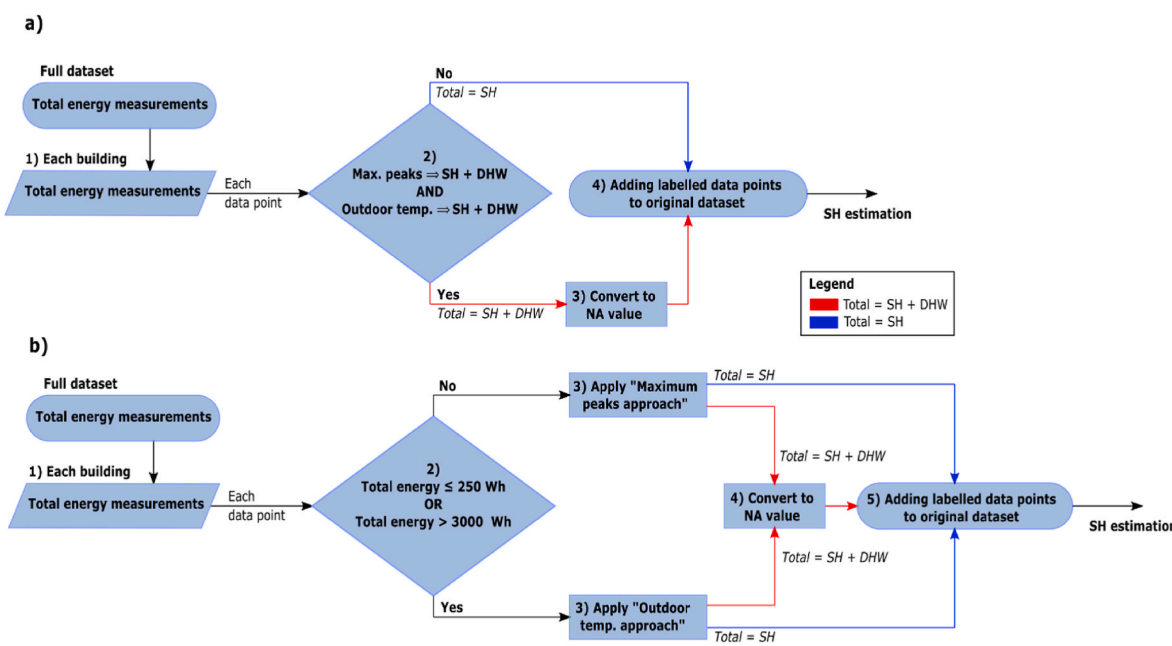


Fig. 5. a) Data flow diagram: Combined 1 approach; b) Data flow diagram: Combined 2 approach.

preliminary test showed that the “Outdoor temperature” method performed better for total heat data points below 250 Wh and above 3,000 Wh. It is advised for each building case to perform a preliminary calculation to establish these heating thresholds accurately because they might differ for each building. The second merged method is presented in Fig. 5b.

Like other approaches, after detecting all data points with DHW usage, they are converted into NA-values, and the household’s dataset is updated with only SH measurements and the NA-values.

2.4. Space heating and DHW estimation

At this stage, the DH dataset consists of NA-values and measurements that only quantify SH ($E_{Total} = E_{SH}$). The next step is to estimate the SH usage ($E_{SH,estim}$) in the NA-values by considering the known SH data points (E_{SH}). After obtaining the $E_{SH,estim}$ of the NA-points, the DHW usage is calculated as $E_{DHW,\ estim} = E_{Total} - E_{SH,estim}$. To calculate the SH demand in the missing points, several methods have been implemented and benchmarked hereafter.

2.4.1. Interpolation – Univariable estimator

Firstly, linear interpolation, cubic spline interpolation, and Stineman interpolation are tested. The linear method calculates the NA-value(s) by assuming a linear relationship between its known neighboring points. To estimate the missing values, the cubic spline method fits a third-order polynomial between the known SH data points. The Stineman interpolation also applies a third-order polynomial into the time series; however, it preserves its monotonicity. These estimation algorithms are derived from the R-package *imputeTS* [29].

2.4.2. Moving average – Univariable estimator

This method is one of the most commonly used in data analysis for smoothing time series. It consists in averaging the values with their neighboring points. The width of neighboring points used to calculate this average is designated as the "window". This window-size variable, or range, must be set beforehand. A range equal to 2 (k = 2) has been selected in this study, which means that any NA-value is estimated by averaging its two previous and two succeeding points. Different weight-averaging techniques are also tested. These techniques are the simple moving average, linear weighted moving average, and exponential weighted moving average.

When estimating the missing points, the simple moving average has equal-weighted neighboring points ($\alpha = 1/4$). The linear weighted moving average follows an arithmetic progression by decreasing the weight of the furthest neighboring points. For a window-size of 2, the previous and succeeding points to the NA-value weigh 0.5 ($\alpha = 1/2$), and the second previous (and succeeding) points weigh 0.33 ($\alpha = 1/3$). For an exponential weighted moving average, the nearest points weigh 0.5 ($\alpha = 1/2^1$), followed by a weight of 0.25 ($\alpha = 1/2^2 = 1/4$). These estimation algorithms are derived from the R-package *imputeTS* [29].

2.4.3. Kalman filtering – Univariable estimator

A Kalman filter is tested with four different model implementations: a structural time series model with and without smoothing and an ARIMA (Autoregressive Integrated Moving Average) model with and without smoothing.

The structural time series model is based on the function "StructTS", which consists of a linear Gaussian state-space model for univariate time series. The ARIMA model is from the function "auto.arima", which finds the best ARIMA model for each building's time series. Both models are tested with and without smoothing. These estimation algorithms are derived from the R-package *imputeTS* [29].

2.4.4. Support vector regression (SVR) – multivariable estimator

Contrary to the aforementioned methods, this estimation technique considers other inputs to calculate SH missing data points. The support vector regression (SVR) is a machine learning method that trains a model with the values labeled as "SH only". The input data to estimate a given SH point is the outdoor temperature, the global solar radiation measured two and one hours prior, and the SH + DHW points (smart meter measurements) before and after the missing point. The SVR model uses a radial kernel function with the parameters C (cost) and γ (gamma) equal to 7 and 0.01, respectively. This estimation algorithm is derived from the R-package *e1071* [30,31].

2.4.5. Combined Kalman filtering and SVR – Univariable/multivariable estimator

From preliminary results, the Kalman smoothing techniques are the best methods to predict space heating from the total heat use. However, as explained, these methods depend on the neighboring data points, which can also be missing in some cases (missing data gap larger than 1 h). To tackle the problem, this algorithm is refined to use the smoothed Kalman filter with the model "StructTS" only when the number of hours missing consecutively is equal to or below 2 (Gap ≤2). If the data gap is larger, the SVR is applied instead with the same parameters described above. These estimation algorithms are derived from the R-packages *imputeTS* and *e1071* [29–31]. One can see in Table 1 all the tested estimation methods and their parameters.

Table 1
Estimation methods.

Method	Parameters	Type	Input	Reference
Interpolation	Linear Cubic spline Stineman	Univariable	E_{Total}	[29]
Moving average	Simple, *k = 2* Linear, *k = 2* Exponential, *k = 2*	Univariable	E_{Total}	[29]
Kalman filtering	Model: *StructTS* Smoothing: *True* Model: *StructTS* Smoothing: *False* Model: *auto.arima* Smoothing: *True* Model: *auto.arima* Smoothing: *False*	Univariable	E_{Total}	[29]
Support vector regression (SVR)	Kernel: *Radial* Cost: *7* Gamma: *0.01*	Multivariable	T_{out} *[i-1, i-2]* *Rad [i-1]* E_{Total} *[i-1, i+1]*	[30,31]
Kalman filtering & SVR	**Gap ≤ 2 h** Model: *StructTS* Smoothing: *True* **Gap > 2 h** Kernel: *Radial* Cost: *7* Gamma: *0.01*	Multivariable	E_{Total} T_{out} *[i-1, i-2]* *Rad [i-1]* E_{Total} *[i-1, i+1]*	[29–31]

With the application of some of these methods, the estimated space heating ($E_{SH,estim}$) can be negative or higher than the total energy measurements. Therefore, if $E_{SH,estim}$ is negative, it is set to zero; and if $E_{SH,estim}$ is larger than E_{Total} (SH + DHW – Smart meter's measurements), it is set to E_{Total}.

2.5. Methodology validation

To benchmark the accuracy of these different estimation methods, two different comparison metrics are computed: the normalized mean bias error (NMBE) and the coefficient of variation of the root mean square error (CVRMSE). These metrics are commonly used to assess numerical models' performance (accuracy) in the energy and building systems field. They can evaluate the distance between the output time series of a numerical simulation and a reference time series [32,33]. The NMBE is given as a percentage (see Equation (1)) and measures the global bias of the estimation methods. If the value is negative, the method is globally underpredicting and overpredicting if positive.

$$NMBE = \frac{\sum_{i=1}^{n}\left(E_{SH,estim}[i] - E_{SH}[i]\right)}{n} \times \frac{100}{E_{SH,max} - E_{SH,min}} \tag{1}$$

The CVRMSE is also given as a percentage and estimates the point-to-point difference between the measurements (ground truth) and estimated values (see Equation (2)).

$$CVRMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(E_{SH,estim}[i] - E_{SH}[i]\right)^2}{n}} \times \frac{100}{\overline{\overline{E_{SH}}}} \tag{2}$$

Where:

NMBE: Normalized mean bias error [%]

CVRMSE: Coefficient of variation of the root mean square error [%]

$E_{SH,estim}[i]$: Estimated space heating [kWh]

$E_{SH}\ [i]$: Measured space heating [kWh]

$E_{SH,max}$: Maximum measured space heating in the dataset [kWh]

$E_{SH,min}$: Minimum measured space heating in the dataset [kWh]

$\overline{E_{SH}}$: Mean measured space heating in the dataset [kWh]

n: Number of measurements in the dataset [−]

After selecting the best estimation method to obtain the SH demand using the above metrics (Equations (1) and (2)), the method is applied to calculate the SH in all apartments and predict the DHW need ($E_{DHW,estim} = E_{Total} - E_{SH,estim}$). The DHW estimated demand is finally compared with the actual DHW measurements and the Danish compliance calculations to investigate if the developed methodology outperforms the current Danish calculations in predicting the DHW household needs.

In Denmark, the DHW consumption in households is currently predicted using the compliance calculation of 250 L/m^2 per year [34]. Similarly, the inlet water (cold) and outlet water (DHW) temperatures are considered to be 10 °C and 55 °C, respectively [35]. By knowing the area of the different apartments, the yearly energy usage for DHW production is calculated through Equation (3):

$$E_{DHW,compl} = \frac{1}{3600} \times 0.25A \times \rho_{water} c_{p,water} \times (T_{DHW} - T_{cold}) \tag{3}$$

Where:

$E_{DHW,compl}$: Estimated DHW energy usage from Danish compliances [kWh/year]

$0.25A$: 0.25 m^3 water volume per m^2 of heated area per year [m^3/year]

$\rho_{water} c_{p,water}$: Water density per water-specific heat capacity – Constant value: 4177 [kJ/m^3°C]

T_{DHW}: DHW supply temperature from Danish standards – Constant value: 55 [°C]

T_{cold}: Cold water supply temperature from Danish standards – Constant value: 10 [°C]

3. Results and discussion

This section presents the results of applying the different methods and their validation to find the best methodology. Moreover, the estimated DHW usage from the best methodology is compared with the current Danish compliance, which estimates the yearly DHW production.

3.1. Energy demand separation

The five DHW separation methods presented herebefore are tested against measurements (ground truth) from 28 apartments in Denmark that have separated metering of SH and DHW energy usage. The validation consists in assessing the identification accuracy of the different approaches.

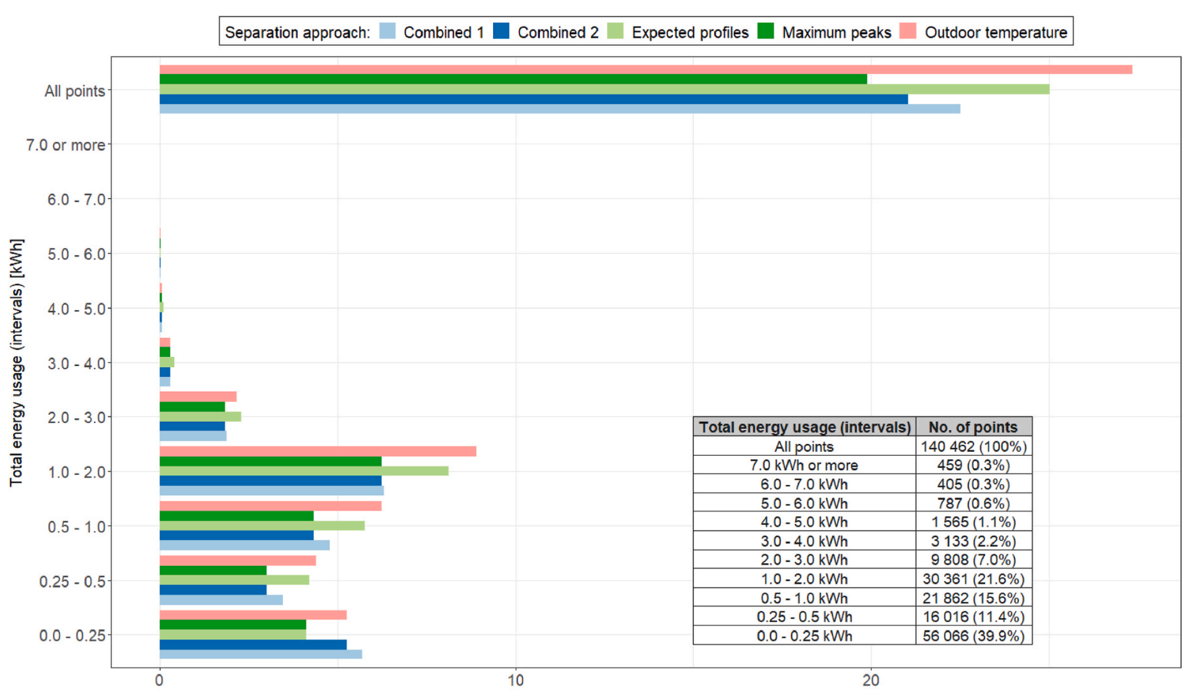
In Fig. 6, one can see the total percentage of incorrectly identified points in all apartments. This percentage is divided into total heating intervals (measured by the smart meters) to see if the methods perform better at different energy demand levels.

In Fig. 6, one can observe that "maximum peaks" and "combined 2" approaches are the best for categorization, with 20% incorrectly identified in all apartments. The method with the highest inaccuracy is the "outdoor temperature" approach, with a value of 27%. It is also seen that for different heating intervals, some approaches performed better than others. However, such differences are too small to conclude that the measured heating intensity affects the approach's performance.

The quantity of correctly and incorrectly labeled (identified) points per approach was also analyzed without dividing by measured energy levels. The explanation of these attributed labels and how they affect the methodology are in Table 2, and its results are in Fig. 7.

One can conclude from the results presented in Fig. 7 that the methods "combined 2" and "maximum peaks" have a similar identification performance. However, the "maximum peaks" approach is preferred as the best separation algorithm from these results because the "combined 2" method is rooted in the "outdoor temperature" approach, which has the largest percentage of incorrectly identified points.

To conclude, the incorrect identified points percentage of each separation approach is calculated for each apartment. This analysis, in



Total energy usage (intervals)	No. of points
All points	140 462 (100%)
7.0 kWh or more	459 (0.3%)
6.0 - 7.0 kWh	405 (0.3%)
5.0 - 6.0 kWh	787 (0.6%)
4.0 - 5.0 kWh	1 565 (1.1%)
3.0 - 4.0 kWh	3 133 (2.2%)
2.0 - 3.0 kWh	9 808 (7.0%)
1.0 - 2.0 kWh	30 361 (21.6%)
0.5 - 1.0 kWh	21 862 (15.6%)
0.25 - 0.5 kWh	16 016 (11.4%)
0.0 - 0.25 kWh	56 066 (39.9%)

Fig. 6. Incorrectly identified points percentage in the overall dataset for each separation approach.

Table 2
Labeling types.

Case	Correct label	Attributed label	Type	Explanation
$E_{DHW}\neq0$	"SH + DHW"	"SH + DHW"	Correct	Points correctly identified and converted into NA-values to be estimated by the SH models
		"Only SH"	Incorrect	Incorrectly identified points that will be used to train the SH models – Will negatively affect the model's training
$E_{DHW}=0$	"Only SH"	"SH + DHW"	Incorrect	Incorrectly identified points that will be converted into NAs and estimated by the models – Will affect the models' accuracy negatively
		"Only SH"	Correct	Correctly identified points that will be used to train the SH models

Fig. 8, is made to understand if the different apartments influence the overall performance of the different methods.

In Fig. 8, one can see the overall incorrect percentage of identified points (x-axis) for each apartment per separation method (y-axis/colored legend). From the figure, it is possible to observe the percentage distribution and extreme cases.

The results show that the different methods have their inaccuracy distributions between 10% and 40%, see Fig. 8. The "outdoor temperature" approach underperforms the most. One can observe that the best approaches are the "combined 2" and "maximum peaks", with a slightly smaller difference in the mean value in the latter. Based on the analysis and application of the methods on the 28 apartments dataset, the preferred method to disaggregate the DH dataset is the "maximum peaks" approach.

3.2. Space heating and DHW estimation

After separating the data points, the following step estimates the SH based on data assumed to be only SH usage. Several methods are tested to determine the most accurate one for this specific application. One can see in Table 3 the NMBE and CVRMSE calculated for each estimation method for the whole dataset (28 apartments).

The results show that most methods have similar values in both metrics, which means that methods differ slightly from each other. The worst-performing method is the cubic spline interpolation, indicating that cubic polynomial is not the best mathematical function to estimate space heating. The best method is the combined Kalman filtering and SVR according to both metrics.

In Fig. 9, one can see the overall error between the estimation and the measurements of SH and DHW of the different apartments. The overall error is calculated by comparing the difference between the aggregated measurements and estimated values during the measurement period.

As one can see from Fig. 9, the overall SH error (green color) is primarily negative (underestimated), with 18 apartments between -10% and 0%. Furthermore, the households with the extreme error values are one apartment with less than -15% error and another with almost +50% error (overpredicted).

Regarding the DHW prediction (blue color), the error distribution is wider than the space heating. In this case, five apartments have an overestimated DHW demand above +25%. The extreme DHW prediction is one household with an overestimation of +85% and four apartments with an underestimation slightly higher than -10%.

Several factors influence the method's estimations and the overall error. Foremost, the separation approach inaccurately identifies some of the points, influencing from the beginning, the estimation accuracy. Another factor is the presence of missing values in the initial dataset. As one can see in section 2.1, the dataset comprises about 25% of missing measurement points. Because the estimation relies on determining the SH demand based on its neighboring points, several missing measurements negatively impact the overall method's performance. Moreover, in section 2.1, it is shown that the retrieved weather data is not at the exact location where the dwellings are located. Besides this, the possibility of different heating systems, a large SH share, the unique dwellers' routines, or the DHW share being equal to zero (no occupancy) may influence the method's performance, and they might be the reasons behind the extreme cases.

The present research also compares the estimated DHW values with

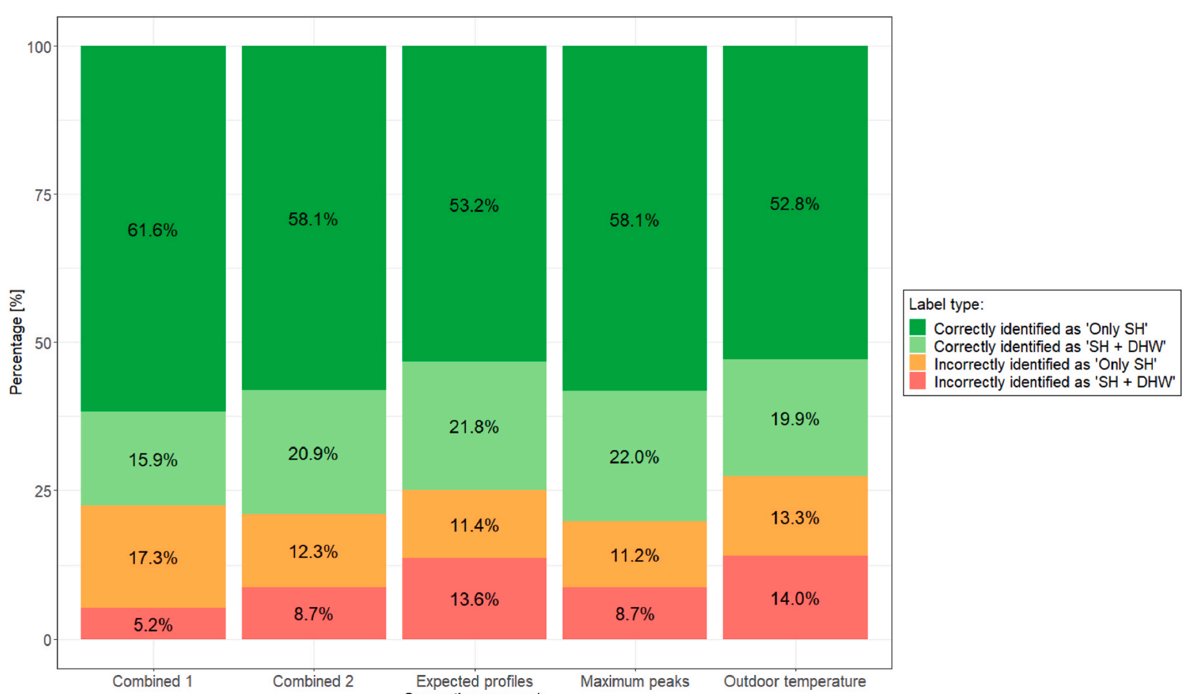


Fig. 7. Attributed labels percentage in the overall dataset for each separation approach.

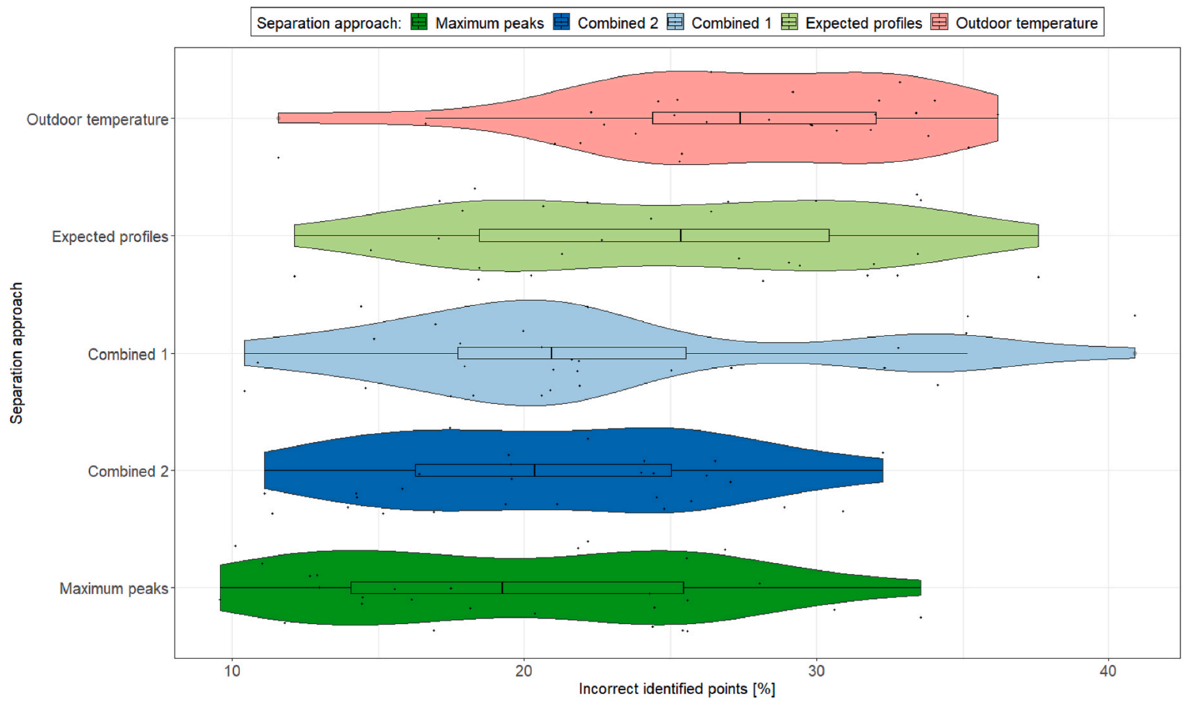


Fig. 8. Incorrectly identified percentage of separation approaches for each apartment (each point is one apartment).

Table 3
Each SH estimation method's NMBE and CVRMSE for all apartments.

Method	Type	Method specifications	NMBE	CVRMSE
Interpolation	Univariable	Linear	−0.25%	54.67%
		Cubic spline	−0.24%	59.60%
		Stineman	−0.25%	54.96%
Moving average	Univariable	Simple	−0.26%	54.62%
		Linear	−0.26%	54.60%
		Exponential	−0.26%	54.67%
Kalman filtering	Univariable	StructTS – Smoothed	−0.27%	53.86%
		StructTS – No smoothed	−0.28%	55.29%
		ARIMA – Smoothed	−0.25%	54.16%
		ARIMA – No smoothed	−0.24%	56.26%
Support vector regression (SVR)	Multivariable	-	−0.18%	54.46%
Kalman filtering & SVR	Multivariable	StructTS – Smoothed -	−0.10%	52.49%

the Danish compliance calculation used to predict the annual DHW demand in households. The results of this comparison are in Table 4.

As shown in Table 4, there are three types of values per DHW usage. The actual DHW demand (E_{DHW}), the compliance calculation of DHW demand used in Denmark ($E_{DHW,\ compl}$), and the estimated DHW from the developed methodology ($E_{DHW,\ estim}$). The "average" values are the aggregated DHW usage divided by the number of data points (hours). For the case of the DHW measurements and estimation, the number of data points is the number of measurement hours in each apartment. For the compliance case, the number of data points is the number of hours in a year. The "average" values are determined to be able to compare all three DHW usage types and calculate the error between the actual measurements and the compliance/estimation values. In most apartments, the developed methodology outperforms (bold values) the current Danish compliance calculations. Even though the disaggregation method has a good performance in estimating the DHW usage for most apartments, there are few cases where the error is significant. The reason behind it might be due to numerous measurement hours missing in the initial dataset or due to the lack of dwellers in the households during the measurement period. However, from the results, it is argued that the method can be applied to predict the household's DHW energy use instead of what has been used to make the dwelling's energy assessment in Denmark. Also, it is clear that basing the Danish DHW compliance calculations only on the building area is imprecise; hence the research must shift towards the occupancy number and its behavior.

Even though some apartments have large SH and DHW estimation errors, this data-driven methodology is quite appreciable when considering its simplicity and the fact that no detailed building information, often unknown, is required (e.g., people habits description, system identification, building envelope characteristics, etc.). Another advantage of this method over some of the existing ones reported in the literature is the possibility of using hourly measurement data. Finally, the method outperforms the floor area-based compliance method currently used in Denmark for estimating DHW production.

4. Conclusion

This article introduces a new data-driven methodology to estimate the SH and DHW from low-resolution heat meter data. The method's novelty is the possibility of applying it to hourly heating measurements without in-depth knowledge of the building and its occupants. The developed method is the combination of two algorithms to i) identify from the total heat measurements the points with DHW production and ii) estimate from the identified DHW usage points, the SH and DHW usage. This research tested several alternative methods for both algorithms to find the best point separation and energy estimation techniques. The different methods can be seen in Table 5:

The validation process shows that the best-performing method to detect when the DHW is being used is the "maximum peaks" approach, with a successful identification rate of approximately 80%. The best algorithm to estimate the SH demand in the identified points is the combination of SVR and Kalman filters (smoothed "StructTS" model). This estimation method has an NMBE of −0.10% and CVRMSE of

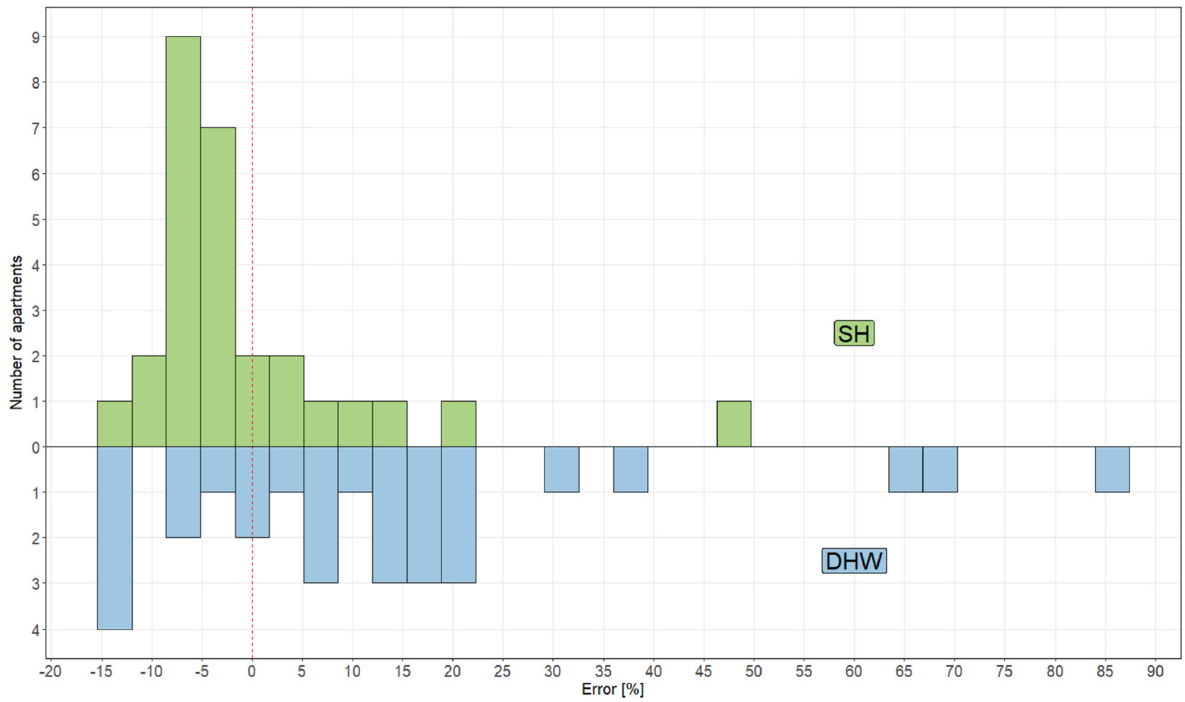


Fig. 9. Overall error of each apartment's SH and DHW estimation.

Table 4
Comparison between the Danish compliance values and the estimation results. The bold error values indicate the best performing method between the novel approach developed in this study and the Danish compliance calculations.

Apartment ID	Area [m^2]	E_{DHW} [kWh/h]	$E_{DHW,\ compl}$ [kWh/h]	$E_{DHW,\ estim}$ [kWh/h]	Error between E_{DHW} and $E_{DHW,\ compl}$	Error between E_{DHW} and $E_{DHW,\ estim}$
666	112	0.314	0.167	0.315	−47%	**0%**
668	111	0.286	0.165	0.346	−42%	**21%**
669	110	0.184	0.164	0.224	**−11%**	22%
670	111	0.588	0.165	0.555	−72%	**−6%**
671	110	0.247	0.164	0.295	−34%	**20%**
697	111	0.692	0.165	0.606	−76%	**−12%**
698	111	0.674	0.165	0.627	−75%	**−7%**
699	110	0.678	0.164	0.588	−76%	**−13%**
700	111	0.074	0.165	0.137	123%	**85%**
701	111	0.167	0.165	0.196	**−1%**	18%
702	110	0.088	0.164	0.115	87%	**32%**
724	110	0.229	0.164	0.255	−28%	**11%**
726	111	0.116	0.165	0.132	43%	**14%**
727	111	0.103	0.165	0.121	61%	**18%**
728	110	0.148	0.164	0.203	**11%**	37%
729	110	0.144	0.164	0.161	14%	**12%**
730	111	0.388	0.165	0.406	−57%	**5%**
731	111	0.087	0.165	0.142	90%	**63%**
732	110	0.406	0.164	0.347	−60%	**−15%**
734	97	0.091	0.145	0.106	59%	**17%**
735	111	0.328	0.165	0.347	−50%	**6%**
736	111	0.336	0.165	0.34	−51%	**1%**
739	111	0.524	0.165	0.561	−68%	**7%**
740	111	0.164	0.165	0.159	**1%**	−3%
741	111	0.237	0.165	0.253	−30%	**7%**
742	97	0.145	0.145	0.167	**−1%**	15%
743	111	0.461	0.165	0.403	−64%	**−13%**
745	111	0.093	0.165	0.157	78%	**69%**

52.49%, being the lowest metric values of all tested SH estimation algorithms. Therefore the chosen overall method to disaggregate SH and DHW demand from the total heat measurements is the "maximum peaks" approach for identification purposes and the combined methods of SVR and Kalman filter to estimate SH needs.

Table 5
List of tested methods in this study.

i) Identification/separation methods	ii) SH estimation methods
Maximum peaks	Interpolation
Expected profiles	Moving average
Outdoor temperature	Kalman filter
Combined 1	Support vector regression (SVR)
Combined 2	Kalman filter & SVR (combined methods)

The overall methodology predicts the SH demand with an error between −10% and 10% for most dwellings. Concerning DHW estimation, the error is slightly wider, with most apartments falling between −15% and 15%. Moreover, this study compared the estimated DHW demand from the method with the actual measurements and the current Danish DHW compliance calculations. This comparison concludes that the developed methodology outperforms the Danish compliance calculations in most cases. Furthermore, it is argued that this disaggregation method can be applied to predict the household's SH and DHW energy shares. The authors also argue that estimating the DHW energy usage by relying solely on the building's area is erroneous (currently being done in Denmark and other European countries). Thus, future research efforts must move toward estimating the heating usage in buildings considering the dwellers' number and more specific building typology regarding DHW use (currently, in Denmark, only two are present: residential and other).

Finally, this data-driven method is simple to compute and understand, and if validated with more building cases and proved to be robust, it can be applied in the future by DH companies and energy auditors. Also, this methodology can be used without having additional detailed information about the building and its dwellers and can be used with 1-h resolution data, which is often the status of the buildings and their metering installations. The authors argue that this method is relevant to the energy and buildings field when considering these advantages, more specifically for the analysis of the energy performance gap, the DHW usage assessment (which has been overlooked until recent years), clustering of different SH usage patterns according to their systems and user's practices, and energy-efficiency decision-making.

5. Further work

A suggestion for further work is the application of this methodology with other datasets for further validation and robustness analysis. Preferably, datasets should come from various countries to ensure the methodology's robustness and applicability in different cases. This study used several algorithms to estimate space heating (e.g., SVR, moving average, etc.). However, this work can be further developed by investigating other estimation methodologies that can be found in the literature (e.g., neural networks, random forest regression, etc.).

It is also suggested to benchmark this novel methodology with other existing disaggregation methods on a common dataset. Furthermore, a more extensive effort must be made to collect good quality datasets – with hourly resolution (or higher) – of separated energy usage for space heating and domestic hot water in buildings with instantaneous hot water production systems.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 893945 (E-DYCE). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 958345 (PRELUDE).

Declaration of competing interest

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

Data availability

The authors are unable or have chosen not to specify which data has been used.

Acknowledgments

The authors would like to acknowledge the Himmerland housing association and Varmekontrol A/S for their support in data collection and to all partners involved in the E-DYCE and PRELUDE projects. Also, the first author would like to express his gratitude to his colleagues (Kamilla H. Andersen, Markus Schaffer, Martin Frandsen, Martin Veit, and Simon P. Melgaard) in the Ph.D. "Corner Office" for the valuable discussions and support.

References

[1] European Commission. *Energy performance of buildings directive* [Online]. 2022. Available: https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en.
[2] European Commission. *Heating and cooling* [Online]. 2022. . Available: https://energy.ec.europa.eu/topics/energy-efficiency/heating-and-cooling_en.
[3] Danish Energy Agency. "Energy Statistics 2020," 2020. [Online]. Available: https://ens.dk/sites/ens.dk/files/Statistik/energy_statistics_2020.pdf.
[4] Werner S. International review of district heating and cooling. Energy 2017;137:617–31. https://doi.org/10.1016/j.energy.2017.04.045. Oct. 15.
[5] IRENA. Renewable Energy Policies in a Time of Transition. 953; 2018.
[6] Paardekooper S, et al. Heat roadmap Europe 4: quantifying the impact of low-carbon heating and cooling roadmaps. Aalborg: Aalborg Universitetsforlag; 2018.
[7] Lund H, et al. "4th Generation District Heating (4GDH). Integrating smart thermal grids into future sustainable energy systems.,". Energy 2014;68:1–11. https://doi.org/10.1016/j.energy.2014.02.089. Apr. 15.
[8] Lund H, et al. Perspectives on fourth and fifth generation district heating. energy 2021;227. https://doi.org/10.1016/j.energy.2021.120520. 120520, Jul.
[9] Dalla Rosa A, et al. General rights IEA DHC annex X report: toward 4th generation district heating experience and potential of low-temperature district heating. 2014.
[10] Li H, Nord N. Transition to the 4th generation district heating - Possibilities, bottlenecks, and challenges. Energy Proc Sep. 2018;149:483–98. https://doi.org/10.1016/j.egypro.2018.08.213.
[11] Volkova A, Mašatin V, Siirde A. Methodology for evaluating the transition process dynamics towards 4th generation district heating networks. energy May 2018;150:253–61. https://doi.org/10.1016/j.energy.2018.02.123.
[12] European Parliament, the Council. Directive (EU) 2018/2002 of the European parliament and of the council of 11 December 2018 amending directive 2012/27/EU on energy efficiency. Off J Eur Union Dec. 2018;328:210–30.
[13] The Danish Government. A Green and Sustainable World - the Danish Government's long-term strategy for global climate action. Oct. 2020.
[14] Energistyrelsen. Danish climate policies [Online]. 2022. Available: https://ens.dk/en/our-responsibilities/energy-climate-politics/danish-climate-policies.
[15] European Commission. Communication from the commission to the European parliament. Brussels: the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions — The European Green Deal,"; 2019.
[16] Do Carmo CMR, Christensen TH. Cluster analysis of residential heat load profiles and the role of technical and household characteristics. Energy Build Aug. 2016;125:171–80. https://doi.org/10.1016/j.enbuild.2016.04.079.
[17] Gram-Hanssen K. "New needs for better understanding of household's energy consumption - behaviour, lifestyle or practices?,". Architect Eng Des Manag 2014;10(1–2):91–107. https://doi.org/10.1080/17452007.2013.837251. Apr.
[18] Pomianowski MZ, Johra H, Marszal-Pomianowska A, Zhang C. Sustainable and energy-efficient domestic hot water systems: a review. Renew Sustain Energy Rev 2020;128(1):109900. https://doi.org/10.1016/j.rser.2020.109900.
[19] Bacher P, de Saint-Aubain PA, Christiansen LE, Madsen H. Non-parametric method for separating domestic hot water heating spikes and space heating. Energy Build Oct. 2016;130:107–12. https://doi.org/10.1016/j.enbuild.2016.08.037.
[20] Marszal-Pomianowska A, Zhang C, Pomianowski M, Heiselberg P, Gram-Hanssen K, Hansen AR. Simple methodology to estimate the mean hourly and the daily profiles of domestic hot water demand from hourly total heating readings. Energy Build Feb. 2019;184:53–64. https://doi.org/10.1016/j.enbuild.2018.11.035.
[21] Lien S, Ivanko D, Sartori I. Domestic hot water decomposition from measured total heat load in Norwegian buildings. In: BuildSIM-nordic 2020. Oslo: Norway; 2020. p. 244–51.
[22] Ivanko D, Sørensen ÅL, Nord N. "Splitting measurements of the total heat demand in a hotel into domestic hot water and space heating heat use,". Energy 2021;219:119685. https://doi.org/10.1016/j.energy.2020.119685. Mar.
[23] Hedegaard RE, Kristensen MH, Petersen S. "Experimental validation of a model-based method for separating the space heating and domestic hot water components from smart-meter consumption data,". E3S Web of Conferences 2020;172:12001. https://doi.org/10.1051/e3sconf/202017212001.
[24] Alzaatreh A, Mahdjoubi L, Gething B, Sierra F. Disaggregating high-resolution gas metering data using pattern recognition. Energy Build Oct. 2018;176:17–32. https://doi.org/10.1016/j.enbuild.2018.07.011.
[25] RStudio. (2022, Mar. 25). RStudio [Online]. Available: https://www.rstudio.com/products/rstudio/.
[26] Cai M, Pipattanasomporn M, Rahman S. Day-ahead building-level load forecasts using deep learning vs. traditional time-series techniques. Appl Energy Feb. 2019;236:1078–88. https://doi.org/10.1016/j.apenergy.2018.12.042.
[27] Crawley MJ. The R book. first ed. John Wiley and Sons Ltd; 2007.
[28] Ulseth R, Lindberg KB, Georges L, Alonso MJ, Utne Å. "Measured load profiles and heat use for 'low energy buildings' with heat supply from district heating,". Energy Proc 2017;116:180–90. https://doi.org/10.1016/j.egypro.2017.05.066.
[29] Moritz S, Bartz-Beielstein T. imputeTS: time series missing value imputation in R. R J 2017;9(1):207–18. https://doi.org/10.32614/rj-2017-009.
[30] Meyer D. Support Vector Machines: The interface to libsvm in package e1071. 2022 [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/papers/ijcnn.ps.gz.
[31] D. Meyer et al. (2022, Feb. 07). CRAN - Package e1071," CRAN – Package [Online]. Available: https://cran.r-project.org/web/packages/e1071/index.html.
[32] Johra H, Mans M, Filonenko K, De Jaeger I, Saelens D, Tvedebrink T. Evaluating different metrics for inter-model comparison of urban-scale building energy simulation time series. In: Proceedings of building simulation 2021: 17th conference of international building performance simulation association. Belgium: Bruges; 2021.
[33] O. Perpinan Lamigueiro. (2022, Feb. 21). tdr: Target Diagram [Online]. Available: https://cran.r-project.org/web/packages/tdr/tdr.pdf.
[34] Aggerholm S, Skovgaard M. SBi-ANVISNING 213: bygningers energibehov. Copenhagen: SBi Forlag; 2018.
[35] Norm for vandinstallationer. Dansk Standard – DS, 439; 2009. 2009.

3.3 Impact of data truncation on SH and DHW estimation

<u>Paper 3</u>

"Validation of a new method to estimate energy use for space heating and hot water production from low-resolution heat meter data"

Daniel Leiria, Hicham Johra, Evangelos Belias, Davide Quaggiotto, Angelo Zarrella, Anna Marszal-Pomianowska, Michal Zbigniew Pomianowski, BuildSim Nordic Conference 2022, Copenhagen, Denmark, https://doi.org/10.1051/e3sconf/202236210001.

Validation of a new method to estimate energy use for space heating and hot water production from low-resolution heat meter data

Daniel Leiria[1*], Hicham Johra[1], Evangelos Belias[2,3], Davide Quaggiotto[4], Angelo Zarrella[4], Anna Marszal-Pomianowska[1], Michal Pomianowski[1]
[1]Department of the Built Environment – Aalborg University, Aalborg, Denmark
[2] ESTIA SA, EPFL Innovation Park, Lausanne, Switzerland
[3]Human-Oriented Built Environment Lab, School of Architecture, Civil and Environmental Engineering, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[4]Department of Industrial Engineering – University of Padova, Padua, Italy
* *corresponding author: dle@build.aau.dk*

Abstract

One of the initiatives to reach the European decarbonization goal is the roll-out of smart heating meters in the building stock. However, these meters often record the total energy usage with only hourly resolution, without distinguishing between space heating (SH) and domestic hot water (DHW) production. To tackle this limitation, this paper presents the validation of a new methodology to estimate the SH and DHW from total measurements in different building types in three countries (Denmark, Switzerland, and Italy). The method employs a combined smoothing algorithm with a support vector regression (SVR) to estimate the different heating uses. The estimation results are compared with the different countries' DHW compliance calculations. The comparison showed that the compliance calculations outperformed this method by considering the validation dataset characteristics.

Introduction

The society is being pressed to become more sustainable. These pressing sustainable challenges are due to the global climate change, pollution issues, and fossil fuel supply curtailment. A "green" transition must occur, especially for the building sector. According to European Commission (2022a), in the European Union (EU), its building sector has an estimated share of 40% of the total energy end-usage, where 79% of it is for space heating (SH) and domestic hot water (DHW) production alone (European Commission (2022b)). It is estimated that 97% of the existing buildings in the EU must be renovated to achieve its 2050 environmental goals (BPIE (2017)). This estimation is based solely on the energy performance certificates (EPC) issued in the different EU member-states. An EPC results from several calculations made by an expert to estimate a building's energy usage and efficiency. These calculations are based on different measurements, assumptions, and standards depending on the country where the building is located. The objective behind these certificates is to raise awareness of energy efficiency among the owners and tenants, promote the refurbishment of the building, and assess the overall country building stock (Iribar et al. (2021); Gonzalez-Caceres et al. (2022)). Even though the EPCs are promising, they usually show a significant difference between the measured and estimated energy usage. This difference is known as the performance gap (Cozza et al. (2021)) and has been studied in several EU countries (Gram-Hanssen and Hansen (2016)). In order to solve this issue, one of the proposed solutions is the usage of actual measurements as additional information for performing the EPC calculations. This manuscript focuses on how the actual building heating measurements can be used to estimate the SH and DHW shares and compare them with the countries' current compliances to estimate the yearly DHW consumption in the EPCs. The countries studied are Denmark, Switzerland, and Italy. Therefore, each country's effort in using energy data to decrease the performance gap is explained below.

As a front-runner country, Denmark is making a great endeavor to install smart heat meters in buildings connected to the district heating (DH) network (Johra et al. (2020)). The resulting data from the meters are the aggregated heating usage (SH and DHW), water consumption, and temperature-weighted volume consumption, resulting in monotonically increasing measurements (Kristensen and Petersen (2021)). Also, these meters typically have hourly measurements, and the data are easily accessible by the utility companies. Although this initiative is a substantial move toward achieving the energetic target set by Denmark (Danish Climate Policies | Energistyrelsen (2022)), it has a downside regarding its data collection. In most buildings, only one device is installed, which collects the total heat usage without differentiating between the energy used for SH or DHW production. Because these two heat uses depend on different factors, it is crucial to disaggregate them to understand better the building and occupancy heat demand (Gram-Hanssen (2014)).

Even though the DHW in Swiss nZEB accounts for 50-70% of the total heat consumption, its monitoring is not required by the local regulations or the EPC (Office fédéral de l'énergie OFEN SuisseEnergie (2016); Flourentzou and Pereira (2021)). On the contrary, in Switzerland, it is common for the buildings to be equipped with one heat meter that measures the total heat consumption, both for SH and DHW, making it challenging to identify the heat required for the DHW or SH production (Flourentzou and Pereira (2021)).

In Italy, in the last years, the Government promoted several energy conservation measures for the building envelope with related incentives due to the prevalence of old buildings. So, as a matter of fact, the energy consumption of the building stock will change in the future (also considering the climate change effect). Consequently, the district heating networks in the main cities of northern Italy, which were built several decades ago and are operating at high temperatures (70-80°C), need to be revised in terms of both production and operating conditions. An example of such an intervention is studied by Vivian, Quaggiotto and Zarrella (2020). The heating and DHW disaggregated profiles will help design and manage these improvements efficiently. In addition, the recent concept of the district heating network integrated with other renewable energy technologies (e.g., heat pumps) in new building districts is a good opportunity (Bordignon et al. (2022)). Also, in this case, the disaggregated profiles can help design and set suitable control strategies to increase energy efficiency.

Another aspect to consider on the importance of knowing these energy shares is regarding refurbishment initiatives. In Pomianowski et al. (2020), the authors argue that global building regulations have stricter SH efficiency rules while overlooking DHW consumption. Therefore, the new buildings, also known as low-energy buildings, have a much higher DHW share due to the continuous decrease of SH usage over the years and the higher levels of comfort concerning heating practices demanded by the residents.

Thus, a better assessment of the thermal appliances can be achieved by disaggregating the energy used in buildings. This contributes to a more detailed understanding and control on the demand side and promotes better decision-making strategies regarding heat production and distribution.

Background

The disaggregation of time-series has been studied since the 1980s regarding electrical appliances metering (Zeifman and Roth (2011)). However, the research has been shifting towards heating meter data. One of the first articles to explore this type of data is Bacher et al. (2016), which presents a statistical methodology to estimate the SH from 10-min resolution total heat measurements. This method is based on the premise that SH demand variates accordingly to the smooth external temperature fluctuations. At the same time, DHW usage fluctuates sporadically with higher peaks due to its short-time hot water draw-off events. The method predicts the SH by applying a kernel smoother to the total measurements, where all values above a defined threshold are due to DHW usage. Although promising, the method needs validation with separated space heating and DHW usage measurements. Also, the need for high-resolution data (10-minutes resolution) to detect the DHW peaks is uncommon to find in the typical installed smart meters.

Unlike the above method, more straightforward methods were developed to disaggregate heating datasets. The articles Lien, Ivanko and Sartori (2020), and Ivanko, Sørensen and Nord (2021) propose different methods to decompose SH and DHW usage based on discovering the DHW profiles when the total heating is assumed to be equal to the DHW usage only (no SH demand). Additionally, considering the relationship between SH demand and the external temperature, the methods were validated with several Norwegian buildings (apartments and hotels) and compared with other existing methods. Also worth mentioning regarding Lien, Ivanko and Sartori (2020) is that besides presenting their developed methodology, they also compared their results with several Norwegian reference data.

In Marszal-Pomianowska et al. (2019), another technique is proposed by assuming that the total heat measurements are equal to the DHW usage during Summer (no SH demand). Their novel approach does not aim to disaggregate the data but to predict the dwelling's daily DHW usage profile. This load-profiling technique seeks to throw light on the customers' DHW practices and how their behavior affects the DH supplier.

In Hedegaard, Kristensen, and Petersen (2021), the weekly SH and DHW usage profiles are predicted using calibrated grey-box models. This method is likely to be the most reliable and accurate of the ones presented in this review, and the authors claim that the model's accuracy can be improved even further. Also worth mentioning is Alzaatreh et al. (2018). A pattern recognition technique was developed and tested in this research work to separate SH measurements from other appliances in two UK single-family dwellings.

This manuscript aims to present the results from the validation of a novel disaggregation methodology described in Leiria et al. (under review). The validation process is constituted by applying the method in three smart heat meter datasets. These datasets are different in terms of measurements resolution (i.e., number of decimal digits), measurements scale (i.e., energy usage in a single apartment or a block of apartments), building type (i.e., residential or commercial buildings), heating systems (i.e., DHW production with or without storage tank) and different countries (i.e., Denmark, Switzerland, and Italy). The current study compares the DHW estimation by the disaggregation methodology with the actual measurements and the DHW compliance calculations of each country.

Following the *Introduction*, the section *Study Case* presents the different validation datasets. In *Methodology*, the applied disaggregation method is explained. The results from the validation are examined in the section *Results and Discussion*. The manuscript closes with *Conclusion* and *Suggestions for Further Work*.

Study Case

For the methodology's validation, three heating datasets are used. All datasets have separated energy measurements of SH, DHW, and the aggregated sum of both (total heat). The differences between datasets are the following.

Danish dataset

This dataset is constituted of 28 single-family apartments. All apartments are from a social housing complex in Aalborg, Denmark. The complex was progressively refurbished to the Nearly Zero-Energy Buildings (NZEB) standard from 2012 to 2020. The interior of the apartments was fully remodeled, and the new SH installation includes radiators in all rooms and kitchens and underfloor heating in the bathrooms and hallways. The heat for SH and DHW is produced at the building block level and distributed to each apartment. Apartments are equipped with single SH and total heat usage meters, and the DHW is calculated through the difference between measurements from the meters. The heated area of the dwellings is between 97 and 112 m^2.

The local weather data (hourly outdoor temperature and the global radiation) is retrieved from the Danish Meteorologic Institute website (Dansk Meterologisk Institut (2022)). The chosen weather station is Tylstrup, the nearest available station to Aalborg.

In this work, the data pre-processing consisted in detecting the number of missing and negative measurements and removing them. In the 28 dwellings dataset (187 123 measurements) with approximately nine months of monitoring for each dwelling, there are 46 661 missing hours (~25% of the dataset). The household with the lowest missing measurements has approximately 3% missing data. Some households have up to 43% of missing data. Regarding negative measurements (incorrect values), few apartments have those. In total, these values only represent 0.013% of the original dataset.

Swiss dataset

This dataset is constituted of an apartment building located in Vevey, Switzerland. The building was built in 20120 and deeply refurbished to reach NZEB standards during 2018-2019. The local district heating network supplies heat for SH and DHW. A heat meter measures the total heat provided by the district heating network. A second, a Flexim ultrasonic portable flowmeter (Fluxus F601), was used to measure the heat consumption for the DHW.

The hourly outdoor temperature and the global radiation data were collected from the Swiss Federal Office of Meteorology and Climatology-MeteoSwiss (Swiss Federal Office of Meteorology and Climatology-MeteoSwiss (2022)). The "Vevey" station was used for the weather data as it was the nearest available. The data were pre-processed in order to identify the missing and negative values and remove them. In total, for 2020, there were five months of available valid data.

Italian dataset

The selected building dataset consists of a theatre and a rehab institution connected to the district heating network of Verona Centro Città, serviced by AGSM. This network supplies heat to residential, tertiary, and industrial customers, operating at constant supply temperature and variable flow rate. Overall there are 247 user substations, but just for 2 (theater and rehab institution) of these 247, the separate monitoring on the use of SH or DHW was provided. The measures all correspond to the primary circuit of the heat exchanger installed at each user substation. The measuring devices installed are all ultrasonic compact energy meters suitable for measuring the energy consumption of district heating systems. The principle of operation of these meters is static and based on the transit time measurement. In particular, ultrasonic meters are characterized by the absence of moving parts, thus preventing mechanical wear of the metering components, low-pressure losses, low start flowrate, and good tolerance to suspended particulates in the water flow. On the whole, the ultrasound principle assures stable and accurate measuring results. The measurement period is from December 1, 2021, to January 31, 2022, for the rehab institution and from January 11 to January 31, 2022, for the theater. The resolution of the measured data is a 15-minute time step.

The local weather data (global solar radiation and air temperature with hourly time step) has been provided by the Arpav Meteorological Institute of Teolo.

Methodology

The methodology starts with the premise that the SH system runs continuously during the heating season while the DHW usage is sporadic throughout the day. Hence, during a day (which has around 24 recorded heating measurements), only a few of those consist of combined SH and DHW usage ($E_{Total} = E_{SH} + E_{DHW}$). The other recorded data points are SH usage alone ($E_{Total} = E_{SH}$). Following this premise, the method has two stages. The first is to segregate the data points with and without DHW production. The second is to estimate the SH share ($E_{SH,estim}$) in the points identified with DHW usage. From the SH estimation, the DHW is calculated through Equation 1:

$$E_{DHW,estim} = E_{Total} - E_{SH,estim} \tag{1}$$

The estimation results are compared with the separated measurements (SH and DHW usage) for each dataset. The DHW values obtained by the disaggregation methodology with the DHW prediction from the different countries' compliance calculations. The disaggregation methodology is disclosed in more detail below. Furthermore, the algorithm presented in this work is coded with the software *Rstudio* (RStudio (2022)).

Energy separation

This first part of the method starts from the same premise as Bacher et al. (2016) that moderate variations of outdoor temperature during the day combined with the inertia of the building environment contribute to smooth SH daily fluctuations. Hence, all peaks recorded by the meters can be accounted for DHW usage. Therefore, the method detects all daily highest points (E_{Total}) and identifies them containing DHW and SH usage ($E_{Total} = E_{SH} + E_{DHW}$). For each day, the method assumes the seven-highest recorded values as DHW usage, while the other measurements are considered SH alone. It is also assumed a sleeping period from 1:00 – 4:00 hours every day. Thus, there is no DHW demand during the sleeping period, and the high values recorded are because of the SH system operation. In Figure 1, one can see a schematic representation of the separation method.

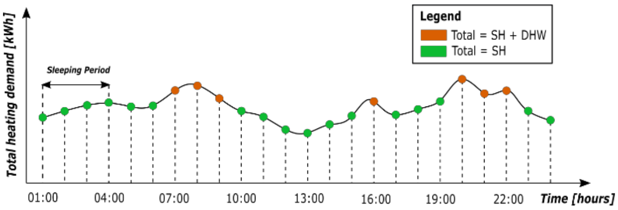


Figure 1: Separation method's representation.

All points identified with DHW production are removed from the dataset in order to have only SH measurements. The remaining SH data points will be used to estimate the SH from the removed recordings. The estimation algorithm is explained in the following subsection.

SH and DHW estimation

At this stage, the smart meters' dataset consists of measurements without DHW production ($E_{Total} = E_{SH}$). The next stage of the methodology is to estimate the SH usage ($E_{SH,estim}$) at the data gaps. After determining the $E_{SH,estim}$, the DHW usage ($E_{DHW,estim}$) is calculated with Equation 1.

From the same starting argument of the energy separation, the SH demand will vary smoothly due to small outdoor temperature oscillations. Therefore, the SH share in the removed data points is predicted from its known neighboring SH measurements that remained in the dataset. To estimate the SH, a smoothed Kalman filter algorithm is applied. This algorithm is based on a structural time series model from the function "StructTS" in the R-package *imputeTS* (Moritz and Bartz-Beielstein (2017)). The package's selected function consists of a linear Gaussian state-space model for univariate time series.

From the results in Leiria et al. (under review), the Kalman smoothing technique is a good method to predict the SH demand in the missing values. However, as mentioned, these values are calculated by their neighboring points. Basing this estimation on the adjacent points raises the risk of inaccuracy when several points are removed sequentially (large gap). To solve this problem, the algorithm is refined to use the smoothed Kalman filter only when the number of hours removed consecutively is equal to or below 2 hours (gap ≤ 2 hours). If the data gap is larger, a support vector regression (SVR) is applied instead. The SVR is a machine learning regressor that is trained with the known SH points that remained in the dataset and takes into account other inputs to calculate the SH usage instead of the neighboring points. The input data to estimate a given SH share is the outdoor temperature and global solar radiation measured two and one hours prior to the missing point and the smart meter measurements before and after the missing point. The SVR model uses a radial kernel function with the parameters C (cost) and γ (gamma) equal to 7 and 0.01, respectively. The SVR algorithm is retrieved from the R-package *e1071* (Meyer et al. (2020)). One can see in Table 1 the details regarding the estimation algorithms.

Table 1: Methods' description.

Method	Parameters	Input	Condition
Kalman filter	Model: StructTS Smoothed: True	$E_{Total}\,[i]$	Gap ≤ 2 hours
SVR	Kernel: Radial $C = 7$ $\gamma = 0.01$	$T_{out}\,[i-1, i-2]$ $Rad\,[i-1]$ $E_{Total}\,[i-1, i+1]$	Gap > 2 hours

The final part of the present methodology compares the estimated values from the methodology and the actual measurements. Also, it explicitly compares the DHW method's prediction on the rounded measurements (present case buildings), prediction on the decimal values (from the study in Leiria et al. (under review)), and the DHW estimation from the compliance calculations in the different countries (as described below).

Danish DHW compliance calculations

In Denmark, the DHW consumption in residential buildings is currently estimated using the compliance calculation of 250 liters/m^2 per year (Aggerholm and Skovgaard (2018)). Similarly, the supplied cold water and DHW temperatures are 10ºC and 55ºC, respectively (Dansk Standard (2000)). By using the floor area of the different apartments, the yearly DHW energy production is calculated through Equation 2:

$$E_{DHW}^{DK} = \frac{1}{3600} \cdot 0.25A \cdot \rho_w c_{p,w} \cdot (T_{DHW} - T_c) \qquad (2)$$

Swiss DHW compliance calculations

In Switzerland, the DHW consumption in residential apartment buildings is currently predicted using the compliance calculation of 35 liters/day per person, and each person is considered to occupy 30 m^2 of floor area (Société suisse des ingénieurs et des architectes (2015)). Similarly, the supplied cold water and DHW water temperatures are 10ºC and 60ºC, respectively (Société suisse des ingénieurs et des architectes (2015)). Thus, by knowing the building's floor area, the yearly DHW energy production is calculated through Equation 3:

$$E_{DHW}^{CH} = \frac{365}{3600} \cdot \frac{0.035}{30} A \cdot n \cdot \rho_w c_{p,w} \cdot (T_{DHW} - T_c) \quad (3)$$

Italian DHW compliance calculations

In Italy, the DHW consumption in specific (commercial) buildings is estimated using particular compliance calculations and standards. The Italian dataset has the heating measurements of a rehab institution and a theatre, therefore, the DHW consumption (volume) is calculated accordingly for each building case. The rehab institution accounts for a water volume (V_w) of 80 liters/day per existing bed in the building (Ente Nazionale Italiano di Unificazione (2014)), and the number of days regarding the calculation period (G) is equal to 365. The theatre's DHW consumption (Vw) is given at 3.8 liters/day per person (ISO (2016)). The theatre is divided into zones where the number of people will variate accordingly. The number of days (G) for this case is 251 (ISO (2016)). For both cases, the supplied cold water and DHW temperatures are 13ºC and 40ºC, respectively (Ente Nazionale Italiano di Unificazione (2014)). Thus, by knowing the buildings' bed number or occupants' number (n), the yearly DHW energy production is calculated through Equation 4:

$$E_{DHW}^{IT} = \frac{G}{3600} \cdot 10^{-3} V_w \cdot n \cdot \rho_w c_{p,w} \cdot (T_{DHW} - T_c) \quad (4)$$

Results and Discussion

The first set of results from this research is from applying the energy separation algorithm to identify the measurements with DHW usage. To assess the identification accuracy, the percentage of incorrectly identified measurements is shown per case building in Figure 2.

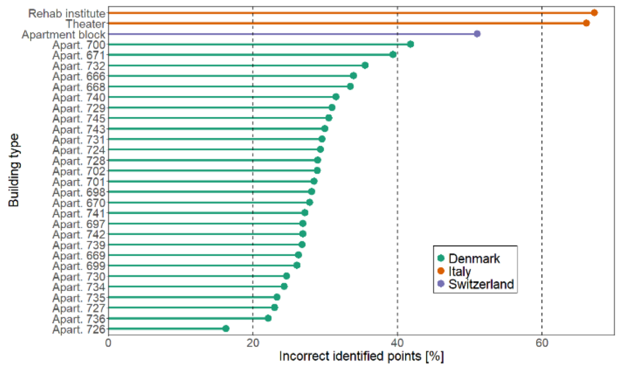


Figure 2: Incorrectly identified points percentage per building type.

The results show that the separation approach is quite inaccurate in identifying the DHW draw-off events. The lowest percentages belong to the Danish cases (single-family apartments), with the lowest value of 16%. The largest inaccurate identified points belong to the Italian cases with the extreme of 67%. The plot corroborates the hypothesis that this separation approach performs better for households than commercial buildings.

The following step in the methodology is the estimation of the SH usage in the detected DHW points. The estimation algorithm combines two methods, smoothed Kalman filter estimator and SVR, as described in the *Methodology*. In Figure 3, it is presented the overall error between the estimated values (SH – upper area; DHW – down area) and the real measurements.

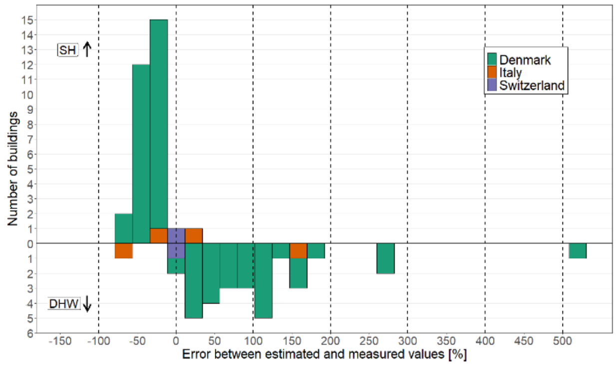


Figure 3: Overall error of the SH and DHW estimation for each building.

As one can see from Figure 3, the overall SH error (upper area) is mainly negative (underestimated) and has a lower error than the estimated DHW, with the buildings having a SH error between -65% and 17%.

Concerning the DHW prediction (down area), the error distribution is much wider than the SH predictions. In this case, 13 of the buildings on the dataset have an overestimation of the DHW demand above +100%. The extreme DHW prediction is one single-family dwelling with an overestimation of +510% and only two apartments being underestimated.

Several reasons can be outlined to explain these energy predictions and their overall error. Foremost, the separation method inaccurately identifies several measurements, decreasing the estimation's accuracy from the start. From Leiria et al. (under review), it is seen that in this research work, the separation approach underperforms more in single-family apartments. This is due to the coarse measurements (rounded values), which hinder the algorithm from finding the maximum values because most data points have the same value (e.g., 1, 2, 3 kWh). Also relevant is that this method has a significant inaccuracy for the commercial buildings. To overcome this challenge, a separation approach can be developed, taking into account the maximum values (as done in this manuscript) and the occupancy schedule. Because these buildings have such strict schedules (e.g., opening and closing hours), a more precise method can be developed to account for these characteristics.

Another factor is the occurrence of missing measurements in the initial dataset. As one can see in the *Study Case*, the different countries' datasets are comprised of large missing measurement gaps (Denmark and Switzerland) or small timespan measurements (Italy). Because the prediction relies on determining the SH usage based on its neighboring points, several missing points negatively impact the overall method's accuracy.

Furthermore, the different heating systems and people's social cultures significantly impact the methodology. As described, some of the DHW systems are of instantaneous heat production (Denmark). However, others have a storage tank (Switzerland), which in itself affects the DHW usage detection. Besides the production system, the unique dwellers' consumption habits or the DHW usage being equal to zero (no occupancy) may influence the method's performance, which might explain the extreme error estimated cases.

The present work also assesses the estimated DHW values from the method with the different countries' compliance calculations used to predict the DHW demand in the buildings. In this comparison, the values presented in the manuscript Leiria et al. (under review) are also displayed ("decimal" column). The results of this comparison are in Table 2.

Table 2: Comparison between the countries' compliance predictions and the method's estimation results. The green-colored cells indicate the best (orange color – the worst) performing method between this research's method (rounded and decimal measurements) and the compliance calculations when comparing with the actual DHW measurements.

Data	Case-building	Error		
		Compliance	Round	Decimal
DK	Apart 666	-47%	97%	0%
DK	Apart 668	-42%	103%	21%
DK	Apart 669	-11%	102%	22%
DK	Apart 670	-72%	21%	-6%
DK	Apart 671	-34%	108%	20%
DK	Apart 697	-76%	12%	-12%
DK	Apart 698	-75%	21%	-7%
DK	Apart 699	-76%	10%	-13%
DK	Apart 700	123%	510%	85%
DK	Apart 701	-1%	93%	18%
DK	Apart 702	87%	182%	32%
DK	Apart 724	-28%	89%	11%
DK	Apart 726	43%	70%	14%
DK	Apart 727	61%	149%	18%
DK	Apart 728	11%	152%	37%
DK	Apart 729	14%	119%	12%
DK	Apart 730	-57%	43%	5%
DK	Apart 731	90%	273%	63%
DK	Apart 732	-60%	24%	-15%
DK	Apart 734	59%	144%	17%
DK	Apart 735	-50%	44%	6%
DK	Apart 736	-51%	40%	1%
DK	Apart 739	-68%	34%	7%
DK	Apart 740	1%	75%	-3%
DK	Apart 741	-30%	59%	7%
DK	Apart 742	0%	121%	15%
DK	Apart 743	-64%	29%	-13%
DK	Apart 745	78%	265%	69%
CH	Apart. block	4%	-9%	-
IT	Rehab inst.	-59%	-79%	-
IT	Theater	-35%	154%	-

As shown in Table 2, there are three calculated errors per DHW usage. The error between the measured DHW usage and the DHW compliance calculations ("Compliance"), the error between the actual measurements and the results from the methodology applied on this manuscript dataset ("Round"), and the error between the DHW measurements and the results from the methodology applied on the Leiria et al. (under review) dataset ("Decimal"). The error calculation is performed using the aggregated DHW usage divided by the number of data points (hours). For the case of the DHW measurements and the disaggregation method, the number of data points is the number of measurement hours in each building. For the compliance case, the number of data points is the number of hours in a year. In most cases, the compliance calculations outperform (green color) the disaggregation methodology when applied to rounded values. However, the total heating values are recorded with decimal number, its accuracy increases and outperforms the DHW compliance prediction. The main reasons behind this performance are stated above. However, it is relevant to highlight that even though the compliance calculations are in some cases more accurate, the methodology should be reviewed or changed because the results are still too high and might be one of the main reasons for the observed building energy performance gap.

Conclusion

This article presents a validation study on a new data-driven methodology to estimate the SH and DHW from hourly resolution heat meters data. The validation novelty is the application of different building cases with different characteristics, e.g., different measurements resolution, building types, heating systems, and countries (consumption habits).

The validation process shows that the method is quite inadequate to detect DHW usage in rounded measurements or commercial buildings. To solve these challenges, it is argued that the measurements cannot be rounded and should be recorded with decimals, and that the separation algorithm must be refined by taking into account the occupancy schedules in large buildings. The overall methodology predicts better the SH demand with an error between -65% and 17%. Concerning DHW prediction, the error is much wider, with most building cases falling between 0% and 200%. Additionally, this study compared the estimated DHW demand predicted by the method with the actual measurements and the DHW compliance calculations used in Denmark, Switzerland, and Italy. This comparison concludes that the compliance estimations outperform this method for most building cases, when the used rounded values. However, it is argued that the compliance calculations must be updated or replaced to estimate more precisely the buildings' DHW demand, hence decreasing the energy performance gap and improving the EPCs' accuracy.

Suggestions for Further Work

A suggestion for further work is the application of this methodology with other datasets for further validation and robustness analysis. Improving the separation methodology for rounded measurements and commercial cases is highly needed.

It is also advised to benchmark this methodology with other existing disaggregation techniques on a common dataset. Additionally, a more extensive endeavor must be made to collect good quality datasets and share them with our research peers.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 893945 (E-DYCE). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 958345 (PRELUDE). This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement number 754354. This project has received funding from the Interdepartmental Centre "G. Levi Cases" of the University of Padova within the Regional Project GHOTEM (POR-FESR 2014-2020, ID 10064601).

The authors would like to acknowledge the Himmerland housing association and Varmekontrol A/S for their support in data collection in Denmark and to all partners involved in the E-DYCE and PRELUDE projects. Also, the first author would like to express his gratitude to his colleagues in the Ph.D. "Corner Office" for the valuable discussions and support.

The authors declare no conflict of interest

Nomenclature

Acronyms	
CH	Switzerland (country code)
DHW	Domestic hot water
DK	Denmark (country code)
EPC	Energy performance certificate
EU	European Union
IT	Italy (country code)
SH	Space heating
SVR	Support vector regression
Symbols and variables	
A	Floor area [m^2]
C	Cost (SVR parameter) [-]
$C_{p,w}$	Water specific heat capacity – Constant: 4.18 [kJ/kgºC]
E_{DHW}	Measured domestic hot water energy usage [kWh]
$E_{DHW,compl}$	Estimated annual DHW energy usage from any compliance [kWh/year]
$E_{DHW,estim}$	Estimated domestic hot water energy usage [kWh]
E_{DHW}^{DK}	Estimated annual DHW energy usage from Danish compliances [kWh/year]
E_{DHW}^{CH}	Estimated annual DHW energy usage from Swiss compliances [kWh/year]
E_{DHW}^{IT}	Estimated annual DHW energy usage from Italian compliances [kWh/year]
E_{SH}	Measured space heating energy usage [kWh]
$E_{SH,estim}$	Estimated space heating energy usage [kWh]
E_{Total}	Measured total heat usage (smart meter measurements) [kWh]
G	Number of days in calculation period [days]
n	Number of people or beds [-]
Rad	Global solar radiation [W/m^2]
T_c	Temperature of inlet cold water [ºC]
T_{DHW}	Temperature of outlet DHW water [ºC]
T_{out}	Outdoor temperature [ºC]
γ	Gamma (SVR parameter) [-]
ρ_w	Water density – Constant: 1000 [kg/m^3]

References

Aggerholm, S. and M. Skovgaard (2018) *SBi-ANVISNING 213: Bygningers energibehov*. Sbi Forlag. Copenhagen (Denmark).

Alzaatreh, A. and L. Mahdjoubi, B. Gething, and F. Sierra (2018). Disaggregating High-Resolution Gas Metering Data Using Pattern Recognition. *Energy and Buildings 176*, 17–32.

Bacher, P., P. Anton de Saint-Aubain, L. E. Christiansen, and H. Madsen (2016). Non-Parametric Method for Separating Domestic Hot Water Heating Spikes and Space Heating. *Energy and Buildings 130*, 107–112.

BPIE (2017). Factsheet: 97% of Buildings in the EU Need to Be Upgraded. *https://www.bpie.eu/publication/97-of-buildings-in-the-eu-need-to-be-upgraded/*.

Cozza, S., J. Chambers, A. Brambilla, and M. K. Patel (2021). In Search of Optimal Consumption: A Review of Causes and Solutions to the Energy Performance Gap in Residential Buildings. *Energy and Buildings*.

Danish Climate Policies | Energistyrelsen (2022). Accessed March 25. *https://ens.dk/en/our-responsibilities/energy-climate-politics/danish-climate-policies*.

Dansk Meterologisk Institut (2022). DMI Frie. Accessed April 5. *https://www.dmi.dk/frie-data/*.

Dansk Standard (2000). *Norm for Vandinstallationer (DS 439:2009)*.

European Commission (2022a). Energy Performance of Buildings Directive. Accessed January 27. *https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en*.

European Commission (2022b). Heating and Cooling. Accessed January 27. *https://energy.ec.europa.eu/topics/energy-efficiency/heating-and-cooling_en*.

Flourentzou, F. and J. Pereira (2021). Domestic Hot Water Optimizing Potential in Existing or Renovated Multifamily Residential Buildings. *In Journal of Physics: Conference Series, 2042:012144*.

Gonzalez-Caceres, A., J. Karlshøj, T. A. Vik, E. Hempel, and T. R. Nielsen (2022). Evaluation of Cost-Effective Measures for the Renovation of Existing Dwellings in the Framework of the Energy Certification System: A Case Study in Norway. *Energy and Buildings*.

Gram-Hanssen, K. (2014). New Needs for Better Understanding of Household's Energy Consumption - Behaviour, Lifestyle or Practices?. *Architectural Engineering and Design Management 10* (1–2).

Gram-Hanssen, K. and A. R. Hansen (2016). Forskellen Mellem Målt Og Beregnet Energiforbrug Til Opvarmning Af Parcelhuse. Copenhagen.

Hedegaard, R. E., M. H. Kristensen, and S. Petersen (2020). Experimental Validation of a Model-Based Method for Separating the Space Heating and Domestic Hot Water Components from Smart-Meter Consumption Data. In *E3S Web of Conferences, 172:12001*.

Iribar, E., I. Sellens, L. Angulo, J. M. Hidalgo, and J. M. Sala (2021). Nonconformities, Deviation and Improvements in the Quality Control of Energy Performance Certificates in the Basque Country. *Sustainable Cities and Society 75*.

International Organisation for Standardisation (2016). *Energy Performance of Buildings-Schedule and Condition of Building, Zone and Space Usage for Energy Calculation-Part 1: Non-Residential Buildings (ISO 18523-1).*

Ivanko, D., Å. L. Sørensen, and N. Nord (2021). Splitting Measurements of the Total Heat Demand in a Hotel into Domestic Hot Water and Space Heating Heat Use. *Energy 219.*

Johra, H., D. Leiria, P. Heiselberg, A. Marszal-Pomianowska, and T. Tvedebrink (2020). Treatment and Analysis of Smart Energy Meter Data from a Cluster of Buildings Connected to District Heating: A Danish Case. *In E3S Web of Conferences, 172:12004.*

Kristensen, M. H., and S. Petersen (2021). District Heating Energy Efficiency of Danish Building Typologies. *Energy and Buildings 231*.

Leiria, D., H. Johra, A. Marszal-Pomianowska, and M. Z. Pomianowski (under review). A Methodology to Estimate Space Heating and Domestic Hot Water Energy Demand Profile in Residential Buildings from Low-Resolution Heat Meter Data. *Energy*.

Lien, S., D. Ivanko, and I. Sartori (2020). Domestic Hot Water Decomposition from Measured Total Heat Load in Norwegian Buildings. *In BuildSim-Nordic 2020 Selected Papers*.

Marszal-Pomianowska, A., C. Zhang, M. Pomianowski, P. Heiselberg, K. Gram-Hanssen, and A. R. Hansen (2019). Simple Methodology to Estimate the Mean Hourly and the Daily Profiles of Domestic Hot Water Demand from Hourly Total Heating Readings. *Energy and Buildings 184*.

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C. Chang, and C. Lin (2020). CRAN - Package E1071. *CRAN - Package. https://cran.r-project.org/web/packages/e1071/index.html.*

Moritz, S. and T. Bartz-Beielstein (2017). ImputeTS: Time Series Missing Value Imputation in R. *R Journal 9 (1),* 207–218.

Bordignon, S., D. Quaggiotto, J. Vivian, G. Emmi, M. De Carli, and A. Zarrella (2022). A Solar-Assisted Low-Temperature District Heating Network Coupled with a Ground-Source Heat Pump. *Energy Conversion and Management 267*.

RStudio (2022). Accessed March 25. *https://www.rstudio.com/products/rstudio/*.

Société suisse des ingénieurs et des architectes (2015). *Données d'utilisation Des Locaux Pour l'énergie et Les Installations Du Bâtiment (SIA 2024).*

The Danish Government (2020). A Green and Sustainable World - Denmark's Global Climate Action Strategy. *https://um.dk/~/media/um/klimastrategi/onepager eng dks globale klimastrategifinal.pdf.*

Ente Nazionale Italiano di Unificazione (2014). *Part 2: Evaluation of Primary Energy Need and of System Efficiencies for Space Heating and Domestic Hot Water Production (UNI/TS 11300).*

Vivian, J., D. Quaggiotto, and A. Zarrella (2020). Increasing the Energy Flexibility of Existing District Heating Networks through Flow Rate Variations. *Applied Energy 275*.

Zeifman, M., and K. Roth (2011). Nonintrusive Appliance Load Monitoring: Review and Outlook. *IEEE Transactions on Consumer Electronics 57 (1)*, 76–84.

3.4 Improving disaggregation methodology performance

Paper 4

"Estimating residential space heating and domestic hot water from truncated smart heat data"

Daniel Leiria, Markus Schaffer, Hicham Johra, Anna Marzsal-Pomianowska, Michal Zbigniew Pomianowski, CISBAT 2023, Lausanne, Switzerland, https://doi.org/10.1088/1742-6596/2600/2/022017.

Estimating residential space heating and domestic hot water from truncated smart heat data

D Leiria[1*], M Schaffer[1], H Johra[1], A Marzsal-Pomianowska[1], and M Z Pomianowski[1]

[1] Aalborg University, Department of the Built Environment, Aalborg, Denmark

* Corresponding author: dle@build.aau.dk

Abstract. The EU aims to digitize the building stock across all member states to better understand energy use and achieve energy efficiency goals to address climate change. Smart heat meters are currently used for billing purposes in district heating (DH) grids. Their data is recorded as integer kWh values, which restricts usability for the modeling and analysis of DH networks. Previous research devised a methodology to estimate space heating (SH) and domestic hot water (DHW) energy from total heating data, but the data truncation process reduced accuracy. This study integrates the SPMS (Smooth–Pointwise Move–Scale) algorithm, which estimates decimal values from DH truncated measurements, to improve the accuracy of the DHW and SH disaggregation methods. The study applies these two methodologies to a dataset of 28 Danish apartments and compares the results against full-resolution and truncated data to evaluate performance. Another dataset, named "optimal dataset" is also assessed to determine overall estimation accuracy. Results show that SPMS reduces the disaggregation methodology error of SH and DHW compared to the truncated data. The optimal dataset outperforms the current methodology, indicating a potential for improving and scaling the methodology for larger datasets.

1. Introduction

The building sector plays a major role in our society as it contributes to 40% of the European Union's (EU) total energy end-usage [1]. District heating (DH) systems are cost-effective, flexible, and sustainable energy solutions to fulfill the buildings' heating demand, as already seen, especially in Europe, the USA, Canada, and Asia [2]. This type of system has been through several innovative technological stages throughout the years (named generations). The most current transition is from the 3rd to the 4th generation of district heating (4GDH) systems. The 4GDH is characterized mainly by a lower-temperature heat-carrier fluid supply, bringing the advantage of reducing heat losses in the DH network, increasing the output capacity when integrating with renewable energy sources, minimizing the risk of pipe leakages due to thermal stress, and meeting new building requirements more effectively.

In the 4GDH transition, smart heat meters (SHM) play a vital role in making it possible to manage the energy production and distribution grid, detect faults, and provide information to DH operators and customers (end-users) [3]. Nevertheless, these meters only measure the total heat demand, and to have a better understanding of the DH end-users' heat usage, it is necessary to estimate the space heating (SH) and domestic hot water (DHW) demand separately. This data disaggregation of the energy use in buildings contributes to better decision-making strategies regarding heat production and distribution due to these two heating demands are dependent on different variables.

1.1. State-of-the-art

Several methodologies have been proposed in the last few years to address the problem of disaggregating the energy for SH and DHW from total heating measurements registered by SHM. In Table 1 is listed and summarized these different methodologies based on the state-of-the-art of Leiria et al., 2023 [4].

Table 1. Summary of the existing methods to disaggregate SH and DHW from total heat measurements reproduced from [4].

Ref.	Country	System	Period	Resolution	Dataset	Methodology
[5]	SE	District heating	8 hours for "winter" and "summer" times	5-sec	An apartment building	DH substation model created using Matlab Simulink.
[6]	DK	District heating	1-month	10-min	One single family house (SFH)	SH estimated with a kernel smoother from the total heat measurements.
[7]	UK	Natural gas boilers	2-months	1-min	Two SFHs	Detects the various heat signatures of the appliances.
[8]	DK UK	District heating	Few months to a full year (depending on the building)	1-hour (DK) 10-min (UK)	Apartments and SFHs	Two methods extract the DHW profiles during the no-heating season.
[9]	NO	District heating	1-3 years (depending on the building)	1-hour	58 apartments blocks and 20 hotels	Identifies DHW profiles during the summer period and calculates the SH by subtracting the extrapolated DHW daily profiles from the total values.
[10]	DK	District heating	1-year	1-hour	44 terraced SFHs	Grey-box model-based method to estimate the SH and DHW usage.
[11]	NO	District heating	1-year	1-hour	A hotel with 260 rooms	Two methods use the energy signature (ES) to estimate the SH and DHW. The second employs the singular spectrum analysis to decompose the SH and DHW derived from the ES.
[4]	DK	District heating	3-9 months (depending on the building)	1-hour	28 SFHs	Disaggregation is based on detecting the daily sporadic DHW production and estimating the SH from the neighboring points in the time series.
[12]	DK CH IT	District heating	3-9 months per building (DK) 5-months (CH) 1-2 months (IT)	1-hour	28 SFHs (DK) An apartment block (CH) Other buildings (IT)	Validation of the method [4] with truncated heat measurements and other types of buildings.

Leiria et al. [4] presented a two-step disaggregation algorithm. The first step (separation stage) refers to distinguishing throughout the day, the measurements with simultaneous DHW and SH (SH+DHW data points) usage from those with only SH demand (only SH data points). The second step (estimation stage) is the estimation of the SH in the SH+DHW points based on the neighboring SH values of the data points with only SH usage. Nevertheless, the work was developed based on high-resolution measurements, which is not the reality in most Danish DH cases. The measurements gathered by the DH companies usually have their values rounded down to a 1 kWh resolution (known as truncation) [13]. This means that if a SHM records a specific heating measurement between 1.1 to 1.9 kWh, it is truncated (decimals removed) and recorded as 1.0 kWh. Due to this context, the authors tested their methodology for truncated values in [12]. The conclusions from this second study are that the methodology underperforms significantly and cannot be used for this type of data. Another question raised throughout these two studies [4,12] is the dependency of the SH estimation accuracy (step 2) from the separation stage (step 1). This means that it is envisioned that the SH is better estimated if the separation stage is more precise in segregating the “SH+DHW” measurements from the “only SH” measurements.

1.2. Contributions and novelty of the current study

To tackle the general issue of truncated data in the DH sector, a novel algorithm (named SPMS) is proposed by [14]. It enhances the usability of truncated data by partially restoring the actual underlying trend of the SHM measurements. The work shows that the proposed method can decrease the error from the truncation process, consequently increasing the usefulness of the data. Based on these results, the current study integrates the SPMS method into the disaggregation method from [4] to increase its estimation accuracy when using hourly truncated SHM data.

Regarding the accuracy of the SH estimation (step 2) based on the precision of the separation stage (step 1) question, this manuscript addresses it by assessing the accuracy of the second step of the method [4] (estimation), if the DHW production data points are perfectly detected in the first step (this was done using the validation dataset by categorizing the measurements as "SH+DHW" when the DHW energy demand was larger than 0 kWh). This dataset with perfect detected "SH+DHW" measurements is named "optimal dataset" and is compared with the accuracy assessed when using the method in decimal, truncated, and SPMS retrieved data measurements.

2. Methods

2.1. Dataset description

The study uses a dataset of 28 apartments located in a social housing complex in Aalborg, Denmark. This dataset is the same one used in [4,8,12,14]. The apartments are equipped with individual SH and total heat demand meters. The DHW is thus calculated as the difference between measurements from those two meters. The weather data, the outdoor temperature, and global radiation were taken from the Danish Meteorologic Institute (DMI). The data pre-processing removes missing and negative energy usage measurements.

2.2. Integration of SPMS with the disaggregation method

This section presents the methodology used in this study to estimate the SH and DHW energy use in Danish dwellings when the data is truncated. The method is based on the previous methodology that was developed to disaggregate and estimate the SH and DHW energy use from the total heating measurements recorded by the SHMs [4,12]. However, the current work combines the disaggregation methodology with a data-recovering algorithm named: Smooth – Pointwise Move – Scale (SPMS) method. It retrieves the decimal values from the truncated measurements.

The first stage in the methodology involves the application of the SPMS algorithm. SPMS is a three-step algorithm with the following steps:

1. The data are smoothed using a linear weighted moving average with a centred window of length 5.
2. For each data point, it is ensured that the new value is within ± 0.4 kWh of the original data and that no value is negative.
3. The values obtained are scaled so that they accumulate over one day to the same amount as the original data.

Steps 2 and 3 are repeated in a loop until all conditions are met [14].

After the processed values are retrieved by the SPMS, the newly generated dataset is used in the disaggregation methodology. The method is based on the following assumptions: (1) the SH is in continuous operation, and (2) the DHW is sporadically produced and responsible for the peaks in the heat demand measurements. These assumptions are based on previous studies on the energy use patterns in Danish dwellings and are used to calculate the SH energy use from the disaggregated total heating measurements [6]. Following the initial assumptions, the method starts by categorizing the highest daily heating demand peaks as "SH and DHW combined usage" and the lower heating measurements as "SH usage". Subsequent to the categorization process, the method estimates the SH energy use by applying a combination of a Kalman estimator and support vector regression (SVR) while taking the outdoor conditions (temperature and global solar radiation) as inputs.

The results are then compared with the estimation when the DHW production measurements are perfectly detected (optimal data points detection).

3. Results and discussion

The proposed disaggregation algorithm was tested for four different cases: the original heating measurements with decimal values ("Decimal"), the truncated heating dataset ("Truncated"), the recovered dataset from the truncated data by the SPMS method ("SPMS"), and heating data where the segregation between data points with and without DHW production is performed perfectly ("Optimal"). The results of the application of the estimation method in the four cases are presented and discussed below.

Overall, the results of the experimental evaluation showed that the proposed energy estimation algorithm with the integration of the SPMS method outperforms the performance of the same method when using truncated data. As one can see in Figure 1, the NMBE and CVRMSE indicators display a significant performance improvement in the SH estimation from the truncated measurements to the SPMS recovered dataset. Nevertheless, this improvement is still lower than when using the original decimal dataset, corroborating the argument that, better than applying the SPMS method, the DH companies should address this problem by recording their heating measurements with a higher resolution.

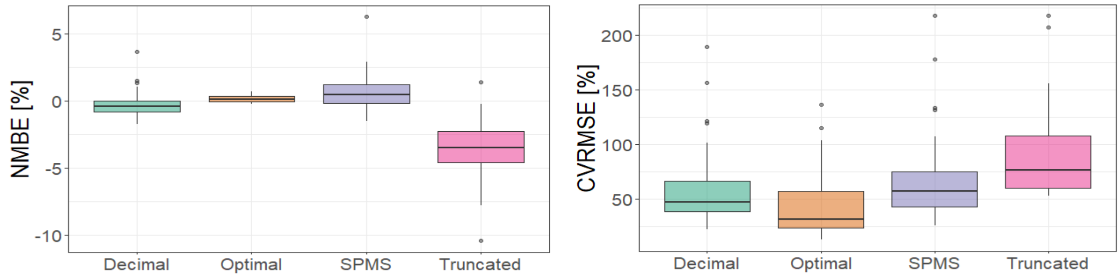


Figure 1. The NMBE and CVRMSE of the SH estimation of the different tested cases.

Regarding the estimation error of overall SH and DHW usage per building, it can be seen in Figure 2 that the best performance is given by the "Optimal" case, while the worst prediction scenario is for the truncated data.

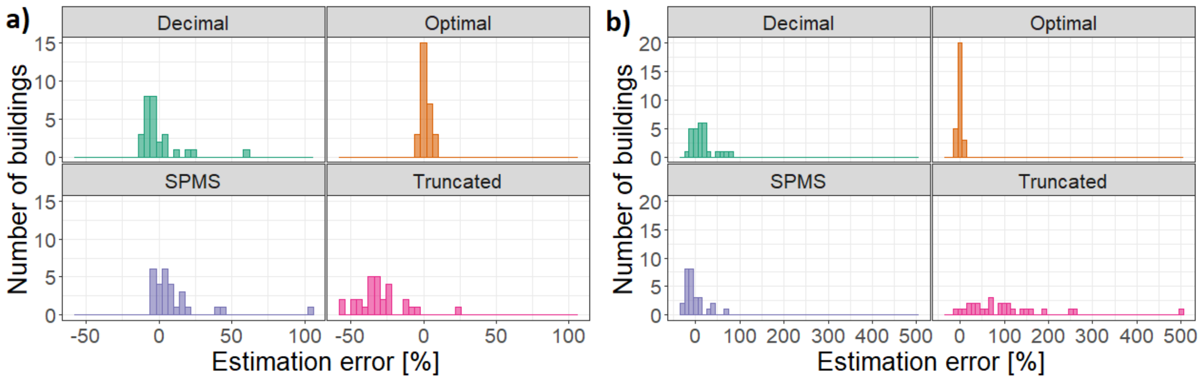


Figure 2. Estimation error for a) SH usage and b) DHW production.

Derived from the results in Figure 2, the Table 2 shows the minimum, median, and maximum error values for the prediction of SH and DHW usage.

Table 2. Minimum, median, and maximum error values per dataset case.

Case	SH			DHW		
	Minimum	Median	Maximum	Minimum	Median	Maximum
Decimal	-13%	-4%	60%	-17%	11%	83%
Truncated	-58%	-32%	23%	-6%	81%	497%

SPMS	-5%	4%	103%	-29%	-10%	73%
Optimal	-6%	1%	9%	-14%	-2%	12%

For SH, the decimal values case has a minimum error of -13% (underprediction) and a maximum error of 60% (overprediction). The truncated scenario has the highest negative error (-58% minimum), while the SPMS scenario has the highest maximum error (103%). The optimal scenario has the smallest range of error (-6% minimum to 9% maximum).

For DHW, the optimal scenario has the smallest absolute median error (-2%), while the truncated data case has the largest median error (81%). The truncated scenario also has the highest maximum error (497%), indicating a large overestimation of DHW energy use. The SPMS scenario has a negative median error (-10%), which means that it more frequently underestimates the DHW production. The optimal scenario has, again, the smallest range of error (-14% minimum to -2% maximum).

Overall, these results suggest that the choice of scenario can significantly impact the accuracy of energy use estimations for both SH and DHW. The results also prove disaggregation methodology can be improved further by developing new DHW usage timing identification algorithms.

4. Conclusion

This article presents the performance evaluation of a heating disaggregation algorithm coupled with a new method to recover (estimate) higher DH data resolution when truncated initially. Four different cases were tested, namely, the original decimal dataset, truncated dataset, SPMS recovered dataset, and the ideal scenario where the data points with and without DHW production are perfectly identified. The results show that the coupled methods outperform the truncated dataset in the estimation of SH and DHW energy use. However, the improvement is still lower than using the original decimal dataset. The optimal scenario performs the best for both SH and DHW estimation, while the truncated dataset case shows the highest median error and maximum error for DHW estimation. These findings highlight the significance of applying the SPMS when having truncated data or using these results to justify the recording of higher-resolution measurements by DH companies for more accurate building energy estimation. This work also emphasizes the potential for further development of the disaggregation algorithm by improving the identification methodology for hourly measurements with DHW usage.

There are several opportunities for further work on the proposed energy disaggregation algorithm. One suggestion is to identify DHW measurement points more accurately, possibly by using additional sensors such as cold-water meters. Another suggestion is to optimize the algorithm for larger datasets by exploring more computationally efficient estimation methods. Another opportunity is the development of a comprehensive evaluation of the present algorithm's performance by comparing it against other existing disaggregation methods (see Table 1) as it could provide insights into its suitability for different scenarios and datasets.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 893945 (E-DYCE). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 958345 (PRELUDE). The authors would like to acknowledge the Himmerland housing association and Varmekontrol A/S for their support in data collection and to all partners involved in the E-DYCE and PRELUDE projects. Also, the first author would like to express his gratitude to his colleagues in the Ph.D. “Corner Office” for the valuable discussions and support.

Credit-author statement

Daniel Leiria: Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Writing - original draft; Visualization. **Hicham Johra:** Conceptualization; Methodology; Resources; Writing - review & editing; Supervision. **Markus Schaffer:** Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Data curation; Writing - review & editing. **Anna Marszal-Pomianowska:** Conceptualization; Resources; Writing - review & editing; Supervision.

Michal Zbigniew Pomianowski: Conceptualization; Resources; Writing - review & editing; Supervision; Project administration; Funding acquisition.

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

References

[1] European Commission *Energy Performance of Buildings Directive* [Internet] 2022 [Accessed Jan 27] Available from: https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en
[2] IRENA, IEA and REN21 2018 *Renewable Energy Policies in a Time of Transition* IRENA, OECD/ IEA and REN21
[3] Volkova A, Mašatin V and Siirde A 2018 Methodology for evaluating the transition process dynamics towards 4th generation district heating networks *Energy* **150** 253–261
[4] Leiria D, Johra H, Marszal-Pomianowska A and Pomianowski M Zbigniew 2023 A methodology to estimate space heating and domestic hot water energy demand profile in residential buildings from low-resolution heat meter data *Energy* **263 B**
[5] Yliniemi K, Delsing J and Deventer J 2009 Experimental verification of a method for estimating energy for domestic hot water production in a 2-stage district heating substation *Energy and Buildings* **41** 169–174
[6] Bacher P, Saint-Aubain P Anton de, Christiansen L Engbo and Madsen H 2016 Non-parametric method for separating domestic hot water heating spikes and space heating *Energy and Buildings* **130** 107–112
[7] Alzaatreh A, Mahdjoubi L, Gething B and Sierra F 2018 Disaggregating high-resolution gas metering data using pattern recognition *Energy and Buildings* **176** 17–32
[8] Marszal-Pomianowska A, Zhang C, Pomianowski M, Heiselberg P, Gram-Hanssen K and Hansen A Rhiger 2019 Simple methodology to estimate the mean hourly and the daily profiles of domestic hot water demand from hourly total heating readings *Energy and Buildings* **184** 53–64
[9] Lien S, Ivanko D and Sartori I 2020 *Domestic Hot Water Decomposition from Measured Total Heat Load in Norwegian Buildings* BuildSIM-Nordic 2020 Oslo Norway OsloMet
[10] Hedegaard R Elbæk, Kristensen M Heine and Petersen S 2020 Experimental validation of a model-based method for separating the space heating and domestic hot water components from smart-meter consumption data *E3S Web of Conferences* **172**
[11] Ivanko D, Sørensen Å Lekang and Nord N 2021 Splitting measurements of the total heat demand in a hotel into domestic hot water and space heating heat use *Energy* **219**
[12] Leiria D, Johra H, Belias E, Quaggiotto D, Zarrella A, Marszal-Pomianowska A and Pomianowski M Zbigniew 2022 Validation of a new method to estimate energy use for space heating and hot water production from low-resolution heat meter data *E3S Web of Conferences* **362**
[13] Broholt T Hauge, Christensen L Rævdal Lund and Petersen S 2022 *Effect of Measurement Resolution on Data-Based Models of Thermodynamic Behaviour of Buildings* REHVA 14th World Congress CLIMA 2022 Rotherdam The Netherlands TU Delft OPEN Publishing
[14] Schaffer M, Leiria D, Vera-Valdés J Eduardo and Marszal-Pomianowska A 2023 *Increasing the Accuracy of Low-Resolution Commercial Smart Heat Meter Data and Analysing Its Error* 2023 European Conference on Computing in Construction Crete Greece (*Accepted*)

3.5 Influence of this research on subsequent studies

Based on this published research work [20-22], other researchers developed subsequent publications. In **Paper 2**, the main assumption to predict SH and DHW is by assuming that certain hours of the day measure SH and DHW usage while others only register SH usage. These hours that have combined measurements are assigned as the ones with the highest energy recordings in the day. This comes with the problem that what if these high values are only due to SH? To solve this issue, Schaffer *et al.* (2024) investigated the proposed disaggregation methodology using cold water meter data and tested this new information on other machine learning models [38].

While Ritosa, Saelens, and Roels (2024) investigated how non-intrusive monitoring and statistical models can assess household energy performance, with a focus on addressing discrepancies in Heat Loss Coefficient calculations through various data conditions and processing techniques [39]. To investigate such topic, the authors applied the disaggregation algorithm proposed in this Ph.D. work.

3.6 Further discussion

The method proposed in this chapter offers a novel approach for estimating SH and DHW usage using low-resolution, hourly data. This method can be used for setting a baseline for understanding energy usage patterns, yet it raises questions about the granularity and completeness of the data used. Specifically, one might wonder about the implications of having access to more detailed measurements. If data on indoor conditions – such as temperature, CO2 levels, and room-specific heat allocation – were available, it would significantly enhance our ability to dissect the contributions of DHW and SH to overall building heating performance.

Moreover, with more detailed data, the influence of occupant behavior on heating usage could be examined more closely. Variations in how individuals use heating systems within their living spaces are often significant but difficult to quantify with coarse data. A more complete dataset would allow for a detailed analysis of these behaviors, providing insights into how personal habits and lifestyle choices impact the building's energy usage.

Another topic that was investigated in Chapter 4 is the application of the energy signature (ES) model and how more detailed data can be used in such a model. Addressing these possibilities, this research sets the stage for a deeper investigation into the interactions between building energy performance and occupant behavior.

Chapter 4. Integrating smart heat meters and indoor sensors data

This chapter discusses the third key finding of the Ph.D. project, which focuses on leveraging data from smart heat meters (SHM) to gain deeper insights into users' behavior and heating practices by integrating indoor sensors data. These indoor sensors refer to temperature, CO_2 concentrations, humidity, window openings, radiators heat allocators, and SH and DHW submeters.

4.1 Indoor analysis: Unlocking insights into heating performance

This chapter attempts to answer some of the questions from Chapter 3 by integrating more detailed measurements into our analysis. Specifically, it is considered how access to comprehensive indoor condition data – such as temperature, CO_2 levels, humidity, and room-specific heat measurements – could transform our understanding of the interplay between DHW and SH contributions to overall energy efficiency.

Furthermore, the role of occupant behavior in shaping energy usage patterns is a complex variable that is often difficult to quantify without detailed data or prior knowledge of their routines. As it is expected that variations in individual heating practices within different rooms can also significantly influence the overall energy of a building. Thus, besides incorporating the indoor sensors, a questionary and talks were made with the different tenants for better assessment of this analysis.

In this work, it was also investigated the potential of adjusting the current energy signature (ES) model from its linear nature to a sigmoid function. As such proposal even though small regarding its model complexity and accuracy could increase the explainability between building heating usage and the outdoor temperature for colder outdoor temperatures, thus paving the way for more targeted and effective energy management strategies.

4.2 Leveraging indoor sensors to amplify SHM data

Paper 5

"From showers to heaters: Evaluating the different factors that play a role in buildings' energy signature"

Daniel Leiria, Hicham Johra, Yue Hu, Olena Kalyanova Larsen, Anna Marszal-Pomianowska, Martin Frandsen, Michal Zbigniew Pomianowski, submitted in *Energy and Buildings*, 2024.

From showers to heaters: Evaluating the different factors that play a role in buildings' energy signature

Daniel Leiria[1], Hicham Johra[1], Yue Hu[1], Olena Kalyanova Larsen[1], Anna Marszal-Pomianowska[1], Martin Frandsen[1], Michal Zbigniew Pomianowski*[1]

[1]Aalborg University, Department of the Built Environment, 9220 Aalborg, Denmark

*Corresponding author: mzp@build.aau.dk

Abstract:

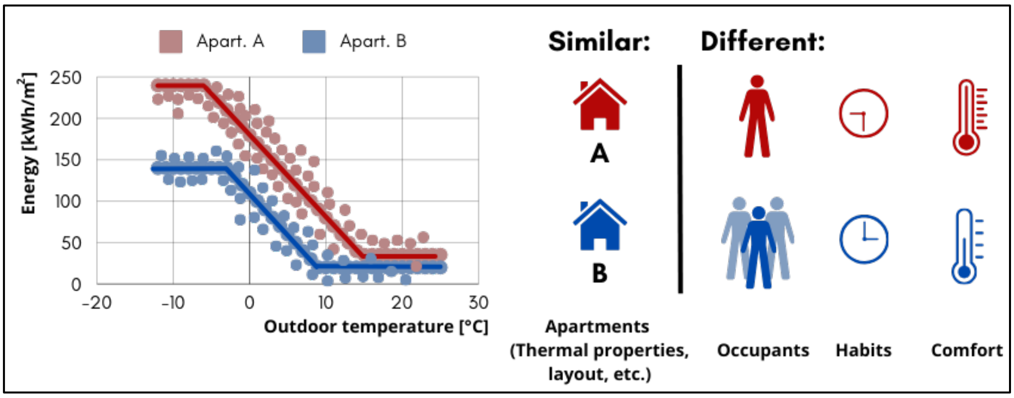
In the realm of energy sustainability, accurately evaluating the impact of domestic hot water (DHW) production on energy signature (ES) models is critical. Traditional approaches have often overlooked or simplified this aspect, potentially skewing efficiency assessments. This study aims to rectify this by dissecting the influence of DHW on ES models and exploring the occupancy heating habits that shape energy usage profiles. Employing a dataset from a small number of Danish apartments, this work delves into the differential energy impacts of DHW production and space heating (SH) operations on the ES model. The study investigates the ES model's applicability across similar apartments and examines whether SH operational patterns significantly alter the model's predictive accuracy. The findings reveal that the incorporation of DHW production into ES models is necessary but not as significant as someone might assume. Occupant behavior, particularly concerning SH habits, seems to be one of the most influential indicators in shaping the ES. Moreover, the study presents the sigmoid ES model as a robust alternative to the prevailing linear ES model, elucidating the reasons behind its distinct asymptote points. The research concludes that the sigmoid ES model displays an improvement in capturing the complexities of building energy dynamics. These insights pave the way for more refined energy assessments, with implications for policy-making and energy management in residential buildings.

Keywords: Linear and sigmoid energy signature, Domestic hot water, User-driven energy usage, Space heating efficiency, Indoor sensor integration.

Highlights:

- Analyzed DHW usage reveals stochastic patterns with consistent median levels.
- ES parameters vary significantly across similar buildings and spaces.
- Occupant behavior impacts heat performance more than thermal characteristics.
- Window opening behavior seems to affect the upper asymptote in the ES model.
- Integrating sensor data provides deeper insights into energy usage patterns.

Graphical abstract:



Introduction

The impact of climate change is being felt across the globe, with rising temperatures, extreme weather events, and the loss of biodiversity. To mitigate these effects and ensure a sustainable future for generations to come, the reduction of carbon emissions and energy usage is crucial. In the European Union, the energy share in buildings accounts for 40% [1]. Due to this large share and knowing that people spend most of their time in their homes [2], a deeper knowledge of the building's energy usage is required. On the Danish side, energy efficiency has been under much focus, and it has been one of the main drivers of building regulations [3]. Energy-efficient buildings have lower energy usage and operating costs and are designed to secure the health and comfort of occupants. However, to attain these benefits, accurate and reliable methods for energy analysis are necessary. This analysis allows building owners and designers to understand how energy is used within a building, identify areas of inefficiency, and develop strategies to improve the building's energy performance. When considering the analysis frameworks in civil engineering, models with different accuracy and complexity are needed for different applications: e.g., early-design parameter optimization, model-predictive control, energy demand baselining, urban-scale modeling, and energy performance assessment, to name a few. Nevertheless, the selection of an appropriate method depends on several factors, e.g., the building's characteristics, energy use patterns, or other specific project objectives. Therefore, the correct choice of suitable, accurate, and reliable methods is essential to ensure that energy-saving measures are effective and provide tangible benefits.

One of these methods is the Energy Signature (ES) model, which is based on the analysis of the correlation between energy usage in buildings and outdoor temperature. This model due to its simplicity and integrability in different analysis frameworks is one of the most applied in the academic and industrial context. Moreover, since the early 1900s, this method has been used [4], and further applied in academic articles, and improvements have been discussed and proposed on how to use this model accurately [5]. However, due to the increasing availability of smart energy meter data, this methodology has been applied much more for the operational energy assessments, and it has been combined with other computational simulations, and also compared with more complex methodologies [6–8]. This study examines the application of the ES model, paying particular attention to certain characteristics that in the authors' opinion, have not been thoroughly explored or have been simplified in previous research

Background

The ES model, also known as the change-point model, expresses the buildings' energy demand as a function of the outside temperature, according to the ASHRAE guideline 14 [9]. In the guideline, it is explained that this model is steady-state and can have from one to several representative parameters, depending on the HVAC systems present in the assessed building. According to the guidelines, it is a useful tool for understanding how a building's energy use changes with the weather, and for identifying opportunities for energy savings. The ES can use different types of energy outputs (e.g., electricity, cooling, heating, etc) and it can also show the operation of different systems working simultaneously or a singular system operating with different operation settings. In this study, we focus on the ES model applied to residential buildings with water-based heating and without cooling systems. The representative parameters that define the ES for only heating systems are seen in Table 1 and represented in Figure 1.

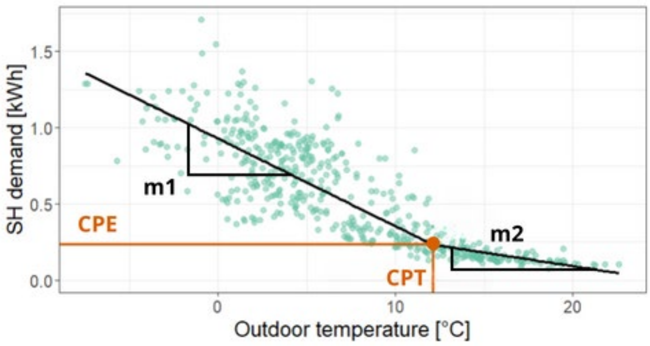


Figure 1: Representation of the ES model for heating purposes.

Table 1: Main characteristics and definitions of the ES model.

ES parameters	Definition
Heating season slope (m_1)	Slope of the ES model concerning the heating season.
No heating season slope (m_2)	Slope of the ES model concerning the no heating season.
Change-point temperature (CPT)	Outdoor temperature limit when the seasons change.
Change-point energy (CPE)	Energy demand limit when the seasons change.

Using the parameters presented in Table 1, the model is defined by a set of two equations corresponding to a continuous piecewise linear function over two distinct ranges, E_h and E_{nh} refer to the heating and no heating seasons, respectively:

$$\begin{cases} E_h = CPE + m_1(CPT - T_{out}), & when\ T_{out} \leq CPT \\ E_{nh} = CPE + m_2(CPT - T_{out}), & when\ T_{out} > CPT \end{cases} \tag{1}$$

In Denmark, a larger portion of residential heating systems are based on a substation connected to the district heating network that fulfills the combined demand of space heating (SH) and domestic hot water (DHW) production. The SH is the energy needed to heat the building's interior and is mainly driven by the dynamic heat balance between the building's thermal inertia, heating and ventilation output systems, miscellaneous heat gains (e.g., solar, people, equipment, etc.), and heat losses (e.g., transmission and ventilation). The DHW is the energy used to heat water for bathing, washing, and other domestic purposes. The DHW demand throughout the day is quite stochastic and can increase significantly during peak usage times, such as in the morning and evening. Due to this, when plotting the ES relation, it is always assumed that the main reason for outliers is the randomness and different intensities inherent in DHW production due to the direct influence of the occupants on tapping usage [10]. However, literature also shows that when taking the ES model, often the DHW is accounted as constant [11]. Thus, this opens the question, of whether the DHW is as constant or irregular as some theorize when applied to the ES. Therefore, this is the first topic that this manuscript attempts to address.

Regarding the ES overall methodology, several research articles in the literature have applied this model to investigate the following topics:

- SH and DHW estimation [11,12].
- The impact of the outdoor temperature on the building's energy use, by analyzing the slope of the linear trend [13-17].
- The potential for energy savings through refurbishment initiatives applied in buildings [18-20].
- Fault detection applications of heating systems [21-23].
- Application of the ES to large urban areas to investigate the building stock [24-26].

Regarding the last topic, some studies have focused on characterizing buildings' thermal characteristics based on the ES model values described in Table 1. However, this opens another question – are these ES properties similar in alike buildings and spaces? This question is also addressed in this study, by analyzing the ES performance for four similar residential apartments with the same construction and thermal characteristics.

Lastly, other research contributions have focused on the specificities of this model itself, where a few works investigated the following key ideas:

- Pros and cons of the ES model [5]
- Improvement of the robustness of the ES model [27-31]
- Proposing another type of ES model based on a sigmoid function [8,32,33]

Regarding the last listed point, these articles proposed that the ES model should be a sigmoid function instead of a linear change-point model. In [33], it is discussed that the sigmoid behavior is observed, and it is theorized to be due to significant changes in venting/ventilation routines by the users or by the supply temperature limitations when reaching extremely low outdoor temperatures. Therefore, another question is raised, regarding the reasons behind the horizontal asymptotes for data points in the sigmoid model. As known, the lower asymptote point is due to a change of heating/no heating seasons, analogous to the CPT in the linear ES model. However, the reason for the upper asymptote point (at low outdoor temperatures) is still unclear. Therefore, this study also attempts to shed some light on the reasons behind the upper limit point when the outdoor temperatures are low using other sensor data retrieved from the measurement campaign [34]. The sigmoid function is represented and discussed further in the Methodology section.

Contributions and novelty of the current study

Due to this combination of different heating demands (SH+DHW) that coexist in the ES representation, this manuscript focuses on the analysis and effect of the DHW production on the ES methodology. As theorized, the DHW energy usage must have an impact on the overall ES model. Until now, this impact was neglected or set constant when applying this methodology. However, for the research's sake, this influence must be tested, which can only be done, if a separate measurement of DHW and SH is available. Also, this work attempts to address the different user heating habits and their influence on the ES model. Lastly, the sigmoid ES model is proposed, analyzed, and compared with the current linear ES model while addressing its properties.

Therefore, the contributions of this paper are the following:

1. The application of the proposed methodology in a small dataset of Danish apartments to evaluate the impact of including/excluding the DHW energy use on the ES model results.
2. To evaluate the applicability of the ES model for the assessment of overall building/apartment heating efficiency and fault detection potential.
3. To investigate and discuss the suitability of the sigmoid ES model in building energy assessment.

Outline

Following the *Introduction*, the *Methodology* section describes the dataset utilized in this work and the assessment methodology developed to answer all questions raised above. The results from the assessment are examined in the section *Results and Discussion*. The manuscript closes with *Conclusions* and *Suggestions for Further Work*.

Methodology

Building case description

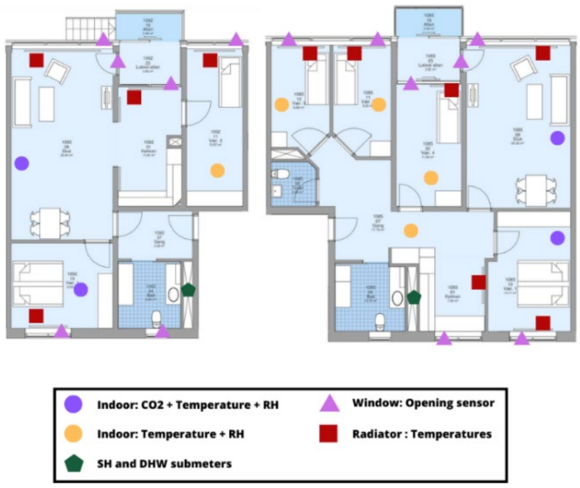
The measurements dataset comes from the Danish demo case of the European project E-DYCE situated in Frederikshavn, Denmark [35]. The overall building was built in 1972 and was renovated in 2011, and encompasses a total of 28 apartments, covering a heated space of 3,120 m^2. It is divided into three different addresses (staircases). Each staircase spans four or six floors, housing two apartments per floor. The present study investigates four apartments on the same staircase where two of them are positioned on the first floor, one on the second floor, and one on the ground floor. In 2021, an energy performance certificate was issued for the building, assigning it an energy label of B. This classification indicates that the building's total energy usage for all appliances is expected to be below 71 kWh/m^2 per year.

Regarding the apartment's heating systems, the SH operates through a central mixing loop powered by district heating, located in the basement's technical room. While all apartments receive the same temperature from this loop, however variations in the actual temperature may occur due to pipe distribution losses. Additionally, each stairwell has mechanical ventilation equipped with heat recovery and a heating coil, also drawing from the main mixing loop. The centralized hot-water production is implemented through a heat exchanger that feeds a circulating loop connected to all apartments. Every apartment is equipped with a water flow meter that measures water consumption. Both the central SH and DHW mechanisms are managed by a Danfoss ECL 310 controller. The building has a control system, called PreHeat [36], which is a self-learning forecasting algorithm implemented in the building's system to save energy. It does this by accumulating knowledge about the home's thermal properties, such as the degree of insulation, the influence of solar radiation, wind, and outside temperature, as well as the household's user behavior. By combining this information with the weather forecast, the control system predicts the apartments' heating demand.

Measurements dataset

The measurements campaign consists of a set of hourly measurements at the apartment level and the room level. Concerning the apartment level, two submeters were installed, one for SH and another for DHW usage. The SH meter measures the hourly flow, supply, and return fluid temperatures. The DHW meter measures hourly flow and supply temperature.

At the room level, it is installed sensors to measure the indoor climate conditions (e.g., indoor temperature, relative humidity, CO_2 concentrations, and window opening rate). Additionally, all radiators were attached with contact temperature sensors to monitor the radiator's heat output throughout the campaign (supply, return, middle of radiator). The measurement period is from the 23rd of September 2021 to the 18th of June 2023. In Figure 2, one can see two of the apartment's layouts with the location of the different sensors combined with a table that shows the different units of the measurements. In Figure 3, one can see two types of sensors installed in one of the apartments.



Level	Shape (color)	Parameter	Units
Apartment	Pentagon (green)	SH flow	m^3/h
		SH supply/return temperature	ºC
		DHW flow	m^3/h
		DHW supply temperature	ºC
Room	Circle (purple or yellow)	Indoor temperature	ºC
		Relative humidity	%
	Circle (purple)	CO2 concentration	ppm
	Triangle (pink)	Window opening rate	%
	Square (red)	Surface pipe supply temperature	ºC
		Radiator surface temperature (middle)	ºC
		Surface pipe return temperature	ºC

Figure 2: Measurement campaign – description (the sensor location is not representative).

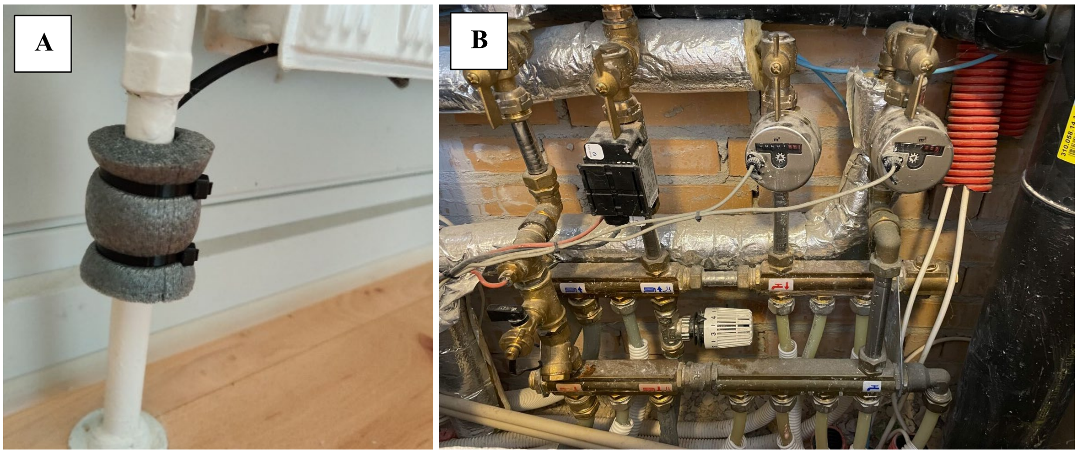


Figure 3: A) Contact sensor mounted on a radiator pipe in the apartment. B) SH and DHW apartment submeters installed in a shaft.

Users' heating habits

To better understand the data retrieved from the sensors' measurements, it was decided to collect additional insight about the occupants' habits. To retrieve this information, a phone survey was conducted with the apartments' occupants. In Table 2, a summary of the different user interactions with their apartment systems is presented. More details regarding these interviews can be found in [34].

Table 2: Occupancy heating habits in each apartment.

	Apt. A	Apt. B (B.1)	Apt. C	Apt. D
OCCUPANTS				
Nr. of adults	1	3	2	1
Nr. of children	0	2	0	0
Weekly occupancy	Not at home from 9-12h and Thursdays from 12-15h	Adults are always at home. Children at school from 8-15h	Always at home	Out of apartments in the afternoons
AIR QUALITY				
Which rooms are vented?	All rooms	All rooms	Bedroom and bathroom	Bedroom (every day) and living room (summer only)
How long and when do you vent the apartment?	Each day bedroom, bathroom, and living room 2-3 times a week. Long venting in summer (every day) and short in winter (10 minutes)	Summer: All day Winter: 1-2 h in the morning	Summer: All day Winter: 1-2 h per day	Summer: All day Winter: 3-4 h in the morning
THERMAL COMFORT				
What is the setting on the radiators' thermostats?	Bedroom is set to be cold (setting 1). Bathroom set on 2.	Different settings in the rooms.	Only the radiator in the living room is open (setting 4-5). Underfloor heating in bathroom (operating).	All radiators are set on 3. Radiator not used in the bedroom. Bathroom underfloor heating is always in use.
Is the temperature in the apartment uniform?	-	Yes, except for one bedroom (where no heating is used)	-	-
ENERGY SAVING MOTIVATION				
Do you pay too much for energy?	Yes	Yes	No	No

Regarding the collection of the data, one should note that there was a change of dwellers during the measurement campaign in apartment B. Therefore, throughout the results and discussion, this apartment is analyzed concerning the different dwellers' periods. Where the first occupants are denoted as B.1 and the occupants that moved after are denoted as B.2. The period when the apartment was vacant, is removed from the analysis. In Table 2, the survey responses for apartment B are associated with the family that resided there prior to moving out (B.1).

Data pre-processing

Regarding the dataset pre-processing, it was ensured that the raw data was transformed and cleaned into a format suitable for further analysis. In this stage, the different sensor hourly measurements were combined into two different datasets. All the data concerning the apartment level (SH and DHW energy submeters) are combined in a single dataset. While all data from indoor sensors installed in the different apartments' rooms (room level) are merged into a second dataset. The next step of the data treatment process is handling missing measurements. The submeters data are cumulative values, therefore the imputation technique applied was a linear interpolation. While "point-in-time" (instantaneous) measurements (e.g., indoor sensors) are imputed using a linear moving average. Lastly, the hourly measurements were aggregated into daily values. The daily values were the ones used throughout this analysis.

Energy signature and its dependence on DHW usage and occupancy SH habits

In the context of analyzing the linear energy signature model concerning DHW usage, the application of a segmented regression approach was used in this article instead of the combination of two linear regressions. This approach, by applying the segmented regression from [37,38] enables the model to account for multiple variables and breakpoints, which in this case, have two linear segments and one breakpoint. This is particularly relevant for DHW usage, where the rate of energy use could vary at different temperature intervals. Such breakpoints signify where the relationship between total energy usage and external temperature shifts, indicating a larger share of DHW usage in the warm periods (e.g., summer) versus a lower share in the cold periods (e.g., winter) due to the simultaneous operation of DHW systems with SH emitters.

Concerning the linear ES model, the impact of the SH systems' interaction with the users was also analyzed. To achieve this, the contact temperature sensors' measurements attached to the radiators in each room of the apartments were investigated. These sensors provided valuable insights that offered an overview of the heat distribution within each apartment. By visualizing this data, it was possible to discern the heat allocation, which in turn enables to understand the similarities or differences in the ES of each apartment.

Sigmoid energy signature

The concept of the sigmoid ES offers a different approach than the linear model to evaluate energy usage in buildings. This method hinges on the sigmoid function, a mathematical S-shape curve that is represented by two asymptotes and a linear trend in between. In the context of building energy assessment, the sigmoid ES function is expected to behave more robustly and, therefore not influenced by outliers and different heating behavior by the dwelling's occupants. Also, it might provide a nuanced way to understand how energy usage patterns change over time in extreme weather conditions (i.e., colder and warmer outdoor temperatures). The mathematical expression that defines the sigmoid ES is given by:

$$E(T_o) = ae^{-e^{cT_o - b}} + d \tag{2}$$

Where the parameters a, b, c, and d affect the function as follows:

- "T_o": Measured outdoor temperature.
- "a": This parameter scales the maximum value of the sigmoid function. It determines the upper asymptote of the curve. If "a" increases, the height of the S-curve increases accordingly. In the context of building physics, it represents the maximum energy demand for heating as the outdoor temperature decreases. As the value of "a" increases, the potential maximum heat loads the building might require in extremely cold conditions also increases.
- "b": This parameter shifts the curve along the outdoor temperature (T_o) axis. If "b" is increased, the S-curve asymptote x-point (CPT) moves to the right (higher T_o-value).
- "c": This parameter controls the steepness or slope of the curve. A higher value of "c" makes the transition from the lower to the upper asymptote steeper, meaning that the change occurs more rapidly. If "c" is smaller, the transition is more gradual. The parameter is related to the building's insulation properties; poor insulation would result in a steeper increase in heat load with a decrease in outdoor temperature.
- "d": This parameter shifts the curve vertically. Increasing "d" will move the entire curve upwards without changing its shape. This affects the lower asymptote of the curve, setting the minimum value that the energy baseline can take, reflecting the base load that is constant regardless of outdoor temperature. Moreover, "d" represents the DHW daily baseload during warmer days, i.e., summer.

An example of the sigmoid of the curve can be seen in Figure 4, with the representation of the different parameters and how they impact the S-shape function. More details of this function can be consulted in [8].

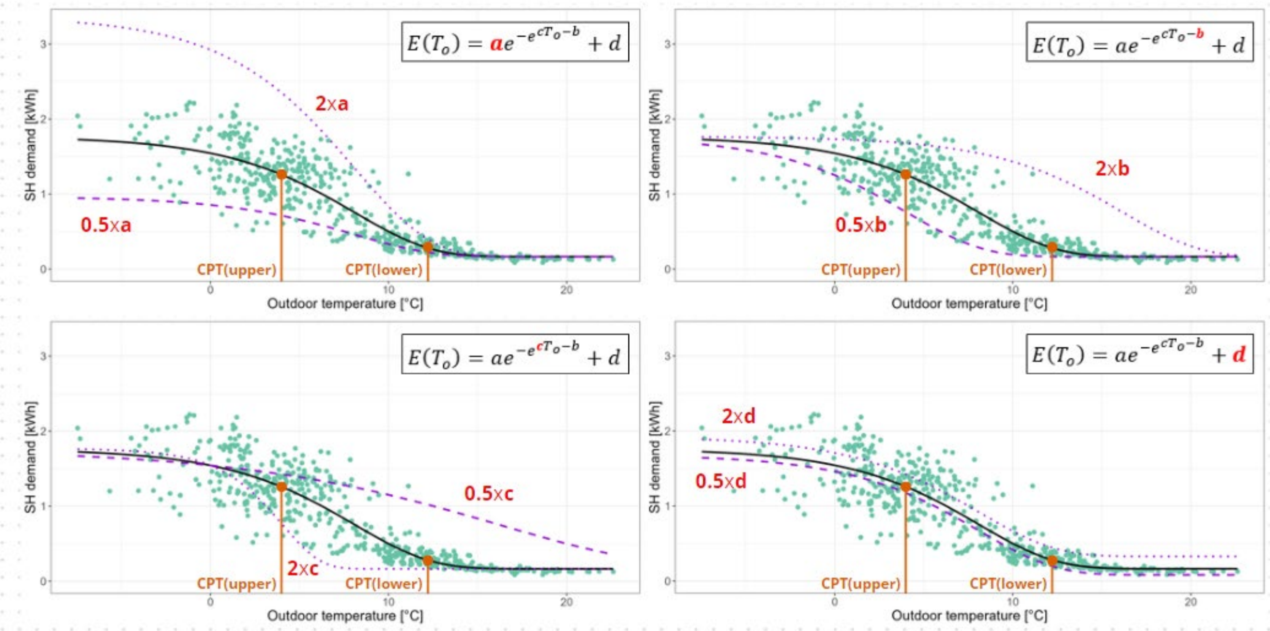


Figure 4: Sigmoid ES model - Representation of its parameters.

Similarly, with the linear energy signature, this function also has change points that can be obtained through equation (3), as shown in the appendix of this manuscript.

$$\begin{cases} CPT_{upper} = \dfrac{1}{c}\ln\left[\dfrac{1}{2}e^{b}(3-\sqrt{5})\right] \\ CPT_{lower} = \dfrac{1}{c}\ln\left[\dfrac{1}{2}e^{b}(3+\sqrt{5})\right] \end{cases} \tag{3}$$

From the equation above, it is observed that the CPT values are only dependent on the parameters “c” and “b”, from equation (2).

Results and Discussion

In the ensuing analysis, we delve into the empirical findings obtained from our assessment of energy demand within the studied apartments. The results underscore the complex interplay between SH and DHW usage, revealing distinct temporal patterns and consumption behaviors. This section aims to break down these findings, drawing on the sigmoid ES model to elucidate the underlying trends and anomalies in the data.

Linear energy signature

The first part of this work is the comparison of the ES model when using the total energy demand (SH+DHW) and the SH energy measurements. However, before, it is necessary to visualize the daily time series of both SH and DHW energy usage for the full extent of the measurement campaign to have a full grasp of ITS values (see Figure 5).

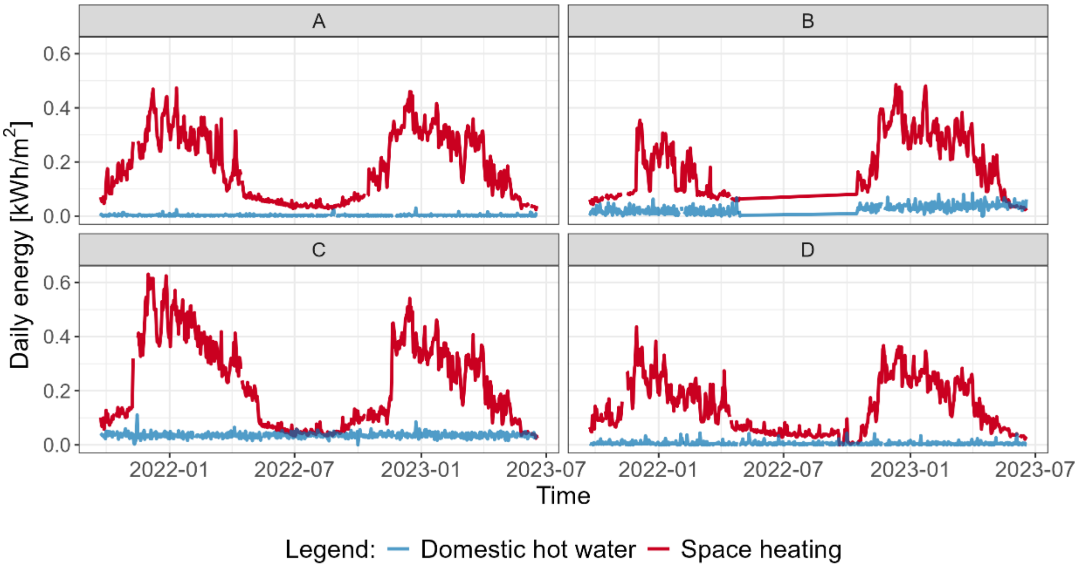


Figure 5: Time-series data from SH and DHW energy measurements.

From Figure 5, one can see that the SH produces a certain seasonality throughout the months, due to a variation of the heating needs due to outdoor temperature changes, while the DHW presents a more stochastic behavior, however, the observed DHW oscillations seem to fluctuate around a consistent average. To better visualize this constant trend of DHW usage, it is plotted in Figure 6, the distribution of the daily DHW volume consumption for each month as a boxplot. The dark dots in the plot are the daily measurements that deviate significantly from the measurements interquartile (outliers).

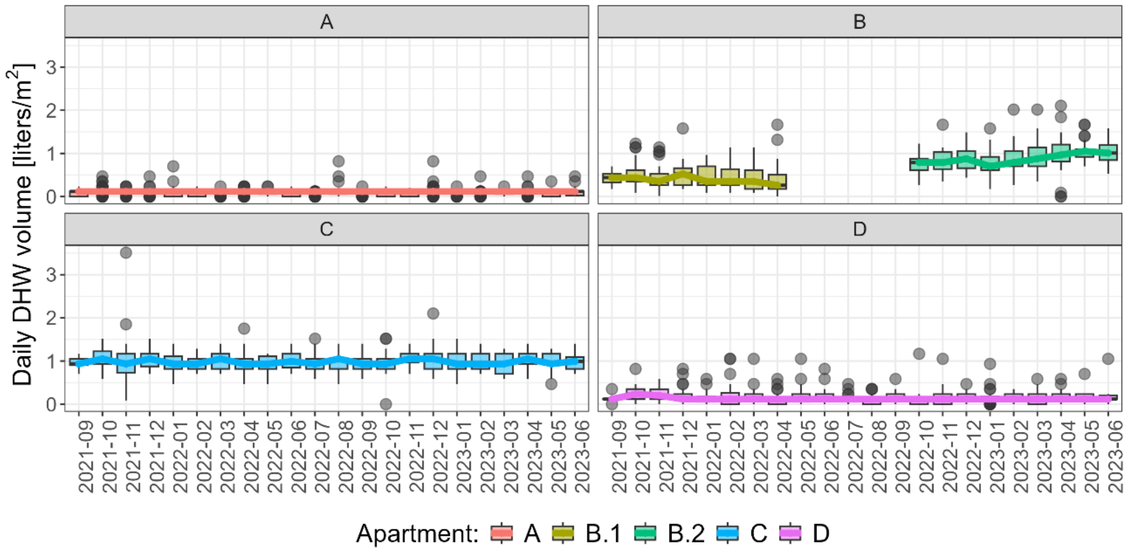


Figure 6: Daily DHW volume consumption per square meter per month.

By visualizing the monthly trend in Figure 6, it becomes clearer that all the apartments exhibit a constant monthly median DHW usage pattern when aggregating the measurements in a lower time resolution. In apartment B, however, there was a change of dwellers during the measurement campaign. After September, a new family moved in with larger DHW demands than the previous one. Another conclusion that can be drawn from this plot is that the magnitude and variation of this trend seem to be based on the number of dwellers in each apartment. As observed, the baseline for both apartments A and D, which have a single dweller each, is much lower than that of the other apartments. Therefore, it can be concluded that the DHW patterns are greatly influenced by the number of occupants and their routines. Nevertheless, apartment C, with two dwellers, shows similar high energy usage for DHW production as apartment B.1 did when it housed five people. This example demonstrates that in some cases, a larger number of occupants in a building does not directly indicate higher DHW consumption.

Focusing on the energy signatures generated from the SH demand alone, one can see in Figure 7, that apartments A, C, and D show distinctly the two seasons due to the change in heating demand. While apartment B has two noticeably different heating season trends, where the smallest heat usage (B.1) regards the family with five dwellers. When applied the same change-point model to the total energy usage (SH + DHW), the final result is plotted in Figure 8.

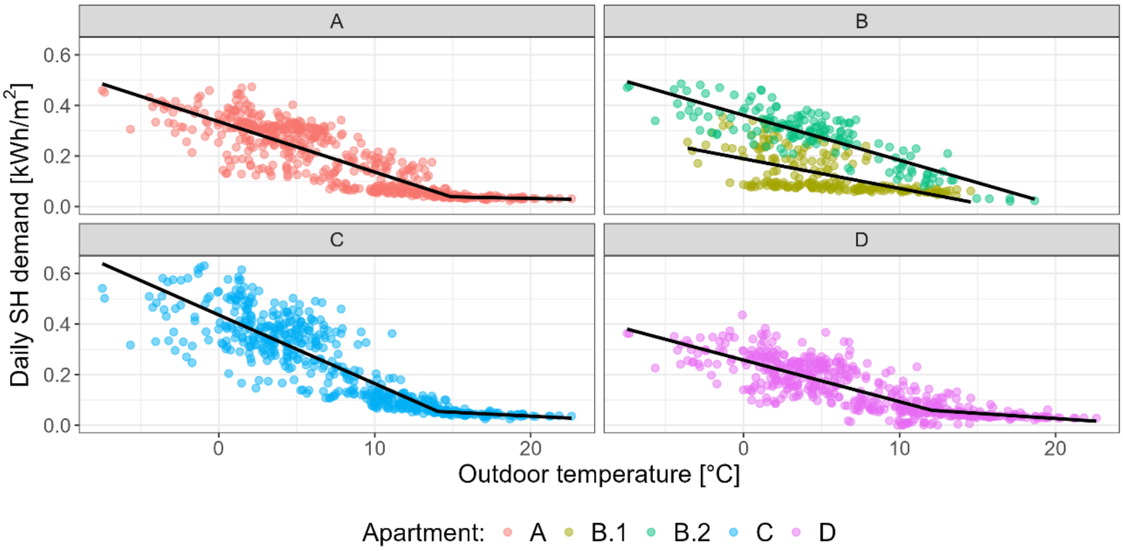


Figure 7: SH energy signature for each apartment.

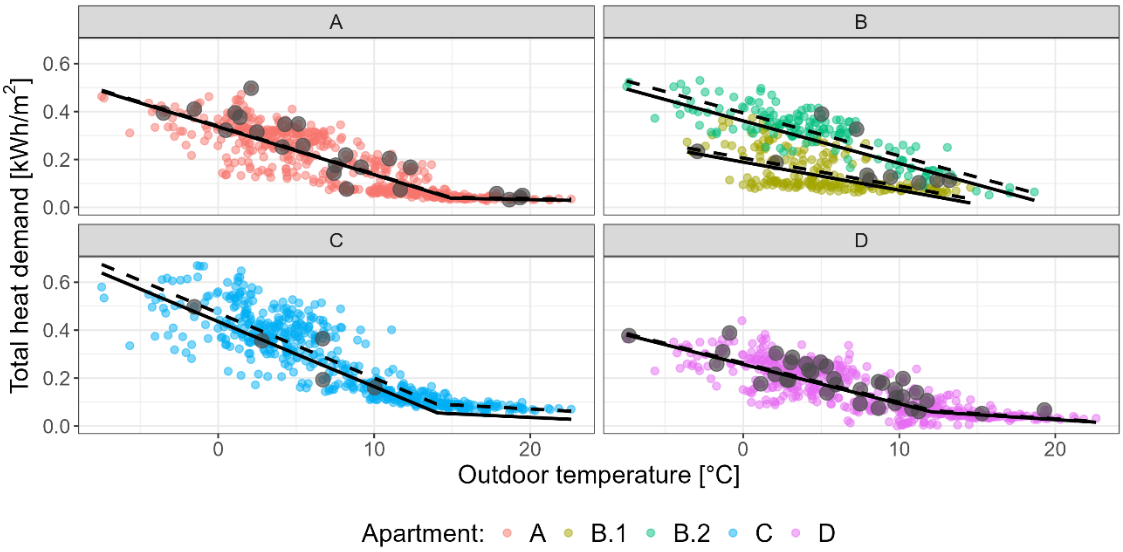


Figure 8: SH (solid line) and Total (dashed line) energy signatures for each apartment. Dark grey highlighted points are measurements assigned as outliers due to the DHW demand (see Figure 6).

In Figure 8, one can observe the SH (solid line) and Total (dashed line) energy signatures for each apartment. While all data points are the total heat measurements. The grey highlighted points indicate measurements identified as outliers due to the spikes in DHW demand as identified in Figure 6. Such outliers, as it seems do not represent the outliers in the total heating measurements. As one can see in Figure 7 and Figure 8, the four analyzed apartments display different linear ES models from each other. This result does not follow the usual hypothesis that similar buildings/spaces should display alike thermal characteristics and thus similar ES. Therefore, these plots show that the building ES properties (e.g., slope, CPT, and CPE) can be different for similar buildings (alike thermal characteristics). The reason for such differences can only be due to occupancy behavior (i.e., different heat setpoints and occupancy presence) or the presence of faults in SH systems inside the household (i.e., unbalanced radiators, broken thermostatic valves, etc.). To understand, how much is the variation of the ES parameters for the different cases, the plot in Figure 9 was generated.

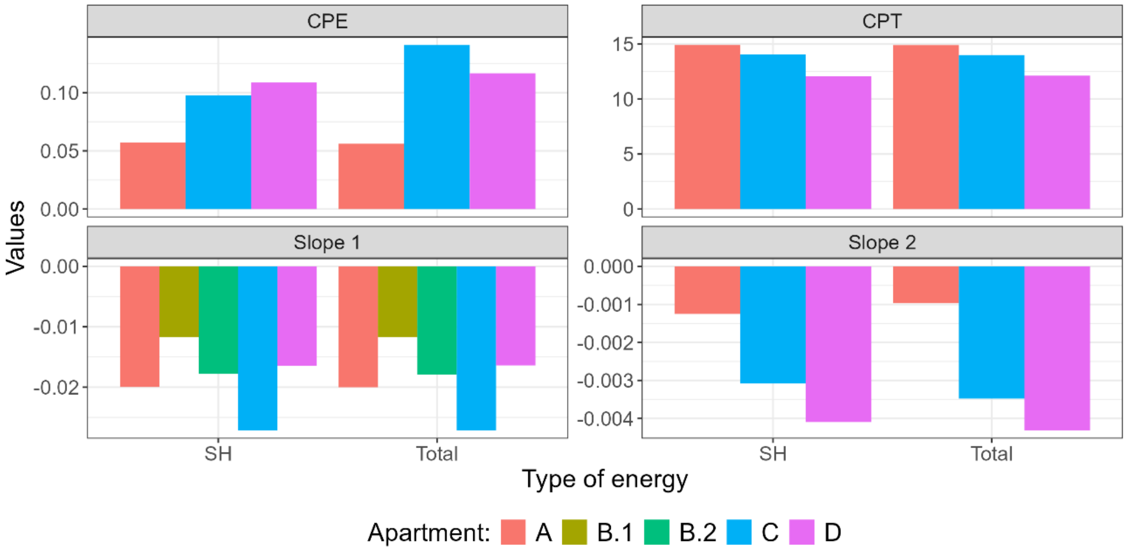


Figure 9: Linear ES parameters of SH and total heat (SH+DHW) demand per each apartment.

Figure 9 shows the ES parameters for different apartments, highlighting the individual energy usage patterns in terms of SH and total energy. The grouped bar charts compare the CPE, the CPT, and two slopes that measure the rate of change in energy demand concerning heat demand seasons (heating and no heating needs). The CPE bars illustrate notable differences among the apartments, a metric that reflects the energy levels at which the heating demand shifts, signaling the apartment's heat baseline. Variations in CPE suggest that each apartment's heating system might engage differently at the DHW production level and minor SH systems (underfloor heating in the bathroom) during summer. The slopes, labeled slope 1 (heating season – m_1) and slope 2 (no heating season – m_2), quantify how energy demand escalates or declines as temperatures rise or fall. A steeper slope would imply a more sensitive change in energy demand with temperature shifts, potentially influenced by the thermal losses (transmission and ventilation) of each apartment. Slope 1 shows a small change between the SH and total per apartment, however, there are significant changes when comparing the same energy demand type between apartments. Slope 2 has similar flagrant differences for the same reasons behind the slope 1 values.

As one can see, the plot shows that in some cases the linear parameters do not change significantly when assessing the dwelling energy demand if using SH or the total demand, where the best case to see this is slope 1. However, when comparing the same energy type for each apartment, these values differ much more. This result is counterintuitive because the most common hypothesis is that similar buildings must thermally behave alike. Because these apartments have similar thermal characteristics (e.g., U-values), architecture, and heating systems, and still display such significant results, one can conclude that other factors are affecting the overall heat performance of the apartments. To better investigate the possible factors noted above, it was investigated the SH and DHW efficiency and profiles with distinct heat meters and interviews with the dwellers of each apartment.

Human-building interaction with the heating systems

Despite the initial hypothesis that apartments with similar physical structures and heating systems would exhibit similar thermal behavior, the observed discrepancies in the energy demand data compel us to look into the heating practices of each apartment's dwellers.

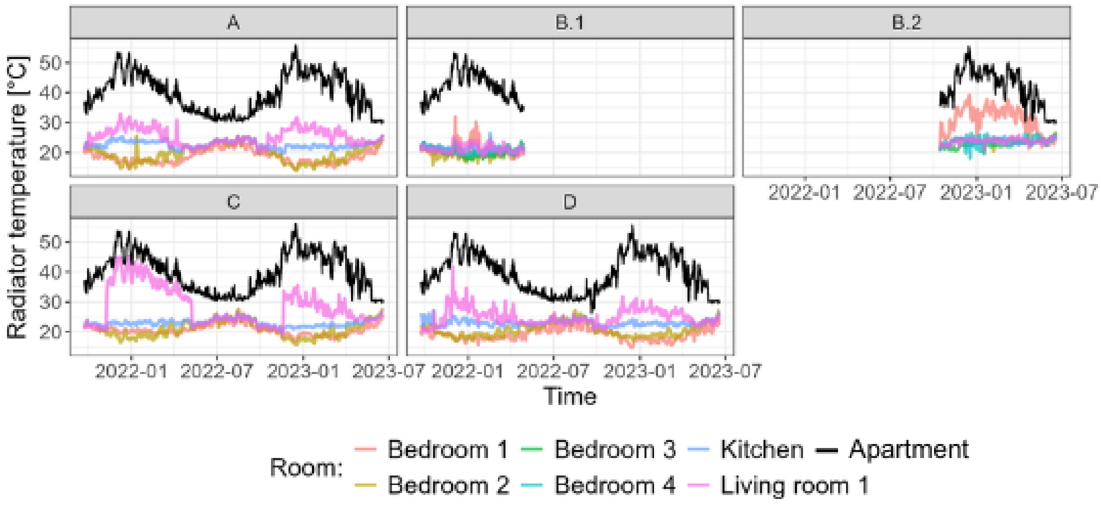


Figure 10: Radiators' temperature per room (colored) compared with the apartment SH supply temperature (black).

Figure 10 presents a comparative analysis of radiator temperatures across different rooms within various apartments, measured over time from January 2022 to just before July 2023. The data shows a seasonal fluctuation in temperatures with peaks suggesting increased usage during the colder months for SH purposes. It is evident that the hot fluid supply temperature of each apartment, marked in black, is set higher than the radiators' temperature in individual rooms. Each colored line represents a different room, with some, like bedroom 1 and the living room, showing consistently higher temperatures, which could point to higher heating setpoints due to the preference for warmer temperatures in these spaces. For instance, bedroom 2's lower temperature readings, as compared to bedroom 1 (in some of the dwellings), reflect less frequent use. The subplots labeled A, B.1, B.2, C, and D represent different apartments where the measurements were taken, showing that the heating profiles vary not only within an apartment but also across different apartments showing that varying occupant interaction with their heating systems are most likely the cause of variations in their ES. Periods of potential overconsumption are suggested in the winter months, where the radiator temperatures are significantly similar to the SH supply temperature (as seen in apartments C and B.2), hinting at inefficiencies in heating distribution or control. This highlights the need for potentially adjusting the heating system to better match the actual demand in each room and apartment to optimize energy efficiency and comfort.

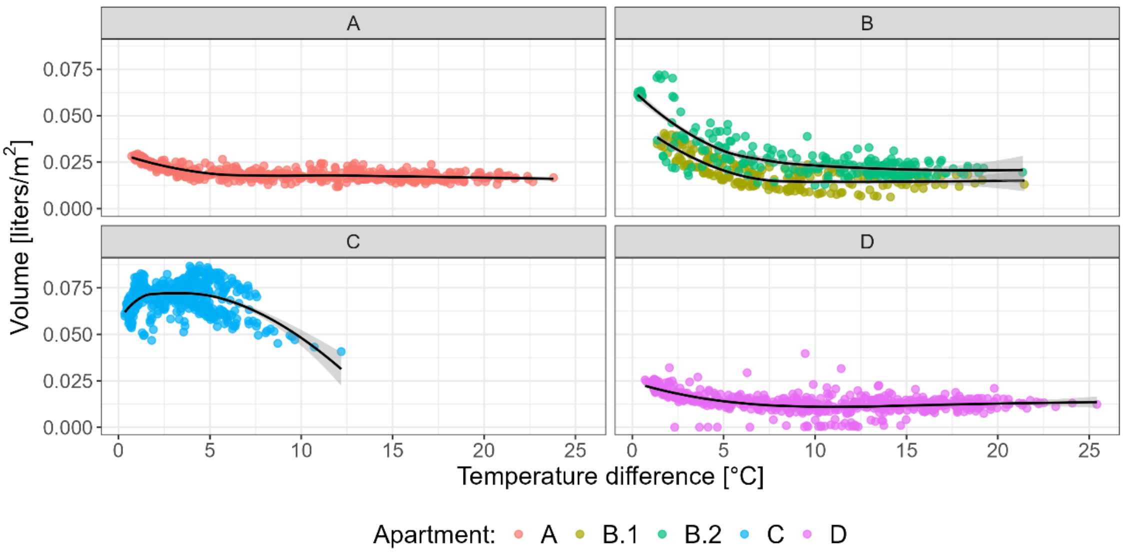


Figure 11: Performance of the overall SH systems per apartment.

In Figure 11, one can see, the performance of SH systems in four different apartments. The performance is assessed by plotting the volume of hot water against the temperature difference (ΔT) between the supply and return temperatures of the hot water provided by the DH grid. Apartment A shows a relatively constant volume of hot water use across the range of temperature differences, suggesting a stable and possibly efficient SH system performance, as also seen in apartment

D. Apartment B shows a decrease in the volume used as the temperature difference increases, which could indicate that the system is responding to the need for less water flow as the difference between supply and return temperatures grows. In apartment C, there is a constant large volume with a lower temperature difference compared with the other apartments. This indicates that apartment C has a much larger overflow with a lower ΔT.

By comparing these measurements with inputs from the surveys in Table 2, it is inferred the reasons behind the patterns in Figure 10 and Figure 11. Apartments A and D have the most similar heat performance of all the apartments. The reasons behind it are that these apartments have the same number of dwellers. However, apartment D has more outliers regarding the volume and ΔT, which are probably related to the fact this resident stays longer in its house, therefore it might change the setpoints more often. This hypothesis is corroborated, as one can see in Figure 10, where the living room (the space that probably is more occupied during the day) radiator has a larger emitted temperature, due to the increasing thermostat setpoint by the user. With a larger intensity but also an alike pattern, it is apartment B, where B.1 occupants have less SH demand than B.2. And it is also seen that B.2 has a higher SH usage due to a larger setpoint in Bedroom 1. Apartment C has the most different pattern from all the dwellings, and according to Figure 11, the worst heat efficiency because of its high overflow and lowest ΔT. The primary cause of this inefficiency is due to the living room radiator being set to the highest thermostat setting while all others are turned off, essentially making it the sole emitter attempting to heat the entire apartment. This also shows that the SH efficiency is not fully dependent on the number of occupants, as B.1 has more people than apartment C.

Sigmoid energy signature

The initial segment of this article investigation focuses on examining the linear ES model, while in this section, the focus shifts toward the proposal of the sigmoid ES model to explain the energy usage in a building. This analysis is critical for understanding the energy demand and the underlying factors influencing its variability in extreme outdoor temperatures. Before delving into the detailed analysis, it is essential to depict the sigmoid ES through graphical representation, which will elucidate the gradual increase and saturation points of energy use over time per apartment (see Figure 12). Apartment B was not investigated in this section, due to the lack of occupants during the no heating season.

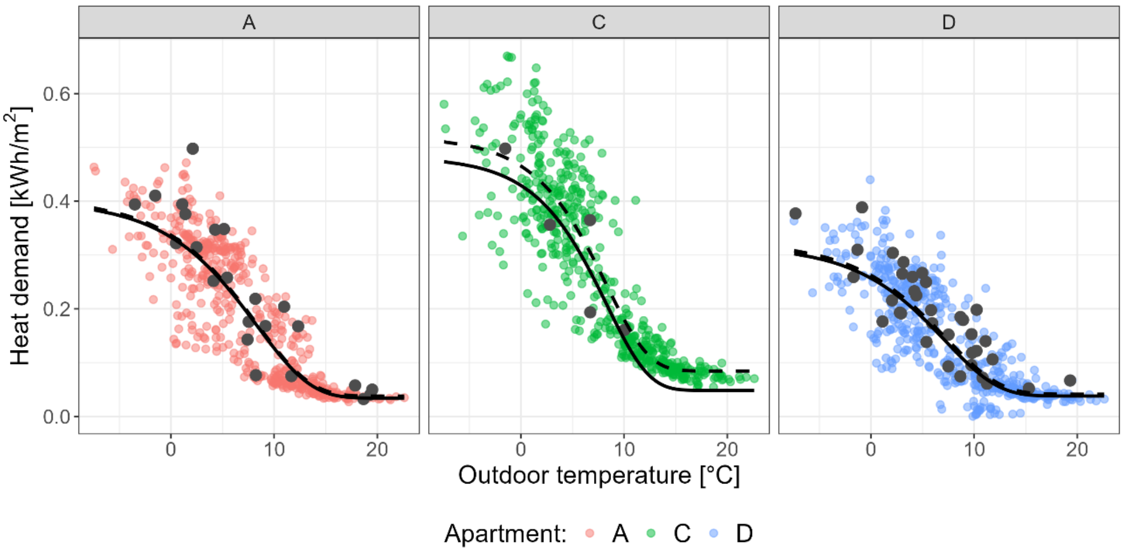


Figure 12: SH (solid line) and Total (dashed line) sigmoid ES for each apartment. Colored data points are the total heat demand recordings. Dark grey highlighted points are measurements assigned as outliers due to the DHW demand (see Figure 6).

The plot shows the sigmoid relationship between outdoor temperature and energy demand for heating across different apartments, with SH shown by a solid line and total heat demand by a dashed line. The energy demand increases as the outdoor temperature decreases, indicating more heating is needed. However, the demand plateaus in extreme outdoor temperature conditions. The grey highlighted points indicate measurements associated with outlier DHW demand however they do not impact the overall energy measurements. To quantify the extent of variation in the sigmoid ES parameters across the different apartments, the subsequent figure was produced (Figure 13).

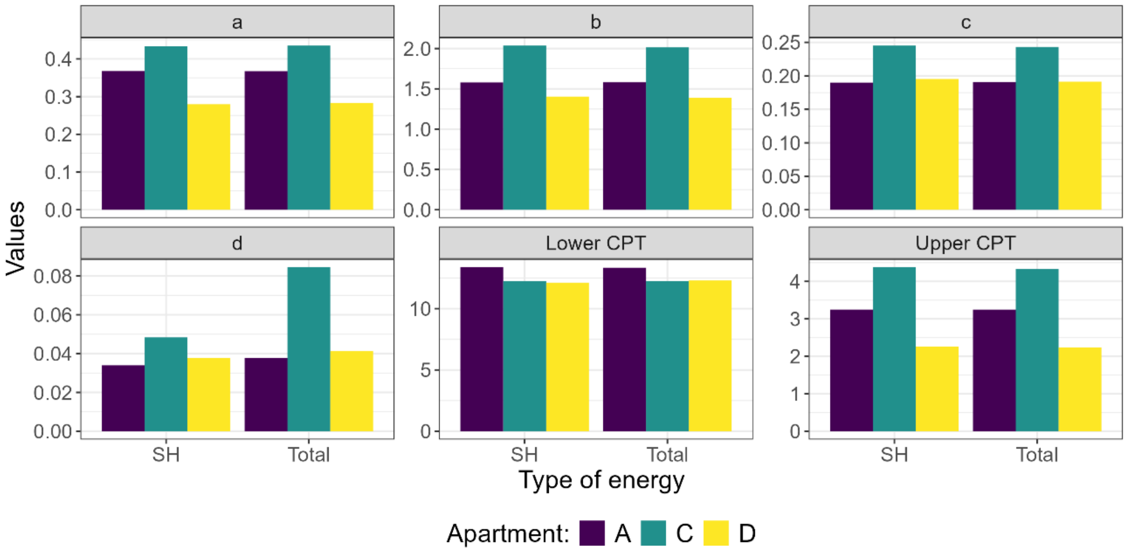


Figure 13: Sigmoid energy signature parameters of SH and total heat (SH+DHW) demand per apartment.

Similarly to the parameters of the linear ES model in Figure 9, sigmoid ES displays a similar display. As one can see in Figure 13, sigmoid parameters have small variations for both SH and the total demand. However, when comparing the same type of energy usage for each apartment, these values differ much more, as explained for the linear ES, it is due to the occupancy interaction with their heating systems.

To compare both models, the root mean squared error (RMSE), R2, and Akaike Information Criterion (AIC) were calculated for the SH measurements. RMSE measures the average magnitude of the errors between predicted and observed values, indicating the model's prediction accuracy. R2 (known as coefficient of determination), indicates the proportion of variance in the dependent variable that is predictable from the independent variables, reflecting the model's goodness of fit. While the AIC evaluates the model's quality by considering both the goodness of fit and the number of parameters used, penalizing more complex models to prevent overfitting. For both RMSE and AIC, the lower these values the better, while R2 the closest this value to 1, the better the model. In Table 3, it is seen the metrics for each model.

Table 3: RMSE, R2, and AIC of both linear and sigmoid ES models for the SH measurements.

	Linear ES			**Sigmoid ES**		
	Apart. A	**Apart. C**	**Apart. D**	**Apart. A**	**Apart. C**	**Apart. D**
RMSE	5.346	6.683	4.596	5.233	6.401	4.548
R2	0.754	0.776	0.703	0.764	0.795	0.709
AIC	3222.8	3461.3	3066.1	3200.8	3416.4	3055.2

The sigmoid ES model outperforms the linear ES model across all apartments, as evidenced by lower RMSE and AIC values, and slightly higher R^2 values. However, the difference between these metrics is quite small. To better corroborate the sigmoid model, 5-fold cross-validation was also performed using the RMSE (see Table 4).

Table 4: Results of the 5-fold cross-validation using RMSE.

		RMSE					
		Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Mean**
Apart. A	**Linear ES**	5.176	5.714	5.829	5.237	4.850	**5.361**
	Sigmoid ES	4.999	5.531	5.817	5.200	4.680	**5.246**
Apart. C	**Linear ES**	6.673	6.198	7.488	6.602	6.445	**6.681**
	Sigmoid ES	6.343	5.544	7.633	6.457	5.819	**6.359**
Apart. D	**Linear ES**	4.273	4.891	4.996	4.631	4.320	**4.622**
	Sigmoid ES	4.185	4.777	5.181	4.512	4.328	**4.596**

From the cross-validation, its results confirm that the sigmoid ES model performs better than the linear ES model. The sigmoid model consistently shows lower mean RMSE values across all apartments (A, C, D), indicating better predictive accuracy, but also better robustness when being trained and tested with the different folds. This reinforces the initial findings that the sigmoid ES model is the preferable choice, even though with slightly better metrics.

Besides these metrics, the primary feature of the sigmoid model is its distinctive dual energy plateaus. The plateau corresponding to the higher outdoor temperatures is already accounted for in the conventional linear ES model, reflecting a decrease in the building's SH demand. However, the plateau occurring at lower outdoor temperatures represents a significant innovation of this model. As noted in references [32,33], the possible explanations for this phenomenon include a) a reduced venting rate due to fewer window openings by occupants, thereby diminishing ventilation losses, or b) reaching the maximum power output by the building's SH systems.

Further analysis is conducted using data from sensors installed during the measurement campaign to elucidate the causes of this plateau. The other sensors measure the following: CO2 concentration, the apartments' SH supply temperature, and the window opening.

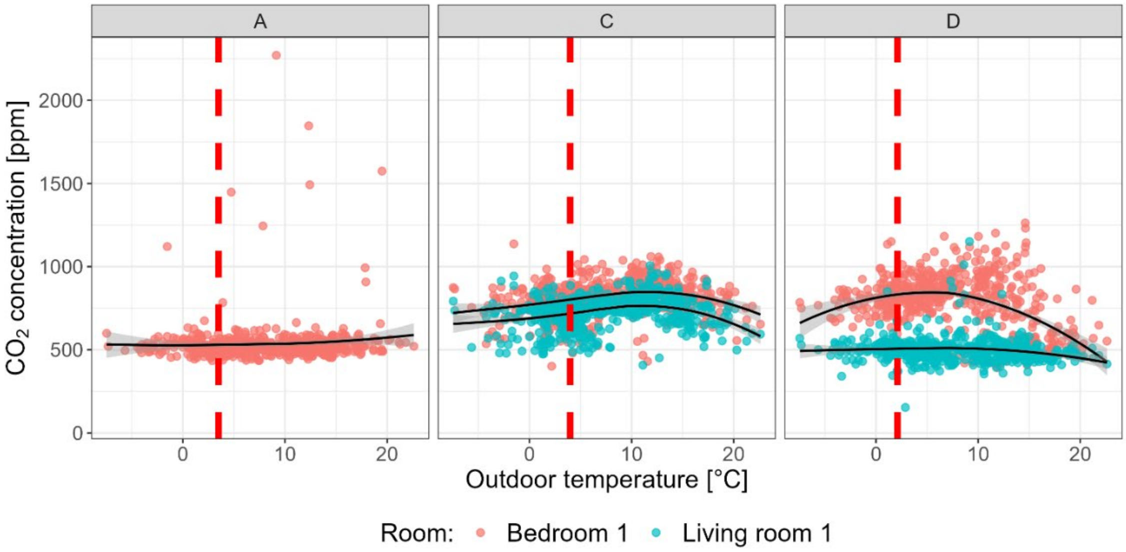


Figure 14: Daily CO_2 concentration for each apartment over the outdoor temperature. The dashed red line represents the upper CPT from the sigmoid ES.

These scatterplots in Figure 14 provide an insightful view into the daily CO2 concentration variations across different rooms within apartments A, C, and D, against the outdoor temperatures. The data, delineated by room type does suggest an almost constant relationship between CO2 concentration and external temperatures. The upper CPT is denoted by a dashed red line across the plots providing the reference point that indicates the colder outdoor temperature plateau from the sigmoid ES of each apartment. It was expected to see in these measurements an increase in CO2 concentration during the colder days due to a decrease in the window opening, however, this was not observed. The main reason is possibly the existing mechanical ventilation system that prevents the CO2 levels from reaching higher values.

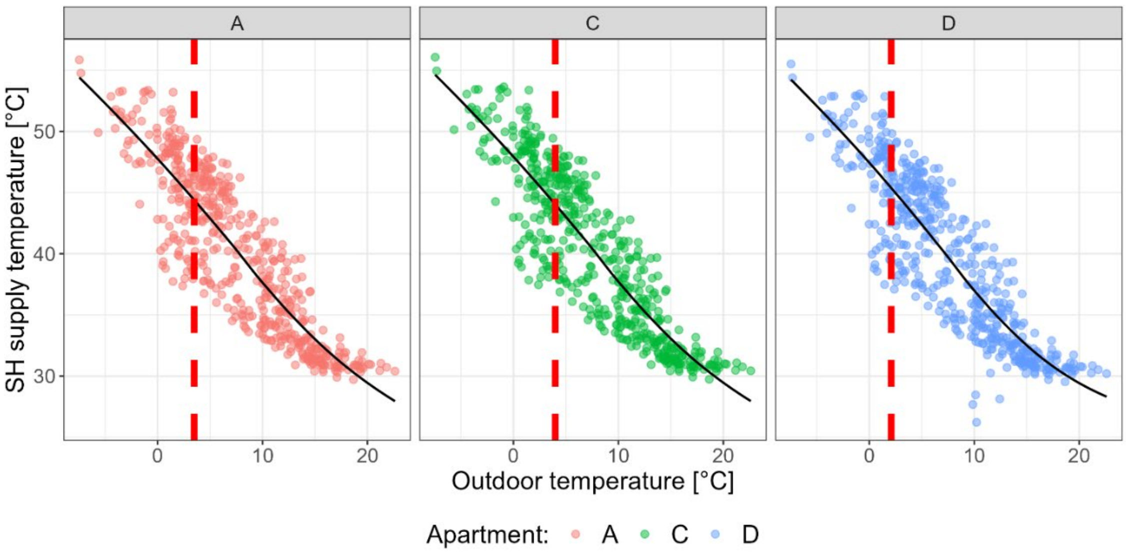


Figure 15: Daily supply temperature of the SH system for each apartment over the outdoor temperature. The dashed red line represents the upper CPT from the sigmoid ES.

While examining the SH supply temperature per apartment, Figure 15 distinctly illustrates an inverse relationship between the daily SH system supply temperature and the outdoor temperature for apartments A, C, and D. As the outdoor temperature increases, the SH system's supply temperature decreases consistently across these apartments, indicating a system that dynamically adjusts its output in response to external temperatures [36]. However, similarly with the CO2 concentration measurements, there is no discernible pattern in the region where the outdoor temperature is below the upper CPT that could explain the sigmoid model's plateau. Consequently, the plateau cannot be attributed to the maximum heating output of the SH system. This leads to the conclusion that the causes of the energy plateau lie elsewhere, necessitating further investigation.

Unfortunately, the window-opening sensors installed in the measurement campaign (see Figure 2) encountered numerous issues, rendering some of the gathered data unusable. Despite these setbacks, the sensors in apartment A yielded the results, as depicted in Figure 16.

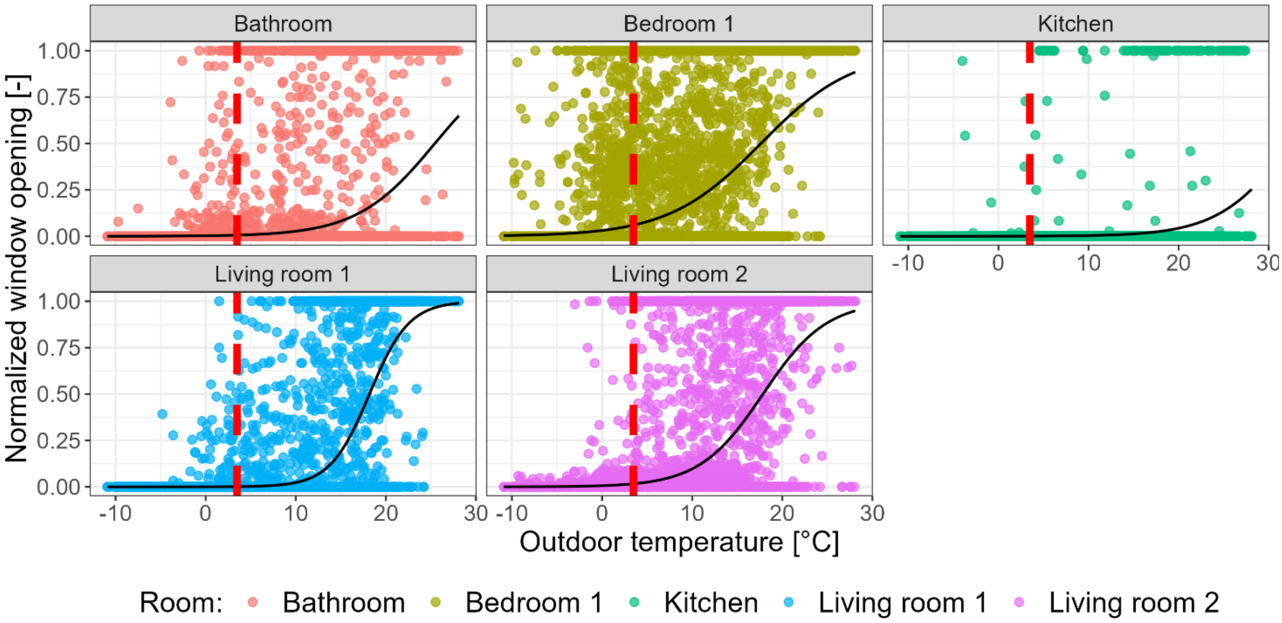


Figure 16: Hourly normalized window opening fraction for apartment A over the outdoor temperature. The dashed red line represents the upper CPT from the sigmoid ES for the SH measurements.

One can see in Figure 16, the hourly normalized window opening fraction for different rooms within apartment A, plotted against the outdoor temperature. The normalization of the window opening fraction allows for a comparison of window usage patterns relative to temperature changes. The data reveals a clear pattern: in both the bathroom and the living rooms, there's a positive correlation between window-opening behavior and outdoor temperature, with window openings becoming more frequent as the temperature outside rises. This trend is especially marked in the living room, where the frequency of window openings significantly increases with higher temperatures. In contrast, the bedroom shows a less pronounced dependency on outdoor temperature, as evidenced by the relatively flat trend lines, indicating that factors other than temperature, such as the need to ventilate the room after waking up, may influence window-opening behavior in these spaces.

This observation suggests a potential explanation for the sigmoid low external temperature plateau observed in the model. As hypothesized in other research articles [32,33], the venting rate appears to be a primary factor behind this plateau. This correlation might have been more evident in the CO2 concentration measurements if the apartments lacked mechanical ventilation, hinting at a direct link between air quality and window-opening habits. Furthermore, these findings underscore the importance of accounting for user behavior, particularly window-opening routines, in building energy models, as they significantly impact energy usage and efficiency.

Conclusion

In the introduction of this article, it was raised three questions that we attempted to answer based on the integration of smart heat meters' data with other indoor sensors data and surveys to have more insights into the occupants. As a product of this analysis, it was concluded the following:

- Is the DHW energy usage constant or irregular when applied to the ES model?

Visual analysis of daily time series data showed that while SH displayed a seasonal pattern, DHW usage was more stochastic but generally fluctuated around a consistent average. Monthly trends further confirmed a constant median

DHW usage pattern, with variations primarily influenced by their different routines and not the number of occupants. Consequently, all DHW outliers, a product of its stochasticity, did not seem to influence the ES model significantly besides increasing slightly its CPE-value (see Figure 9).

- Are the ES characteristics similar in alike buildings and spaces?

Contrary to the initial hypothesis, the ES parameters were not consistent across similar buildings and spaces. Analysis of energy signatures for SH and total energy demand showed significant differences among the apartments, suggesting that factors other than thermal characteristics, such as occupancy behavior and potential faults in SH systems, significantly affect the overall heat performance. This finding was reinforced by examining the heating profiles and efficiency variables, such as water volume and ΔT, from the radiator heat allocators which indicated that occupant interaction with heating systems plays a crucial role in the observed variations.

- What are the reasons behind the upper asymptote point in the sigmoid ES model at low outdoor temperatures?

The reasons behind the upper asymptote point in the sigmoid ES model at low outdoor temperatures were explored using additional sensor data. The upper asymptote, representing a plateau in energy demand at low temperatures, could not be explained by supply temperature limitations of the heating systems or CO_2 concentration variations. However, data from window opening sensors suggested that reduced venting due to less window opening at lower temperatures is likely a primary factor, nevertheless, more data is required to confirm this assumption. This behavior, influenced by occupant routines, underscores the importance of accounting for user behavior in building energy models to better understand and predict energy usage patterns.

Suggestions for Further Work

Based on the findings of this study, several paths for further research are proposed. A primary focus should be on enhancing fault detection and diagnosis (FDD) in buildings' heating systems and district heating substations. Given the significant influence of occupant behavior towards their systems on the ES model, future research should integrate advanced FDD methods that leverage the variability in energy usage patterns to identify anomalies and inefficiencies in heating systems. This could involve the development of machine learning algorithms trained on extensive datasets from smart heat meters and indoor sensors, aiming to detect subtle signs of system malfunctions or suboptimal performance before they escalate into more significant issues.

Additionally, exploring the integration of real-time monitoring systems with user-friendly interfaces could empower occupants to better understand and manage their energy usage, potentially reducing the occurrence of faults due to user error. Research should also investigate the impact of various occupant behaviors on the efficiency of heating systems, using detailed surveys to refine predictive models.

Another promising direction involves the application of the sigmoid ES model to a broader range of buildings and climates to validate its robustness and versatility. This includes assessing the model's performance in commercial and industrial buildings, where energy usage patterns and HVAC systems differ from residential settings. Moreover, expanding the dataset to include more diverse building types and regions would provide a comprehensive understanding of this ES model's applicability and limitations. Further investigation into the reasons behind the upper asymptote in the sigmoid ES is also needed. More extensive data collection, particularly focusing on window opening behavior and its correlation with indoor air quality metrics, could help confirm the initial findings and provide deeper insights into occupant-driven energy dynamics.

Credit author statement

Daniel Leiria: Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Data curation; Writing - original draft; Writing - review & editing; Visualization. **Hicham Johra:** Conceptualization; Methodology; Resources; Writing - review & editing; Supervision. **Yue Hu:** Data curation; Resources; Writing - review & editing. **Olena Kalyanova Larsen:** Conceptualization; Methodology; Writing - review & editing. **Anna Marszal-Pomianowska:** Conceptualization; Resources; Writing - review & editing; Supervision. **Martin Frandsen:** Resources; Writing - review & editing. **Michal Zbigniew Pomianowski:** Conceptualization; Resources; Writing - review & editing; Supervision; Project administration; Funding acquisition.

All authors have read and agreed to the published version of the manuscript.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 893945 (E-DYCE).

Declaration of competing interest

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

Acknowledgments

The authors would like to acknowledge all partners involved in the E-DYCE and PRELUDE projects. Also, the first author would like to express his gratitude to his colleagues (Kamilla H. Andersen, Markus Schaffer, Martin Frandsen, and Simon P. Melgaard) in the Ph.D. "Corner Office" for the valuable discussions and support.

References

[1] European Comission, "Energy performance of buildings directive," *Energy - European Commission*, 2019. [Online]. Available: https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en. [Accessed: June 25, 2024].

[2] N.E. Klepeis *et al.*, "The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants," *Journal of Exposure Science & Environmental Epidemiology*, vol. 11, pp. 231–252, 2001, doi: https://doi.org/10.1038/sj.jea.7500165.

[3] L. Rohde, T. S. Larsen, R. L. Jensen, O. K. Larsen, K. T. Jønsson, and E. Loukou, "Determining indoor environmental criteria weights through expert panels and surveys," *Building Research & Information*, vol. 48, no. 4, pp. 415–428, 2019, doi: https://doi.org/10.1080/09613218.2019.1655630.

[4] M.F. Fels, "PRISM: An introduction," *Energy and Buildings*, vol. 9, no. 1–2, pp. 5–18, 1986, doi: https://doi.org/10.1016/0378-7788(86)90003-4.

[5] S. Hammarsten, "A critical appraisal of energy-signature models," *Applied Energy*, vol. 26, no. 2, pp. 97–110, 1987, doi: https://doi.org/10.1016/0306-2619(87)90012-2.

[6] P. Westermann, C. Deb, A. Schlueter, and R. Evins, "Unsupervised learning of energy signatures to identify the heating system and building type using smart meter data," *Applied Energy*, vol. 264, p. 114715, 2020, doi: https://doi.org/10.1016/j.apenergy.2020.114715.

[7] V. Milić, P. Rohdin, and B. Moshfegh, "Further development of the change-point model – Differentiating thermal power characteristics for a residential district in a cold climate," *Energy and Buildings*, vol. 231, p. 110639, 2021, doi: https://doi.org/10.1016/j.enbuild.2020.110639.

[8] J. Palmer Real, J. K. Møller, R. Li, and H. Madsen, "A data-driven framework for characterising building archetypes: A mixed effects modelling approach," *Energy*, vol. 254, p. 124278, 2022, doi: https://doi.org/10.1016/j.energy.2022.124278.

[9] ASHRAE, "Measurement of Energy and Demand Savings," ASHRAE Guideline 14-2002, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta, GA, 2002.

[10] E. Fuentes, L. Arce, and J. Salom, "A review of domestic hot water consumption profiles for application in systems and buildings energy performance analysis," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1530–1547, 2018, doi: https://doi.org/10.1016/j.rser.2017.05.229.

[11] D. Ivanko, Å. L. Sørensen, and N. Nord, "Splitting measurements of the total heat demand in a hotel into domestic hot water and space heating heat use," *Energy*, vol. 219, p. 119685, 2021, doi: https://doi.org/10.1016/j.energy.2020.119685.

[12] D. Leiria, H. Johra, A. Marszal-Pomianowska, and M. Z. Pomianowski, "A methodology to estimate space heating and domestic hot water energy demand profile in residential buildings from low-resolution heat meter data," *Energy*, vol. 263, p. 125705, 2023, doi: https://doi.org/10.1016/j.energy.2022.125705.

[13] T. Cholewa et al., "A simple building energy model in form of an equivalent outdoor temperature," *Energy and Buildings*, vol. 236, p. 110766, 2021, doi: https://doi.org/10.1016/j.enbuild.2021.110766.

[14] J. Vesterberg, S. Andersson, and T. Olofsson, "Robustness of a regression approach, aimed for calibration of whole building energy simulation tools," *Energy and Buildings*, vol. 81, pp. 430–434, 2014, doi: https://doi.org/10.1016/j.enbuild.2014.06.035.

[15] D. Leiria, H. Johra, A. Marszal-Pomianowska, M. Z. Pomianowski, and P. Kvols Heiselberg, "Using data from smart energy meters to gain knowledge about households connected to the district heating network: A Danish case," *Smart Energy*, vol. 3, p. 100035, 2021, doi: https://doi.org/10.1016/j.segy.2021.100035.

[16] Jan-Ulric Sjögren, S. Andersson, and T. Olofsson, "Sensitivity of the total heat loss coefficient determined by the energy signature approach to different time periods and gained energy," *Energy and Buildings*, vol. 41, no. 7, pp. 801–808, 2009, doi: https://doi.org/10.1016/j.enbuild.2009.03.001.

[17] M. Eriksson, J. Akander, and B. Moshfegh, "Development and validation of energy signature method – Case study on a multi-family building in Sweden before and after deep renovation," *Energy and Buildings*, vol. 210, p. 109756, 2020, doi: https://doi.org/10.1016/j.enbuild.2020.109756.

[18] B. Arregi and R. Garay, “Regression analysis of the energy consumption of tertiary buildings,” *Energy Procedia*, vol. 122, pp. 9–14, 2017, doi: https://doi.org/10.1016/j.egypro.2017.07.290.

[19] J. Rose, J. Kragh, and K. F. Nielsen, “Passive house renovation of a block of flats – Measured performance and energy signature analysis,” *Energy and Buildings*, vol. 256, p. 111679, 2022, doi: https://doi.org/10.1016/j.enbuild.2021.111679.

[20] T. Csoknyai, J. Legardeur, A. A. Akle, and M. Horváth, “Analysis of energy consumption profiles in residential buildings and impact assessment of a serious game on occupants’ behavior,” *Energy and Buildings*, vol. 196, pp. 1–20, 2019, doi: https://doi.org/10.1016/j.enbuild.2019.05.009.

[21] H. Gadd and S. Werner, “Fault detection in district heating substations,” *Applied Energy*, vol. 157, pp. 51–59, 2015, doi: https://doi.org/10.1016/j.apenergy.2015.07.061.

[22] F. W. Yu and K. T. Chan, “Energy signatures for assessing the energy performance of chillers,” *Energy and Buildings*, vol. 37, no. 7, pp. 739–746, 2005, doi: https://doi.org/10.1016/j.enbuild.2004.10.004.

[23] L. Belussi and L. Danza, “Method for the prediction of malfunctions of buildings through real energy consumption analysis: Holistic and multidisciplinary approach of Energy Signature,” *Energy and Buildings*, vol. 55, pp. 715–720, 2012, doi: https://doi.org/10.1016/j.enbuild.2012.09.003.

[24] A. Anjomshoaa and M. Salmanzadeh, “Estimation of the changeover times and degree-days balance point temperatures of a city using energy signatures,” *Sustainable Cities and Society*, vol. 35, pp. 538–543, 2017, doi: https://doi.org/10.1016/j.scs.2017.08.028.

[25] O. Pasichnyi, J. Wallin, and O. Kordas, “Data-driven building archetypes for urban building energy modelling,” *Energy*, vol. 181, pp. 360–377, 2019, doi: https://doi.org/10.1016/j.energy.2019.04.197.

[26] P. Gianniou, C. Reinhart, D. Hsu, A. Heller, and C. Rode, “Estimation of temperature setpoints and heat transfer coefficients among residential buildings in Denmark based on smart meter data,” *Building and Environment*, vol. 139, pp. 125–133, 2018, doi: https://doi.org/10.1016/j.buildenv.2018.05.016.

[27] Q. Meng et al., “Change-point multivariable quantile regression to explore effect of weather variables on building energy consumption and estimate base temperature range,” *Sustainable cities and society*, vol. 53, pp. 101900–101900, 2020, doi: https://doi.org/10.1016/j.scs.2019.101900.

[28] C. Ghiaus, “Experimental estimation of building energy performance by robust regression,” *Energy and Buildings*, vol. 38, no. 6, pp. 582–587, 2006, doi: https://doi.org/10.1016/j.enbuild.2005.08.014.

[29] F. Flouquet, “Local weather correlations and bias in building parameter estimates from energy-signature models,” *Energy and Buildings*, vol. 19, no. 2, pp. 113–123, 1992, doi: https://doi.org/10.1016/0378-7788(92)90005-2.

[30] J. Vesterberg, S. Andersson, and T. Olofsson, “A single-variate building energy signature approach for periods with substantial solar gain,” *Energy and Buildings*, vol. 122, pp. 185–191, 2016, doi: https://doi.org/10.1016/j.enbuild.2016.04.040.

[31] A. Rabl and A. Rialhe, “Energy signature models for commercial buildings: test with measured data and interpretation,” *Energy and Buildings*, vol. 19, no. 2, pp. 143–154, 1992, doi: https://doi.org/10.1016/0378-7788(92)90008-5.

[32] P. Nageler et al., “Comparison of dynamic urban building energy models (UBEM): Sigmoid energy signature and physical modelling approach,” *Energy and Buildings*, vol. 179, pp. 333–343, 2018, doi: https://doi.org/10.1016/j.enbuild.2018.09.034.

[33] E. A. Koch, “Continuous Simulation for Urban Energy Planning Based on a Non-Linear Data-Driven Modelling Approach,” Ph.D. thesis, Karlsruher Institut für Technologie, Germany, 2016.

[34] M. Pomianowski, O. K. Larsen, D. Leiria, P. Vogler-Finck, and P. D. Pedersen, “Danish case studies report,” 2022. Accessed: Jun. 25, 2024. [Online]. Available: https://edyce.eu/wp-content/uploads/2022/09/E-DYCE_D5.5_Danish_case_studies_report_29.08.2022_Final.pdf

[35] “ E-DYCE Project,” E-DYCE. https://edyce.eu/ (accessed Jun. 25, 2024).

[36] Neogrid, “PreHEAT,” Intelligent energistyring af bygningens varmeforbrug. https://neogrid.dk/preheat/ (accessed Jun. 25, 2024).

[37] V. M. R. Muggeo, “Estimating regression models with unknown break-points,” *Statistics in Medicine*, vol. 22, no. 19, pp. 3055–3071, 2003, doi: https://doi.org/10.1002/sim.1545.

[38] V. M. R. Muggeo, “segmented: Regression Models with Break-Points / Change-Points Estimation (with Possibly Random Effects),” R-Packages, 2024, https://cran.r-project.org/web/packages/segmented/index.html (accessed Jun. 25, 2024).

Appendix

Determination of the changepoint temperatures (CPT) of the sigmoid energy signature model

The sigmoid energy signature is an inverted-s curve function, and it is defined mathematically as:

$$y = ae^{-e^{cx-b}} + d \qquad (A.1)$$

This function has two x-points where the graph's asymptote "starts". In mathematical terms, it can be represented where the third derivative of the function is zero (as one can see in A.2).

$$y'''(x) = \frac{d^3y}{dx^3} = 0 \qquad (A.2)$$

The first derivative of function y is defined as:

$$y' = -ac \cdot e^{-e^{cx-b}+cx-b} \qquad (A.3)$$

By derivating A.3, it is reached the second derivative of y (y''):

$$y'' = ac^2 \cdot e^{-e^{cx-b}+cx-2b}\left(e^{cx} - e^{b}\right) \qquad (A.4)$$

By derivating A.4, it is reached the third derivative of y (y'''):

$$y''' = -ac^3 \cdot e^{-e^{cx-b}+cx-3b}\left(-3e^{cx+b} + e^{2cx} + e^{2b}\right) \qquad (A.5)$$

Using A.5, it is necessary to find the x values that fulfill the condition A.2:

$$-ac^3 \cdot e^{-e^{cx-b}+cx-3b}\left(-3e^{cx+b} + e^{2cx} + e^{2b}\right) = 0 \qquad (A.6)$$

In the expression A.6, the coefficients a and c must not be equal to 0. Therefore, to find x-values that fulfill the condition, only the expression inside the parenthesis must be equal to zero:

$$-3e^{cx+b} + e^{2cx} + e^{2b} = 0 \qquad (A.7)$$

Converting the following:

$$\begin{cases} \beta = e^{cx} \\ \gamma = e^{b} \end{cases} \qquad (A.8)$$

Therefore, substituting in equation A.7 by the variables in A.8.

$$-3\beta\gamma + \beta^2 + \gamma^2 = 0 \qquad (A.9)$$

Solving the first condition above for β:

$$\beta = \frac{3\gamma \pm \sqrt{9\gamma^2 - 4\gamma^2}}{2} = \frac{3}{2}\gamma \pm \frac{1}{2}\gamma\sqrt{5} \qquad (A.10)$$

Substituting in A.10 with the values in A.8:

$$e^{cx} = \frac{1}{2}e^{b}\left(3 \pm \sqrt{5}\right) \qquad (A.11)$$

Finally obtaining:

$$\begin{cases} x_1 = \frac{1}{c}\ln\left[\frac{1}{2}e^{b}(3 - \sqrt{5})\right] = CPT_{upper} \\ x_2 = \frac{1}{c}\ln\left[\frac{1}{2}e^{b}(3 + \sqrt{5})\right] = CPT_{lower} \end{cases} \qquad (A.12)$$

These two x-values are the points that represent the two outdoor temperature values where the sigmoid energy signature becomes constant.

4.3 Further discussion

From **Paper 5**, it is concluded that integrating SHM data with indoor sensor measurements can provide a more comprehensive understanding of the operation of residential apartments and buildings [23]. This integration enables a detailed analysis of the interactions between building heating systems and their occupants, offering insights that are crucial for optimizing building performance. Furthermore, patterns in the SHM recordings reveal a significant impact on residents' behavior towards their heating systems performance which can be regarded as potential faults within these systems. Such faults can markedly affect energy efficiency, as deviations in heating performance often manifest in the SHM data.

Recognizing these patterns in the SHM data underscores the need for a more focused effort on understanding and addressing heating system faults. Detecting and diagnosing these faults is essential to mitigating their impact on a building's individual energy usage and overall grid efficiency [40]. As the ability to quantify the effects of these faults not only aids in improving energy efficiency but also enhances the reliability of heating systems within the residential sector. This research emphasizes the importance of fault detection and diagnosis (FDD) within district heating (DH) customer substations and their heating systems, highlighting the need for advanced methodologies to identify and diagnose these faults.

The following chapter delves deeper into this investigation, exploring innovative data-driven approaches for effective FDD. The aim is to optimize building operations to support more efficient DH network operation, thereby enhancing overall operational efficiency and contributing to more sustainable energy management practices.

Chapter 5. Fault detection and diagnosis within district heating customers

This chapter discusses the fourth and last research topic of the Ph.D. project, which focuses on leveraging smart heat meters (SHM) data for fault detection and diagnosis (FDD) in buildings connected to the district heating (DH) network.

5.1 The challenge of faults in the district heating systems

The importance of FDD in DH systems can be seen through several key aspects. Firstly, energy efficiency is greatly influenced by the operational integrity of these systems. Faults such as leaks, blockages, or control malfunctions in components like heat exchangers can result in substantial energy losses [41]. By implementing effective FDD processes, these issues can be detected early, allowing for timely interventions that minimize unnecessary energy usage and lower fluid-return temperatures. Secondly, the reliability of the heat supply is vital for the comfort and productivity of its end-users. As undetected, faults can disrupt the consistent delivery of heat, leading to thermal discomfort and potential economic losses, such as extra heating tariffs [42].

Despite the clear benefits of FDD, traditional manual methods of fault detection have significant limitations. Manual detection relies heavily on routine inspections and the expertise of maintenance personnel, which can be delayed, time-consuming, and prone to human error [30]. The complexity and scale of DH networks further aggravate these challenges, making it difficult to monitor and analyze the multitude of parameters and customers in the grid.

To overcome the limitations of manual fault detection, there is a growing need for automated FDD approaches. Automated systems can continuously monitor the performance of DH systems in real-time, using advanced algorithms and data analytics to identify anomalies and potential faults with high precision [43]. These systems can analyze large volumes of data from various sensors and control points, enabling the early detection of faults that might be missed by manual inspections. Furthermore, automated FDD can provide diagnostic insights and actionable recommendations, facilitating quicker and more accurate responses to identified issues. This not only enhances the overall efficiency and reliability of DH systems but also reduces the reliance on human intervention. To address this topic, the research conducted in this area is presented and summarized in two articles.

Paper 6 focuses on examining data from 158 fault reports and SHM data from residential buildings in Denmark to investigate prevalent faults and assess their impact. The analysis included a subset of 90 buildings, comprising both faulty and non-faulty data points, to benchmark various fault detection indicators [24].

Paper 7 further investigates the challenges and advancements in collecting ground-truth data to improve fault diagnosis models for DH systems. Motivated by the need to overcome limitations in previous data collections, the research utilizes an improved dataset to identify faults in household substations. Despite encountering issues such as inaccurate fault categorizations, obscured fault patterns, and truncated measurements, the analysis of 50 detailed cases out of 127 fault reports reveals that return temperature and volume flow are reliable fault indicators, whereas energy usage patterns are not. Regarding fault diagnosis, the research categorizes fault symptoms and patterns, showcasing the effectiveness of high-dimensional data clustering in fault diagnosis using self-organizing maps. However, it is important to note that with the current data available, SHM measurements resolution, and the number of fault reports, it is not possible to diagnose faults at the component level (e.g., specific valve), but rather at the system type, such as the radiator, underfloor heating, or domestic hot water [25].

5.2 Initial assessment of FDD in the district heating

Paper 6

"Towards automated fault detection and diagnosis in district heating customers: generation and analysis of a labeled dataset with ground truth"

Daniel Leiria, Kamilla Heimar Andersen, Simon Pommerencke Melgaard, Hicham Johra, Anna Marszal-Pomianowska, Marco Piscitelli, Alfonso Capozzoli, Michal Zbigniew Pomianowski, IBPSA Building Simulation 2023, Shanghai, China, https://doi.org/10.26868/25222708.2023.1576.

Towards automated fault detection and diagnosis in district heating customers: generation and analysis of a labeled dataset with ground truth

Daniel Leiria[*1], Kamilla Heimar Andersen[1], Simon Pommerencke Melgaard[1], Hicham Johra[1], Anna Marszal-Pomianowska[1], Marco Savino Piscitelli[2], Alfonso Capozzoli[2], and Michal Zbigniew Pomianowski[1]
[1] Aalborg University, Department of the Built Environment, Aalborg, Denmark
[2] Politecnico di Torino, Department of Energy, TEBE Research Group, BAEDA Lab, Turin, Italy

Abstract

This study aims to develop a framework for automated fault detection and diagnosis (AFDD) in district heating (DH) substations by comprehensively understanding typical faults. AFDD is presently dependent on manual detection and diagnosis, leading to limitations. To address this issue, the study utilized data from 158 fault reports and smart heat meter data from residential buildings in Denmark to investigate common faults and conduct a fault impact assessment. The study suggests additional indicators for use by DH utility companies to detect anomalies in the future. The findings indicate that greater attention to fault detection and diagnosis can decrease energy usage and return temperatures, demonstrating the significance of AFDD implementation.

Highlights

- The study analyzed 158 fault reports from DH systems, categorized according to the type of fault detected.
- A subset of 90 buildings with both faulty and non-faulty data points was used to benchmark different fault detection (FD) indicators.
- A novel FD indicator was proposed based on water volume and temperature difference in DH substations.
- DH system faults significantly impact energy use, and more efficient detection can reduce DH customers' energy usage on average by 14%.
- The proposed FD method shows promise in detecting anomalies for future use by DH utility companies and integration into an AFDD framework.

Introduction

District heating (DH) systems are becoming increasingly popular in many countries as an effective and centralized solution to provide heating to large agglomerates of buildings while reducing energy waste and integrating with other renewable energy sources. Nevertheless, like any complex system, DH networks are prone to faults that can lead to reduced energy efficiency and increased costs. By identifying and repairing these faults in DH systems, we can move towards more efficient and sustainable heating solutions, particularly in the context of 4th generation district heating (4GDH) systems prioritizing lower supply temperatures and higher energy efficiency (Li and Nord, 2018).

A successful transition towards 4GDH systems calls for the parallel transformation of the primary/supply and secondary/demand sides. Apart from the renovation of the building stock, the low-hanging fruits are the detection, diagnosis and fixing of the faults in the end-users' heating installations, which often cause high energy use and high return temperatures. According to Gadd and Werner (2015), from the analysis of a smart heat meters (SHM) dataset of 135 substations, 100 (74%) presented patterns that could indicate a faulty operation of the system. If this result is representative for the existing DH systems in operation today, it means that three-quarters of the heating distribution grid is underperforming. This presents a sizeable economic potential (approx. 0.05 to 0.5€/MWh·ºC) by fixing the existing faults in the DH end-user installations and consequently reducing the return temperature (Frederiksen and Werner, 2013; Gadd and Werner, 2014).

Background

There is significant research on fault detection and diagnosis (FDD) in DH systems, with studies focusing on developing automated algorithms to identify different symptoms and detect and diagnose faults in DH substations using SHM and other sensor data.

Gadd and Werner (2015) established a fault detection (FD) framework based on SHM and outdoor temperature measurements of a dataset of 135 Swedish substations. Månsson et al. (2019) presented a survey-based study to describe the current status of the Swedish DH utilities regarding FDD. The survey concluded that the common methods used by the utilities to evaluate the faulty customers are from within the Gadd and Werner (2015) framework. This FD framework finds customers with faulty substations through hourly SHM data when it is observed at least one of the three fault symptoms: a) *Unsuitable heat load pattern:* Any building with heating usage measurements different from what is expected from their occupancy profiles; b) *Low average annual temperature difference:* Any building with a yearly average ΔT significantly lower than 45ºC; c) *Poor substation control:* Any building with irregular energy oscillations and low correlation between heating usage and outdoor temperature.

The *overflow* indicator, presented by Frederiksen and Werner (2013), is also used by few utilities for fault detection purposes. However, even though these methods have proven useful in detecting faulty systems based on SHM measurements, they lack the knowledge of what type of faults the different indicators detect.

https://doi.org/10.26868/25222708.2023.1576

Also, based on Swedish DH systems and taking into account the FD framework mentioned above, Månsson et al. (2018a) developed a statistical algorithm to detect faulty substations based on their patterns of energy, temperature difference between supply and return DH water (ΔT), and return temperature signatures (i.e., variables' variation due to outdoor temperature). A similar approach was proposed by Calikus et al. (2018), which analyzed the energy signature generated by a robust linear regression model of the substations to detect abnormal data points that are significantly deviated from the linear trend. Both methods relied on threshold settings and statistical methodologies to detect anomalous data points (outliers), which are most likely due to faulty heating system operations caused by technical or occupant behavior (Schaffer et al., 2022). Focusing on specific faults in components of DH substations, Guelpa and Verdi (2020), developed and tested a methodology based on measurements in the primary and secondary circuits to detect fouling in the heat exchanger in some buildings connected to the DH network of Turin, Italy.

In the machine learning (ML) discipline, Månsson et al. (2018b) proposed an ML algorithm using gradient boosting regressor to model the flowrate of a well-performing substation based on the external temperature and ΔT to compare it with the same dataset with faulty data points induced by the authors. Xue et al. (2017) proposed a data mining algorithm consisting of data cleaning, clustering, and analysis of two substations in Changchun, China, to extract more information from SHM data. On a larger scale, Calikus et al. (2019) advanced the field of FDD in the DH systems by applying unsupervised clustering methodologies in a 1,385 Swedish buildings dataset to detect buildings with abnormal heating profiles, and after they investigated the reasons behind the anomalous measurement profiles.

The described methods above are significant steps towards detecting and diagnosing the various faults in DH substations more efficiently than today's process. However, to a certain extent, all these methods lacked adequately labeled data regarding the occurring faults, thus hindering the further development and testing of these FDD methodologies. As Månsson et al. (2021) outlined, labeled data with ground truth is a historical register of specific faults that occurred in a particular substation. Furthermore, according to the authors, these ground truth datasets must be gathered on a large scale and be combined with SHM data to understand the difference between "faulty" and "well-performing" heating installation operations.

Contributions

Some DH utilities apply the indicators proposed by Gadd and Werner (2015) framework for automated FD while a human user still carries out the diagnosis process. This is due to, as explained above, the lack of adequately labeled data with ground truth or an infrastructure to record and gather this information. This study attempts to investigate this matter by:

- Coupling approximately 351 fault reports (labeled dataset with ground truth) issued by technicians with measurements from SHM data.
- Providing an overview of typical faults occurring in DH end-user heating installations.
- Assess the efficacy of the current indicators for FDD purposes and propose a new one to assess the DH customers.
- Perform a fault impact assessment of the different faults to investigate their consequences on the buildings' energy usage.

Methodology

Dataset description

The dataset used in this study combines SHM data from 351 residential buildings connected to the DH grid with faults assessment reports made by technicians when visiting the installations in Aalborg, Denmark. The SHM dataset consists of the energy use, water volume, and ΔT measurements with an hourly resolution for 2022 for the heating demand of space heating (SH) and domestic hot water (DHW). Each building has an individual meter ID associated with the customer, and its fault assessment report is generated after a visit by a utility company technician. These visits were triggered when the substation's SHM registers ΔT-values below 10ºC. The weather data is extracted from the Danish Meteorologic Institute (DMI) portal. The selected weather station is Tylstrup, the station nearest to Aalborg, available in the DMI database.

In this work, the data pre-processing consisted of analyzing each report individually to categorize the different faults. Some of the reports were found to be ambiguous in terms of describing the real cause of the fault and were therefore disregarded from this study. Later in the study, it is only assessed the installations where it is indicated by the technicians that the fault was fixed after the visit. Therefore, from the initial dataset of 351 buildings, only 90 were used for this work. The data treatment of SHM measurements was performed by following the framework described in Schaffer et al. (2022). Customer installations differ based on country and DH system. Some countries have direct connections without hydraulic separation in the space heating system, while others have a prevalent indirect connection where the SH and DHW systems are separated from the primary circuit by heat exchangers (Figure 1). In this study, the heating installation was divided into two sections, the primary circuit (substation) and the secondary circuit (systems inside the household).

https://doi.org/10.26868/25222708.2023.1576

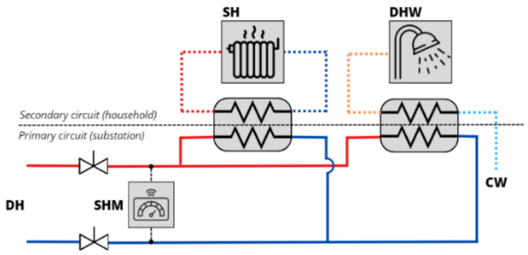


Figure 1: Illustration of a residential heating installation.

Definitions and FDD concepts

The immense body of literature on FDD employs several different naming conventions. In this study, the authors have adopted the ontology from Andersen et al. (2023) for definitions of a *symptom of a fault, fault, cause,* and *consequence* for the fault report analysis. Furthermore, labeled dataset with faulty ground truth is defined as annotations of each fault within the different defined fault categories presented in Table 1.

Furthermore, the framework proposed by Melgaard et al. (2022), originally developed for building systems, is applied in this study due to its overlap with the DH systems. The framework outlines three levels of FDD: fault detection, fault isolation and identification, and fault evaluation. Fault detection forms the fundamental basis of FDD, which involves determining the existence of a fault. Fault isolation and identification, in conjunction with fault detection, comprise the FDD process by pinpointing the fault location and identifying the cause of the fault. Finally, the fault evaluation stage assesses the impact of the fault, which can include estimating the excess energy or financial resources used due to the fault.

Applied FDD indicators

In this work, five different indicators are used to assess the installations' performance. Four of these are already in use by several DH companies, while the fifth indicator is a new one proposed. Moreover, to be able to compare the different buildings, the energy and volume measurements used in the indicators are divided by the building's heated floor area. Regarding the SHM data, the measurements have hourly resolution. However, the analysis was performed with a minimum resolution of one day due to measurement truncation problems raised when calculating the ΔT. Additionally, some of the indicators were assessed for the seasons of heating (from January to May and from September to December) and no-heating (from June to August). The indicators are the following:

Ind. 1: Overall substation's operation: Assessing the annual average measurements (energy, volume, and ΔT) to understand each customer's overall performance.

Ind. 2: Temperature difference intervals: Determining the number of days with different ΔT-values during the heating and no-heating seasons.

Ind. 3: Heating and outdoor temperature correlation (energy signature): Analysing the relationship between the daily heating demand and the outside temperature, as described in Gadd and Werner (2015).

Ind. 4: Overflow: Assessing the overconsumption (V_{over}) of the different buildings using equation 1 and considering the ΔT_{ideal} constant for all buildings equal to 45ºC (Gadd and Werner, 2014).

$$V_{over} = V_i - V_{ideal} = V_i - \frac{E_i}{\rho c_p \Delta T_{ideal}} \; [m^3/m^2] \qquad (1)$$

This work also proposes a novel indicator (equation 2) to identify faulty substations by calculating the daily ratio between the measured water volume per m² (V_i) and the measured temperature difference (ΔT_i) as a function of the daily average outdoor temperature (T_{out}).

Ind. 5: Volume-temperature ratio is the proposed indicator to assess the heating installation behavior, and it is calculated with equation 2:

$$V_{Temp} = f(T_{out}) = \frac{V_i}{\Delta T_i} \; [m^3/m^2 \, ℃] \qquad (2)$$

According to the proposed indicator, a well-performing substation displays a ratio that changes linearly throughout the outdoor temperature conditions variation. Therefore, all data points that do not follow this linear profile are marked as faulty.

This research assesses how these indicators perform in analyzing the different symptoms measured by the SHM and attempts to correlate them with the faults identified during the visits to the installations. Furthermore, this study also endeavors to discuss the presented indicators' advantages and disadvantages and perform the fault impact assessment by estimating the energy savings obtained from the intervention made in the heating installations. The energy savings calculation is made through the difference in the sum of energy used before and after the visit divided by the building's heated area and the heating degree-days (in order to normalize it over the building size and outdoor temperature) and two case buildings are discussed using the indicator volume-temperature ratio.

Results and Discussion

Overview of the faults in DH substations

From a set of 351 fault assessment reports, only 158 reports of different residential buildings were used in the initial stage of the study. The visit to each installation was prompted because the SHM registered a ΔT below 10ºC, which according to the literature, is a symptom of an underperforming DH substation. Table 1 shows the types of faults, their frequency, and their identified causes. The categories are based on the ones proposed by Månsson et al. (2019).

Table 1: Types of faults and their description.

Fault categories	Nr. of faults	Cause of fault
Wrong general settings in the system	62 (39%)	Consists mainly of high settings in valves that control DHW heat exchanger/storage tank, SH system, etc.
SH system (Secondary side)	38 (24%)	Consisted mainly of defective components in SH

https://doi.org/10.26868/25222708.2023.1576

		systems inside the building (secondary side).
Control valves	31 (20%)	Consisted mainly of defective valves in the DHW heat exchanger/storage tank.
User behavior/practices	15 (10%)	Consisted mainly of sporadic usage of the SH systems by the occupants.
Controllers or sensors	7 (4%)	Consisted only of cases where the battery of the thermostat of SH systems in the building ran out.
DHW exchanger/tank	5 (3%)	Consisted only of defective DHW heat exchanger/storage tank.

Explicitly, the highest percentage of existing faults in the dataset are caused by high settings in the heating systems and defective components in the SH system indoors. This corroborates two important arguments in the field of FDD in the DH systems. Firstly, most of the existing faults on the end-user side can be easily solved by changing the systems' settings to better values representing a simple but rather impactful measure to increase the overall performance of the DH network (Gadd and Werner, 2015). Secondly, these values show that most faults occur on the secondary side of the heating circuit (i.e., inside the building instead of the substation). This implies that DH utilities need physical access to the customer's installation to fix most of its network faults. Utilities can accomplish this by establishing with the customers a service agreement on maintenance and repair of the heating installations, thus promoting good communication between the company and the customer and, in the long-term, maintaining a well-performing substation (Månsson et al., 2019).

After the intervention in each installation, some technicians were explicit in describing if the detected fault was solved. All the faults caused by occupants' abnormal use or high settings of the systems were fixed by the technicians, while the faults caused by a defective component that must be replaced were not fixed and must be addressed by a plumbing company hired by the customer. Figure 2, summarises the status of the heating installations after the visit, showing that 57% of the faults were solved, while only 27% could not be fixed by the technicians.

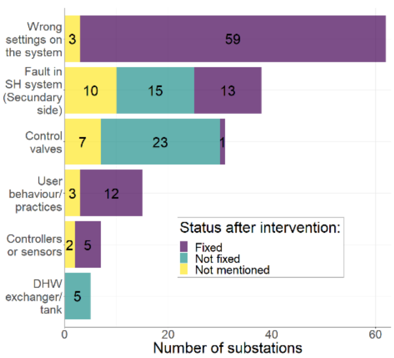


Figure 2: Type of each fault according to the installation's status after the intervention.

To summarize the described faults in the DH system at the end-user level, some labels are proposed in this study and applied further. Figure 3 shows these five proposed labels for each fault type.

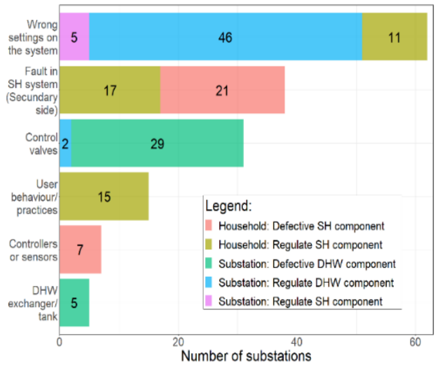


Figure 3: Type of each fault according to the proposed labels.

The labels are comprised of three parts. The first is whether the fault occurred in the central system (substation) or inside the building (household). The second part focuses on whether the fault refers to a broken component (defective) or the need to adjust the component's settings. The third part indicates which system the fault occurred in, either SH or DHW. This work also focuses in understanding which indicators are accurate in detecting and diagnosing the occurring faults in customer's level by assessing the system's symptoms through the SHM data. This task is performed by calculating and analyzing the indicators described in the *Methodology* section in the period prior and following the intervention.

FDD indicators performance assessment

Indicator 1: Overall substation's operation

The first tested indicator is the one used to understand the overall substation's operation. Figure 4, shows how the different fault labels are displaced according to the annual average volume and ΔT before and after the intervention. From these results, it is concluded that there is an overall increase of the ΔT and reduction of the energy and volume after the intervention – as expected.

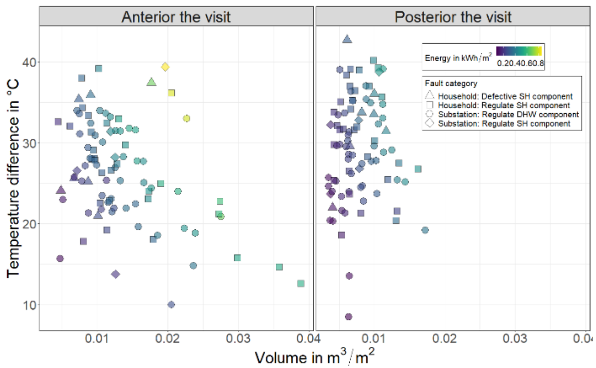


Figure 4: Representation of the annual average ΔT and water volume over energy usage and fault labels.

However, these results are as predicted when compared to the period before and after the intervention. There are a few remarks that must be highlighted regarding this

https://doi.org/10.26868/25222708.2023.1576

indicator. As observed in the period before the intervention, a few buildings have good average measurements of energy, volume, and ΔT (right-top corner). Therefore, this indicator would not flag this group of buildings as faulty in the first instance. Another remark is the existence of buildings with low ΔT after the intervention that have low heating usage due to low water volume consumption (bottom-left corner). One of the reasons behind this symptom is the DHW settings that were readjusted after the visit to higher values by the occupants. Another drawback of this indicator is that it must have a large sample of measurements for the average values of the variables to be representative. Therefore, this indicator is relevant to detect faults for extreme cases (i.e., buildings with extremely low/high energy, volume, and ΔT-values) but might not be enough for detecting faults that occur sporadically during a few days of a year (as it happens in occupancy-based faults and high systems' settings).

Indicator 2: Temperature difference intervals

Focusing on the temperatures measured by the SHM, it is investigated the density distribution of the daily ΔT-values for the heating and no-heating seasons before and after the intervention in the heating installations by the graphs in Figure 5 and Figure 6.

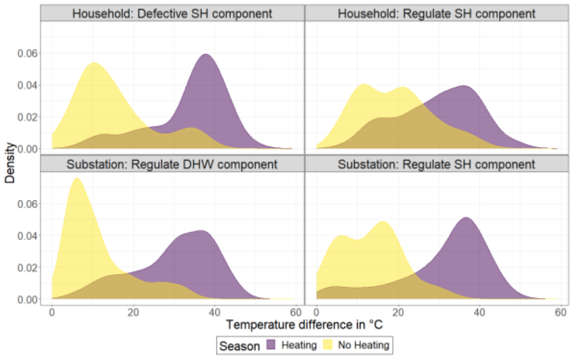


Figure 5: Density of ΔT during heating and no-heating seasons prior to the intervention.

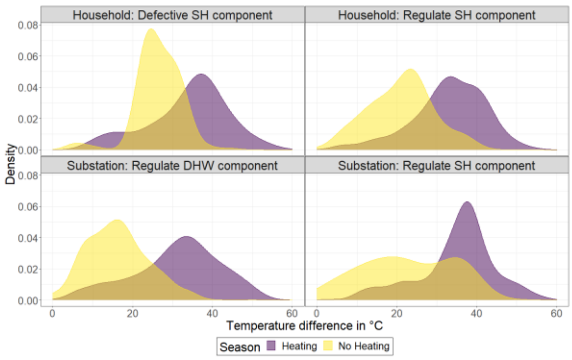


Figure 6: Density of ΔT during heating and no-heating seasons following the intervention.

According to the results, it is observed that all fault labels are characterised by lower daily ΔT-values for both seasons before the intervention. The main difference between both seasons is that during the heating season, the ΔT is higher than for warmer months. The main reason is that the SH and DHW systems are operating simultaneously during the heating season, while during the no-heating season, the heating demand is predominantly due to DHW usage. Therefore, it is also seen that the faults due to DHW decrease even more the ΔT measurements throughout both seasons, but mainly during the warmer months. It is also seen that after the visit, the overall ΔT increases for both seasons in all cases. However, following the intervention and the fault fixed, some buildings still have small ΔT, meaning that this indicator alone is not enough to assess the heating installation performance because it might highlight well-performing substations as faulty.

Indicator 3: Heating and outdoor temperature correlation

Gadd and Werner (2015) described that poor energy correlation with the outside temperature indicates poor substation control. According to Figure 7 and Figure 8, it is observed that after the visit, the correlation between the heating demand and the outdoor temperature increases, proving that this correlation is, in fact, related to the well or poor function of the substation.

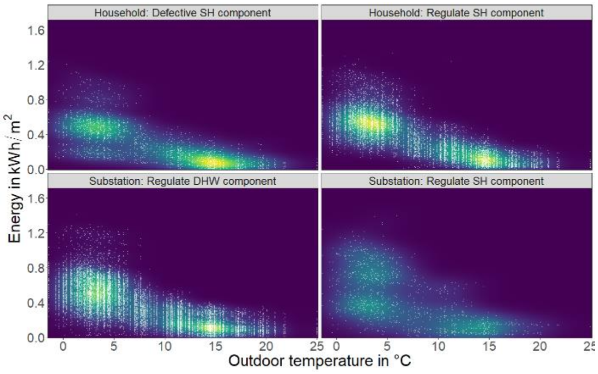


Figure 7: Energy signatures per fault category prior to the intervention.

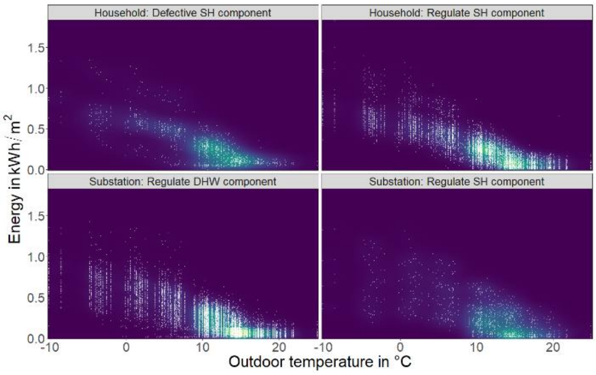


Figure 8: Energy signatures per fault category following the intervention.

For a better overview of the correlation increase after the visit, the coefficient of determination (R^2) obtained from the energy signature fitting was determined for each building before and after intervention for the heating and no-heating season (Figure 9).

https://doi.org/10.26868/25222708.2023.1576

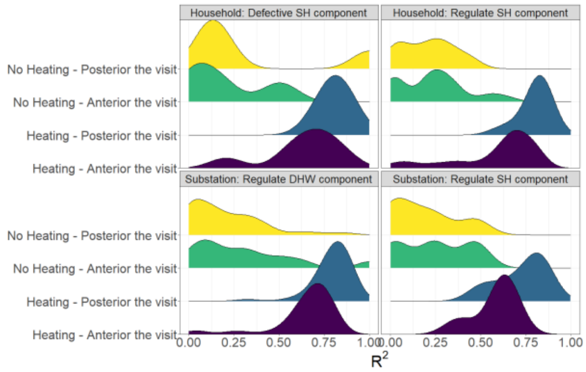


Figure 9: The R^2 of the energy signatures of each building per fault category for the different seasons prior to and following the intervention.

In Figure 9, one can see an increase of R^2 after the intervention for the heating season, while in the no-heating season, the R^2 tends to decrease toward zero. It is also observed that the correlation is not only dependent on the substation level but on the components in the building (e.g., underfloor heating control) because in all fault labels, a significant change in the R^2 is observed in both seasons after the visit. Therefore, it can be concluded that a well-performing system must have R^2 closer to one during the colder months while an R^2 closer to zero when only the DHW is mainly operating in the no-heating season. If this is not observed, it shows that the weather conditions are not the main driver of the substation's performance, but also a faulty component or occupants controlling the system poorly.

Even though this indicator is relevant for detecting faulty heating installations, it also requires a large sample of data points to describe this correlation accurately. Therefore, this indicator can only be used for long-time periods of collected data but not for short-term measurements. Another essential factor to consider for this indicator is that it can cause false alarms while attempting to detect faults due to the correlation being affected by two systems (SH and DHW) operating simultaneously. This is seen in the no-heating season, where few buildings have an $R^2 > 0.5$ after intervention due to the SH still operating (e.g., bathrooms with underfloor heating or water-heated towel dryers operating during summer).

Indicator 4: Overflow

The overflow indicator can be calculated as the difference between the measured volume by the SHM and the ideal volume when considered an ideal ΔT (usually 45 ºC). From all indicators above, the overflow does not need large data samples to be calculated, making it a great performance value for short-term measurements. As one can see in Figure 10, the overflow of four different buildings is calculated over time. From the results, it is clear that the overflow is higher for faults involving DHW systems when this system is mainly operating (non-heating season). While it is observed when the fault is due to a defective component, the overflow has a rapid large spike over a short period of time.

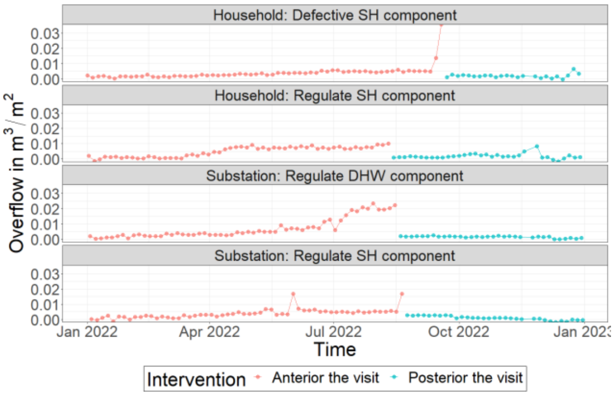


Figure 10: Overflow over time of four building cases per fault category prior to and after the intervention.

As expected, it is observed that there is a significant reduction of the overflow when a fault is fixed. This reduction occurs because the water volume measured in the building decreases while the ΔT increases. As seen, the overflow indicator is effective for DH customers' analysis for large and small sample measurements. Additionally, it can evaluate the installations and rank them based on their overflow (Månsson et al., 2019). However, this indicator has two drawbacks when applying it. Firstly, it is a value highly dependent on the ΔT variation without considering possible changes in the volume that are not dependent on ΔT. This is due to equation 1, where the ideal volume is calculated only considering an ideal ΔT. The second drawback is the predefinition of an ideal ΔT as a constant value. This preselection may hinder the comparison between buildings where their ideal ΔT might be different due to their location in the network, and for the same reason, buildings with lower ideal ΔT may be accounted as faulty when in reality, their substations are performing well.

Indicator 5: Volume-temperature ratio

Because of the necessity of comparing several buildings in the network and knowing that each building might have different ideal standards of volume and ΔT, this study proposes a new indicator to be used by the DH companies when assessing their customers. The indicator is called *volume-temperature ratio* and is based on the fact that there is a direct proportion between the volume usage and the ΔT, which is linear throughout the outdoor temperature variation when the substation is well-performing. This relation can be observed for four different building cases in Figure 11.

https://doi.org/10.26868/25222708.2023.1576

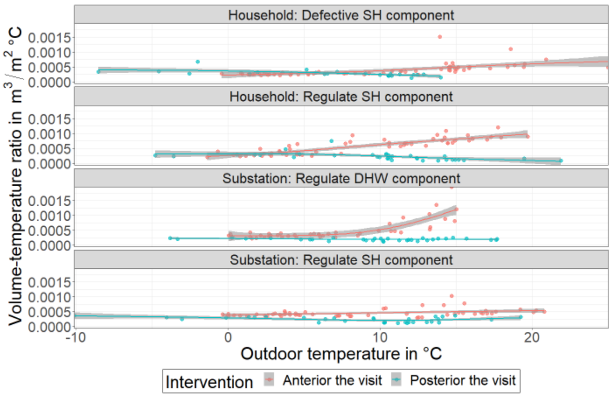


Figure 11: Volume-temperature ratio over outdoor temperature of four building cases per fault category prior to and after the intervention.

Figure 11 shows that for all the assessed buildings, the ratio of volume and ΔT should be constant for well-performing substations, regardless of the system. Besides, it is also noted that there are no significant differences between heating and no heating seasons. To better overview the values calculated from this ratio, Figure 12 shows the distribution of the volume-temperature ratio for each fault label before and after the visit. Compared to the overflow, the volume-temperature ratio attempts to solve the drawbacks of the former indicator while maintaining its benefits. Because there is no requirement to establish an ideal volume or ΔT, the ratio can be applied to compare all the buildings regardless of their location in the grid. It also, as the overflow, can be used to detect faulty singular data points without needing a large data sample.

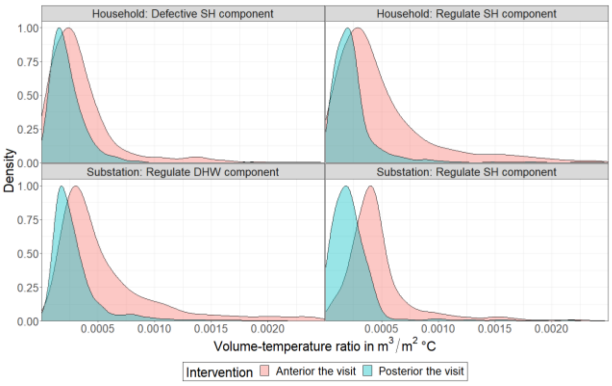


Figure 12: Density of the volume-temperature ratio for 158 buildings dataset prior to and following the intervention.

Energy savings due to FDD intervention

As previously shown above, many are the types of faults that can occur in the heating installations in a household connected to the district heating leading consequently to significant energy bills. From the initial dataset of 351 buildings, in 90 of them, it is mentioned in their reports that the fault was fixed by the technician or by the customer himself. Of these 90 buildings, 59 of them have positive energy saved after the intervention (refer to Figure 13), with an average energy saving of 14%, while 31 buildings do not report energy savings (negative values)

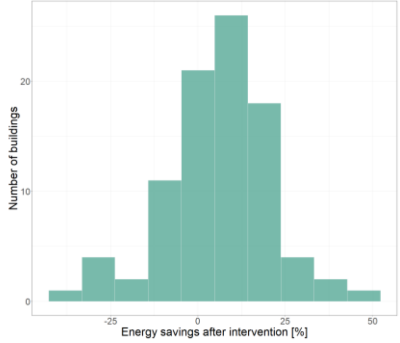


Figure 13: Energy saving after the intervention.

The main reason behind these negative values is due to some of the technical interventions were not enough to improve the system's performance, e.g., occupants continuing using the system suboptimal after the intervention, or operation changes.

It is selected two buildings where one had a large saving post-intervention (+25%), and the other displayed negative energy savings after the technician's visit (-30%). Both cases are analyzed with the indicator volume-temperature ratio, as one can see in Figure 14 and Figure 15.

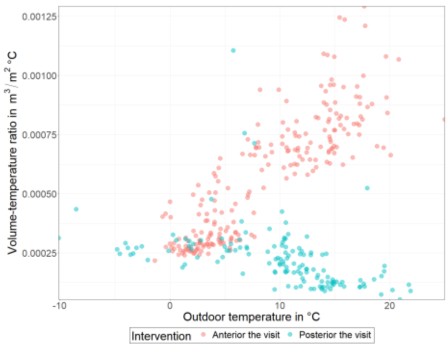


Figure 14: Volume-temperature ratio of a building with positive energy savings.

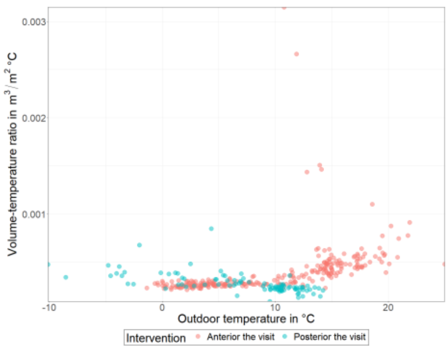


Figure 15: Volume-temperature ratio of a building with negative energy savings.

The first case pertains to a building where an incorrect setting was detected in the SH system, while the second case involved a defective radiator thermostat valve. These instances highlight the significance of identifying and diagnosing faults to ensure the efficient and safe operation of buildings and their systems. It was observed that after the fault being repaired, the data points conform to a linear trend, as expected for this indicator. Notably, the significant difference in energy savings between the two

https://doi.org/10.26868/25222708.2023.1576

cases is primarily attributable to the first case having a significantly greater number of data points that deviated from the trend compared to the second. This indicates that certain faults in DH customer installations may persist over extended periods while others may be momentary. Thus, the fault impact assessment per building must be performed on a data point basis rather than by time periods (e.g., pre- and post-intervention).

Conclusion

This study has demonstrated the importance of fault detection in DH systems and its potential impact on the buildings' energy usage. Through the analysis of 158 fault reports, it was concluded that the faults in the highest amount concerned high settings on the system by poor regulation or occupants' practices and defective SH system components indoors. This study also benchmarked different FD indicators using a subset of 90 buildings with faulty and non-faulty data points. A novel FD indicator based on water volume and temperature difference in DH substations was proposed, showing promise in detecting anomalies for future use by DH utility companies and integration into an automated fault detection and diagnosis (AFDD) framework. The findings suggest that DH system faults significantly impact energy use, and more efficient FD has the potential to reduce energy usage by customers by an average of 14%.

Suggestions for further work

There are numerous potential directions for further research in FDD in DH systems based on the findings of this study. One possible avenue is to explore the use of ML supervised algorithms to train classification models to analyze the relation of the volume-temperature ratio with its magnitude and occurrence period to diagnose different types of faults at the DH customer's level. Furthermore, researchers could analyze these indicators (e.g., energy signature, volume-temperature ratio, etc.) in the different existing heating system solutions implemented in buildings worldwide (e.g., direct/indirect connection, storage/instantaneous system, underfloor/radiator SH systems, etc.), and further on explore the potential for integrating FDD with other smart technologies in DH systems. However, to continue with this work, the DH companies must make a more extensive effort to collect good quality datasets with ground truth to progress in developing automated and implementable FDD methodologies.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 893945 (E-DYCE). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 958345 (PRELUDE). The authors would like to acknowledge Aalborg Forsyning for their support in data collection.

References

Andersen, K. H., Melgaard S. P., and Leiria D. (2023). *Summary of Existing FDD Frameworks for Building Systems*. DCE Technical Reports . Department of the Built Environment, Aalborg University (DK).

Calikus, E., S. Nowaczyk, A. Sant'Anna, and S. Byttner (2018). Ranking Abnormal Substations by Power Signature Dispersion. *Energy Procedia 149*.

Calikus, E., S. Nowaczyk, A. Sant'Anna, H. Gadd, and S. Werner (2019). A data-driven approach for discovering heat load patterns in district heating. *Applied Energy 252*.

Frederiksen S., Werner S. (2013). *District Heating and Cooling*. Studentlitteratur AB. Lund (SE).

Gadd, H., and S. Werner (2014). Achieving low return temperatures from district heating substations. *Applied Energy 136*.

Gadd, H., and S. Werner (2015). Fault detection in district heating substations. *Applied Energy 157*.

Guelpa, E., and V. Verda (2020). Automatic fouling detection in district heating substations: Methodology and tests. *Applied Energy 258.*

Li, H., and N. Nord (2018). Transition to the 4th generation district heating - possibilities, bottlenecks, and challenges. *Energy Procedia 149*.

Melgaard, S. P., K. H. Andersen, A. Marszal-Pomianowska, R. L. Jensen, and P. K. Heiselberg (2022). Fault Detection and Diagnosis Encyclopedia for Building Systems: A Systematic Review. *Energies 15*(12).

Månsson, S., K. Davidsson, P. Lauenburg, and M. Thern (2018a). Automated Statistical Methods for Fault Detection in District Heating Customer Installations. *Energies 12*(1).

Månsson, S., P. J. Kallioniemi, K. Sernhed, and M. Thern (2018b). A machine learning approach to fault detection in district heating substations. *Energy Procedia 149*.

Månsson, S., P. J. Kallioniemi, M. Thern, T. V. Oevelen, and K. Sernhed (2019). Faults in district heating customer installations and ways to approach them: Experiences from Swedish utilities. *Energy 180*.

Månsson, S., I. L. Benzi, M. Thern, R. Salenbien, K. Sernhed, P. J. Kallioniemi (2021). A taxonomy for labeling deviations in district heating customer data. *Smart Energy 2*.

Schaffer, M., T. Tvedebrink, and A. Marszal-Pomianowska (2022). Three years of hourly data from 3021 smart heat meters installed in Danish residential buildings. *Scientific Data 9*(420).

Xue, P., Z. Zhou, X. Fang, X. Chen, L. Liu, Y. Liu, and J. Liu (2017). Fault detection and operation optimization in district heating substations based on data mining techniques. *Applied Energy 20.*

https://doi.org/10.26868/25222708.2023.1576

5.3 Application of data patterns to diagnose DH building faults

<u>Paper 7</u>

"Is it returning too hot? Time series segmentation and feature clustering of end-user substation faults in district heating systems"

Daniel Leiria, Hicham Johra, Justus Anoruo, Imants Praulins, Marco Savino Piscitelli, Alfonso Capozzoli, Anna Marszal-Pomianowska, Michal Zbigniew Pomianowski, submitted in *Applied Energy*, 2024.

Is it returning too hot?

Time series segmentation and feature clustering of end-user substation faults in district heating systems

Daniel Leiria[1], Hicham Johra[1], Justus Anoruo[1], Imants Praulins[1], Marco Savino Piscitelli[2], Alfonso Capozzoli[2], Anna Marszal-Pomianowska[1], Michal Zbigniew Pomianowski[1]*

[1]Aalborg University, Department of the Built Environment, 9220 Aalborg, Denmark

[2]Politecnico di Torino, Department of Energy, TEBE Research Group, BAEDA Lab, Turin, Italy

*Corresponding author: mzp@build.aau.dk

Abstract

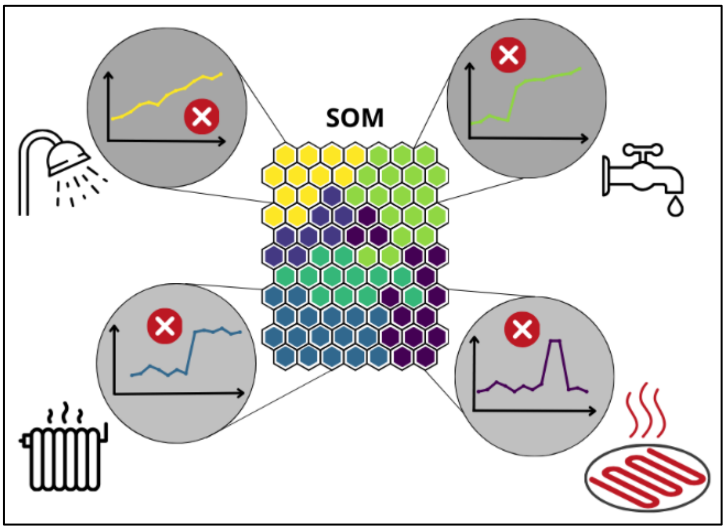
This study explores the challenges and advancements in collecting ground-truth data to enhance fault diagnosis models for district heating systems. Initiated by the need to address limitations in previous data collections, this research leverages an enriched dataset from Aalborg utility to identify faults in household substations. Despite some inaccurate fault categorizations, obscured fault patterns, and truncated measurements, the analysis of 50 detailed cases out of 127 fault reports reveals that while return temperature reliably indicates faults, energy usage patterns do not. Employing self-organizing maps combined with k-means clustering, the research categorizes fault symptoms and patterns, demonstrating the utility of high-dimensional data clustering in fault diagnosis. Additionally, an algorithm using time series decomposition is suggested to autonomously identify both extreme and subtle anomalies, enhancing fault detection capabilities. The paper concludes that these methodologies significantly augment the precision and dependability of fault diagnostics in heating systems, paving the way for more efficient operational management.

Keywords: Fault detection and diagnosis, District heating, Self-organizing maps, Unsupervised learning, Time series decomposition, Substation performance assessment.

Highlights:

- Self-organizing maps and k-means for fault clustering in heating systems.
- 50 cases were analyzed, revealing distinct features related to a fault occurrence.
- Demonstrating time series decomposition for autonomous anomaly identification.
- Highlighting challenges in data truncation and fault diagnosis accuracy.
- Proving the effectiveness of data segmentation in detecting various fault anomalies.

Graphical Abstract:



1. Introduction

In the European Union, district heating (DH) and cooling provide 12% of the energy to the building stock [1]. Despite this relatively small share, DH is considered vital for decarbonizing the heating sector, and its share is expected to increase in the following years [2]. With this implementation increase, one of the future goals in the DH sector is the integration of renewable energy sources for heat production instead of traditional fossil fuel-based ones [3]. However, lowering the fluid-supply temperature of the thermal grids is required to incorporate these alternative energy sources and industrial heat waste/surplus [4,5]. This is at the core of the 4th generation of DH systems, and it represents the next natural step toward decarbonizing the building sector.

The return fluid temperature is mainly dependent on the end-user side and how effectively it can extract heat from the heat-carrier fluid, which is dependent on the installation design, control, and the presence of faults in the building heating system [6]. In an attempt to maintain the return temperature at acceptable levels, the DH utility companies might charge their high-return temperature consumers an additional fee to motivate them to change their behavior or fix the existing system's faults [7,8]. Nevertheless, research has been conducted to investigate whether utilities should change their business model in providing their services and expertise to their clients by offering constant monitoring to detect, diagnose, and solve faults [8,9]. Aligning the business model to establish a closer relationship between utilities and consumers can expand access to building heat substations and secondary heat systems, thus enabling a deeper understanding of the different faults that might occur in the latter. Combining this new set of information on the customers' installations with the smart heat meters (SHM) data that has been already collected on a large scale in several countries (e.g., Denmark, Sweden), can lead to the next step of fault detection and diagnosis (FDD) in the DH, by consolidating the causality between anomalous symptoms observed in the measurements and the actual cause of these symptoms (ground truth). Having this ground truth is vital to the development of automated FDD algorithms that can be deployed at a large scale by utility companies, thus improving even further their business model and the sustainability of the DH system by keeping their customers' installations at optimal performance [10,11]. Moreover, the benefits of a DH network optimal performance obtained by the automation of the FDD processes are the following:

- Optimization of energy usage and system efficiency: FDD can help optimize energy usage by ensuring that all components of the DH system are functioning correctly. This optimization leads to improved overall system efficiency, reducing energy waste and lowering operational costs [12].
- Early detection of faults and anomalies: FDD algorithms can detect potential faults or anomalies early on, allowing DH operators to take corrective actions before these issues escalate into more significant problems. Early detection helps prevent major system failures and prolongs the lifespan of DH components [13].
- Minimization of downtime and service interruptions: Early detection and diagnosis of faults enable prompt intervention and maintenance activities, minimizing downtime and service interruptions. This ensures a more reliable heating supply to end-users, enhancing customer satisfaction [13].
- Effective prioritization of maintenance activities: Through the identification of recurring issues, DH operators can prioritize maintenance activities more effectively. This targeted approach to maintenance ensures that the most critical issues are addressed first, optimizing resource allocation [14].
- Reduction of operating costs: By reducing the need for emergency repairs, FDD helps lower operating costs for district heating providers. Preventive maintenance enabled by FDD reduces the frequency and severity of repairs, contributing to cost savings [15].

1.1. Related work

Two different review scientific articles [16,17] were published to map out the current developments of data-driven FDD methodologies in the DH sector. One of their main conclusions is that to push forward this field, more high-quality data needs to be collected and comprise ground truth about fault occurrence and nature. Despite this lack of a large-scale collection of labeled data of occurring faults complemented with SHM measurements, several studies propose different FDD algorithms to be implemented in the DH grid. This subsection reviews a few of these methods targeting the DH end-users (buildings connected to the grid).

One of the first works conducted in this field is [18], where a statistical assessment was performed for 135 substations in different types of buildings to find the ones with at least one of the three fault symptoms: anomalous heat patterns, low average annual temperature difference, and poor substation control (i.e., poor correlation between heat demand and outdoor temperature – energy signature). Around 75% of buildings presented at least one of these symptoms showing the potential and need for systematic fault detection campaigns. Similarly, the authors proposed in [5] a novel threshold-based method for fault detection using the temperature difference signature (relation between the DH temperature difference of supply and return and outdoor temperature). Calikus at al. (2018) employ a similar method to spot hindered building substations via their energy signatures and introduce a ranking method to sort the substations with large abnormality symptoms [19]. These methods are all dependent on set thresholds triggering alarms. This means that they cannot account for the dynamic nature of heating systems, where normal operating ranges can vary significantly under different conditions, leading to false alarms or undetected faults [20]. Contrastingly, modern FDD methods leverage other machine learning (ML) algorithms to overcome these limitations: they can learn from the historical data to identify specific patterns and anomalies [21].

Within ML, supervised learning trains algorithms on a labeled dataset containing ground truth on the target outputs, meaning that it learns from data that already contains the answers (outputs) associated with given inputs. This approach is particularly powerful for predictive tasks, as the model can infer relationships between features and outcomes, making it capable of making predictions on new, unseen data. In the context of FDD for buildings connected to the DH, Månsson et al. (2018) developed a model of well-performing DH substations using gradient boosting regressor and SHM data to detect deviation patterns indicating possible faulty operation [22]. Similar works can be found in [23-27]. On the other hand, unsupervised learning deals with identifying patterns in a dataset without pre-existing categorization, i.e., the algorithm tries to infer the structure from the dataset itself. In FDD for buildings connected to DH systems, unsupervised learning is particularly useful for identifying unusual patterns or anomalies that could indicate faults. For instance, clustering techniques can group similar operational profiles of heating systems and identify outliers. These outliers represent operational anomalies, which could be linked to system failures or distinct systems operation by the occupants. Due to the lack of labeled datasets with ground truth on faults, unsupervised learning is, by far, the dominating method in the field. Calikus at al. (2019) clustered heat profiles using k-shape algorithm, and the anomalous profiles were segregated and investigated further [28]. Xue et al. (2017) employ a methodology that uses several clustering algorithms with association rules analysis to find abnormal behaviors in substations from SHM data [29]. Other unsupervised methods for FDD in DH data can also be found in [30-35].

Lastly, there is a bigger trend toward implementing deep learning techniques to analyze buildings connected to the grid to diagnose suboptimal substations. Deep learning is a subset of ML that utilizes artificial neural networks to model complex patterns and relationships in data and can be employed for supervised and unsupervised learning purposes. These models are particularly effective in handling large volumes of data and time series. Choi and Yoon (2021) developed an autoencoder to generate relevant features and detect faults and applied a multilayer perceptron (MLP) for classifying faults in a multi-family residential building in South Korea [36]. Kim et al. (2021) propose a DH substation fouling detection and diagnosis method using k-means clustering, MLP, and virtual sensor-assisted. The k-means method is used for pattern identification to segment data for training and testing, while an MLP model, incorporates measurements from virtual sensors, and predicts system variables to detect fouling based on threshold violations [37]. Other deep learning methods applied in FDD are proposed in [38-40]. Despite the merits of these methodologies, they typically share a significant limitation: they do not provide a comprehensive view of fault occurrences at both the substation and building (indoors) levels. Studies of system faults, involving diverse sensors and fault types, tend to cover only a limited number of building cases. Conversely, studies encompassing a broad range of buildings generally offer less detailed information about, among others, occupants, fault types, or heating installations. This issue is widely acknowledged in the field as a major challenge [12, 41-42]. To address this, a standardized methodology for collecting fault label information was proposed [43]. Furthermore, van Dreven et al. (2021), due to the need for this type of data, have developed an experimental setup to gather ground truth data from faulty heating systems. This initiative aimed to generate high-quality data that distinguishes between normal and abnormal system behaviors, identifying the specific faults associated with different operational profiles [44]. Despite this advancement, the overall landscape of ground truth data remains scarce and this knowledge gap underscores the necessity for more extensive and varied data collection efforts to enhance the reliability and accuracy of FDD processes in DH systems.

1.2. Contributions and novelty of the present study

In response to this need, Aalborg Forsyning, a Danish DH utility company, launched its own initiative to engage with key customers on their heating grid who exhibited low-temperature difference between supply and return. During these engagements, technicians visited these households, diagnosing their primary issues and documenting them with any corrective actions taken in intervention reports. Consequently, access to these reports, coupled with the measurements collected from the buildings' SHM, permits to drawing of correlations between the heating data and the nature of the faults identified. Thus, the work presented in this article advances the field of FDD in the DH sector with the following:

1. Explanation of the characteristics and challenges in collecting fault reports for ground truth in fault diagnosis models:

This article in sections 2.1 and 2.2 sheds light on the challenges arising when collecting fault reports to create models that can classify and diagnose system faults. It delves into the complexity of establishing a robust ground truth dataset, underlining difficulties such as variable reporting standards, the subjective nature of human diagnoses, and the impact of incomplete data.

2. Comprehensive analysis of SHM data from 50 DH household substations with verified faults:

A thorough investigation is presented on SHM data collected from a cohort of 50 DH substations. Each dwelling substation has documented instances of malfunctions verified by a technician. This meticulous analysis described in sections 2.3 and 2.4 and presented in section 3.1 aims to correlate specific data patterns with the confirmed faults, thereby refining the predictive accuracy of maintenance protocols.

3. Proposal for employing a time series decomposition methodology to identify and isolate anomalous data:

A proposal described in section 2.5 and presented in 3.2 is made advocating for the utilization of a time series decomposition method, already developed for satellite image recognition [45], aiming to identify and delineate anomalous data within monitored systems. By decomposing the time series measurements into its fundamental components—trend, seasonality, abrupt changes, outliers, and residuals – the methodology possibilities to segment the data that signify operational faults, thus supporting this applicability in DH fault detection systems.

4. Introduction of a clustering framework for categorizing symptoms and patterns of faults in DH customer systems:

The paper proposes a clustering methodology in section 2.6 with its results in section 3.3 to systematically group the diverse symptoms of faults reported by DH customers. This novel approach, based on self-organization maps (SOM) and k-means, lays the groundwork for the initial stages of fault diagnosis, potentially streamlining the identification process and enhancing the accuracy of subsequent maintenance efforts.

5. Discussion of lessons learned and the next steps for automated FDD integration in DH systems:

The paper concludes in section 3.4 with a discussion of the insights gained from this research and outlines the next steps necessary for the optimal integration of automated FDD processes in DH systems. It emphasizes the importance of leveraging the clustering framework to improve the accuracy and efficiency of fault detection and suggests future enhancements in data collection, analytics, and real-time monitoring to fully realize the potential of FDD in these systems.

1.3. Outline

Following the introduction in section 1, section 2 describes the methodology behind the data collection and treatment process, the manual and automated time series segmentation for fault detection, and the applied clustering algorithm for fault diagnosis. The results from the investigation and discussion on the lessons learned are presented in section 3. The article closes with conclusions in section 4.

2. Methodology

This section presents in section 2.1) a description of the case study and how the data was collected from SHM with its associated fault reports filed by technicians; 2.2) the major challenges encountered with these datasets; 2.3) a brief explanation of how the data was pre-processed for analysis; 2.4) an overview of the manual analysis conducted to inspect and visualize various measurements alongside their associated faults; 2.5) a suggestion on how to automate fault detection from these SHM data; and 2.6) a proposal of a method to cluster the studied faults according to their measurement and diagnose them according to their systems.

2.1. Case study description

Regarding the DH metering in this study, a standard SHM comprises several key components: two temperature sensors, a flow sensor, and an integrated computer that calculates and transmits the energy data. The flow sensor records the water flow through the primary side, while the temperature sensors measure the temperatures of both the supply and return hot water on the primary side. The meter calculates the energy transfer from the primary side to the secondary side (customer), and this energy is the one billed by the heat provider. The control systems and sensors on the secondary side are generally not owned by the utility company, making these data often unavailable.

This study exploits SHM data from 127 residential buildings connected to the DH network, along with 356 fault assessment reports. All buildings are residential, predominantly single-family homes, located in Aalborg municipality (Denmark). They are equipped with space heating (SH) systems, which may consist of radiators and/or underfloor heating (UFH), and domestic hot water (DHW) production based on heat exchangers or storage tanks. Originally, the measurements are recorded on an hourly basis resolution with some buildings having measurements from 2017 until 2023, depending on their SHM installation date. The fault reports in this study encompass the assessment outcome from the utilities technicians after visiting the faulty installations of the targeted consumers. This assessment was made during on-site visits or telephone calls with the occupants. This reporting process started in 2022, and the report's structure was simplified to optimize completion by the technicians. These reports, therefore, consisted of only three important inputs (see Table 1).

Table 1: First iteration of the fault assessment structure.

Parameters	Type of input	Definition
Meter ID	Individual single number	A unique identifier of the SHM installed in the building
Assessment date	Date	The date of the technicians' visit, formatted as day/month/year
Fault description	Open text answer	Open-ended comments for the technician to describe the fault

This report format was the one investigated in [42], and as this study has shown, having open-text comments to report a fault might cause ambiguous descriptions depending on the degree of detail written by the technicians. Furthermore, this type of reporting makes it cumbersome to select and group similar faults when coding, as their description text, is different from each other. Therefore, a second iteration of the faults reports format was developed and applied afterward by the DH company. This new format of the report includes a combination of multiple drop-down menus featuring pre-set options of possible faults, along with a section for free-text comments to be used if necessary. Originally composed in Danish, these reports have been translated into English and subsequently reviewed throughout this analysis for accuracy. One can see in Table 2, the structure of such new report.

Table 2: Second iteration of the fault assessment structure.

Parameters	Type of input	Definition
Meter ID	Individual single number	A unique identifier of the SHM installed in the building
Assessment date	Date	The date of the technicians' visit, formatted as day/month/year
Contact type	Predefined multiple-choice answer	The method used to contact customers, with the given options: • Telephone/E-mail • Physical visit
Hydraulic connection type	Predefined multiple-choice answer	The existing type of DH connection to the heating systems in the building, with the given options: • Direct • Indirect
SH system	Predefined multiple-choice answer	The existing type of SH systems in the building, with the given options: • Radiators • UFH • Combined (both radiators and UFH)
DHW system	Predefined multiple-choice answer	The existing type of DHW system for heat production in the building, with the given options: • Heat exchanger • Storage tank
Faulty component	Predefined multiple-choice answer	The component where the fault was identified by the technician, is categorized into specific labels, with the given options: • In SH system: o Pressure differential regulator o Radiator thermostat o UFH shunt o Etc. • In DHW system: o Temperature regulation valve o Incorrect settings in the temperature regulation valve o Incorrect pump settings o Etc.
Fault description	Open text answer	Open-ended comments for the personnel to describe in detail the fault.
Fault identification status	Predefined multiple-choice answer	The status of fault analysis, with the given options: • Proven (fault identified and confirmed by the technician) • Suspicion (unverified assumption of a fault by the technician)
Technician action	Predefined multiple-choice answer	Actions undertaken to rectify the fault, with the given options: • Error is resolved (fault fixed) • The customer must contact VVS (customer advised to contact a plumber to resolve the fault) • No action (no measures taken)

Despite the extensive scope of the second iteration dataset, it is not without its limitations. The following section outlines the challenges encountered, which could potentially complicate the data analysis and interpretation.

2.2. Challenges

Inconsistent standards and quality of the reports: The dataset suffers from a lack of uniform standards and quality control. This is observed even after the implementation of the second iteration reporting format, e.g., some technicians might choose different answers for the same fault. Moreover, some technicians tend to have different levels of detail when providing information in the fault description commentaries, e.g., not mentioning if a specific valve is broken in the fully open or closed position. This inconsistency introduces variability that complicates the analysis of SH and DHW systems across the diverse range of residential buildings when combined with the SHM data and the specific patterns these faults generate.

Multiple faults occurring at the same time: In some cases, there might be more than one fault during the intervention visit; however, the technician only described one of them, e.g., a broken SH system component with high settings in the DHW production system. "High settings" refers to the operational parameters of the DHW system, such as temperature or flow rate, being set to higher-than-normal levels. These settings can strain the system and lead to inefficiencies or additional faults. However, during a visit, if these two faults are found, a technician might opt to report only the broken component and not address the high settings in the system. Thus, it becomes difficult to discern the full extent and development of each individual fault's pattern. As this complexity can mask the interactions between the different faults and their impact on the overall DH system's performance.

Timing of interventions: The interventions, often occurring soon after the detection of a fault, may prevent the complete evolution of the fault pattern from being captured. As a result, the SHM dataset may not fully reflect the progression and potential impact of these faults over time. In terms of energy efficiency, a quick intervention is positive but not for the aspect of collection, analysis, and categorization of faults.

Most of the faults go unnoticed by residents: Most of the faults in this dataset have likely persisted over long periods without causing noticeable disruptions to the occupants. Therefore, the dataset lacks instances of faults with immediate and evident consequences, such as system leakages or deficiency in DHW production, which prompts a more urgent response by the dwellers, therefore being solved before DH company takes any intervention measures.

2.3. Data pre-processing

Initially, the dataset comprised hourly measurements, which were subsequently aggregated into daily values. This transformation was crucial for two main reasons. Firstly, it helped reduce computational complexity, making the dataset more manageable for analysis. Secondly, it addressed issues arising from potential data truncation by utility companies [46]. Hourly measurements often suffer from truncation (measurements being rounded down to the nearest integer), leading to inaccurate representations of energy, volume flow, and temperature measurements. By aggregating them into daily values, the impact of truncation errors is minimized, and it can be obtained more reliable data for analysis. Nevertheless, despite the conversion to daily values, the dataset still contained a few erroneous values, primarily due to truncation. To rectify this, these incorrect data points are replaced with 'NA', denoting missing measurements.

The next step in this process involved expert analysis to segment the time series corresponding to fault occurrences. Given the complexity of the systems under study, multiple operating systems (e.g., radiators combined with DHW production) could potentially contribute to fluctuations in measurements. Therefore, expert analysis by the authors was necessary to pinpoint the most reliable periods when faults occurred. This segmentation process enables to isolate and focus on the specific intervals relevant to fault analysis, facilitating more targeted and insightful investigations. To maintain the integrity and reliability of the dataset, buildings with incomplete data – marked by missing measurements – were excluded. This exclusion criterion ensured that the dataset used for analysis comprised only complete and accurate information, minimizing the risk of bias or erroneous conclusions. Finally, the dataset was refined by filtering out instances where fault reports lacked verification from on-site technician visits. Only faults confirmed and assessed by technicians were retained for analysis. Additionally, fault reports were considered valid only if the fault was conclusively "proven" in the report as the genuine cause of the faulty measurements. This stringent filtering process ensured that the dataset contained high-quality, verified fault labels, enhancing the credibility and reliability of subsequent analysis. By undertaking all these steps, the original dataset of 127 SHM and 356 fault reports was reduced to 50 cases with SHM data with combined reports.

2.4. Manual detection and analysis of the fault patterns

This section aims to provide insights into the patterns associated with various faults and how these patterns correlate with the operational metrics of the heating system. This is achieved by analyzing individually each SHM time series of the 50 cases. By exploring each case individually, the analysis seeks to highlight the specific occurrences and characteristics of fault patterns, offering a comprehensive overview of their temporal distribution, their impact on the different measurement variables, and their overall significance in system performance.

This analysis begins by aggregating the data to count and categorize occurrences by fault label, identifying the frequency of each fault across all monitored buildings. This step helps in understanding which faults are most prevalent within the system (see Table 4). Next, it is determined the month in which each fault is predominant. This calculated month is neither the month when the fault starts nor when it is identified by the technicians. It is instead the mean value between these two. E.g., if the fault starts in January and the intervention date (technician's visit) is March, this value will be assigned as February. This enables us to identify any potential seasonal trends that could influence system performance (see Figure 2).

During the manual analysis of the time series of each case, a key characteristic that emerged was the principal pattern of the measurements recorded during the fault period. The return temperature, volume flow, and energy were the analyzed variables. By examining these variables, it was consistently observed specific patterns manifest across each one (as one seen in Figure 1).

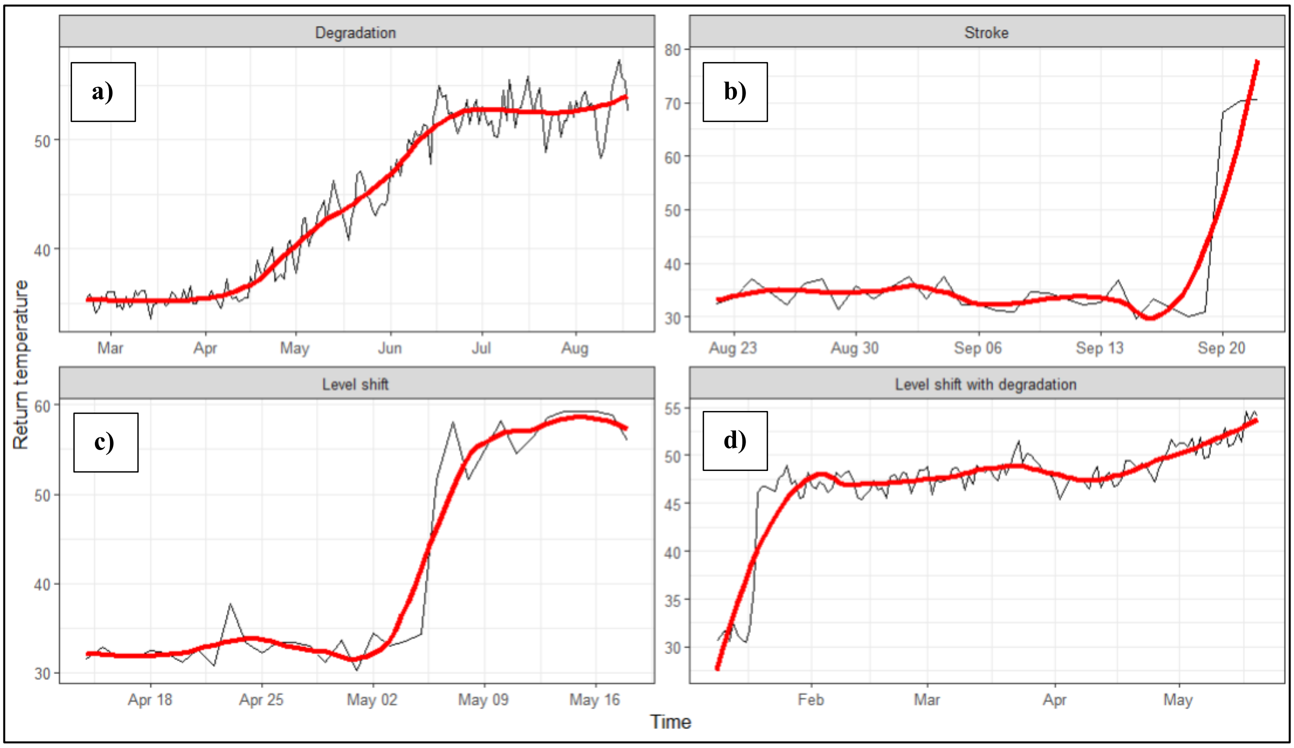


Figure 1: Representation of the different patterns (free scales).

As interpreted from the figure, various patterns in the measurement variables during the fault period become discernible:

a) The degradation pattern is characterized by a gradual and sustained increase in the variable over time, signifying a progressive decline in system performance. This is the most extended pattern among those identified, indicating a slow beginning of the fault.

b) In contrast, the stroke pattern emerges abruptly, marked by a rapid and intense change. This pattern represents a swift and significant event, suggesting a sudden fault occurrence.

c) The level shift pattern is akin to the stroke in its immediate change; however, it distinguishes itself by stabilizing at the new level post-fault. This implies that after the initial sharp increase, the variable maintains a consistent higher value.

d) A level shift with subsequent degradation is observed when, following the initial abrupt increase, the variable continues to trend upwards. This denotes a compound fault scenario where the system not only experiences a quick shift but also continues to degrade over time.

It is noteworthy that the stroke pattern may not always be as it appears. In some instances, what is identified as a stroke could in fact be a level shift misidentified due to the proximity of the intervention date to the fault event (as mentioned in one of the challenges for this type of data in section 2.2). Hence, these cases may exhibit similarities. Additionally, it is crucial to mention that there are instances where no discernible pattern is observed, which are categorized as "Not observed". This lack of pattern can be as significant as the others, indicating scenarios where the fault's manifestation does not conform to the typical behaviors captured by the other categories. Further analysis explores whether the observed patterns in each measurement's variables are associated with each other (see Figure 4). Also, the patterns are investigated to identify whether or not they are repeated across different periods (see Figure 5).

Furthermore, the duration of each fault and the difference between the starting and ending measurement during the segmented period are calculated (see Figure 6 and Figure 7). Lastly, another characteristic that was calculated in the different segments is volatility (see Figure 8). The volatility (Equation 2) was calculated as the standard deviation of all the time series measurements obtained by Equation 1 for each measured parameter:

$$y_t = \log(x_t) - \log(x_{t-1}) \tag{1}$$

$$volatility = \sqrt{\frac{\sum(y_t - \bar{y_t})^2}{n-1}} \tag{2}$$

Where x_t and x_{t-1}, are the measurements of a specific parameter, e.g., return temperature, at a specific day and its prior. While the value y is the logarithmic difference between these two values, and volatility is the standard deviation of the all-calculated y-values. The volatility is obtained in this form to disregard the trend in the time series, and only account for the daily variation. Another consideration in this calculation is the period where it is calculated. If a pattern is identified as a degradation, its volatility is calculated when the degradation starts. If the pattern is a level shift or a level shift with degradation, the volatility is calculated after the larger spike. In all buildings with a stroke pattern, the volatility is not calculated.

2.5. Automated detection of the fault patterns (time series segmentation)

The FDD analysis discussed in section 2.4 heavily depends on the segmentation and pattern recognition of time series data. Originally, the segmentation and pattern recognition tasks were performed manually by the authors based on expert knowledge, which is not scalable to larger datasets. To manage these extensive datasets effectively and identify data segments based on their main features, the application of BEAST method from "Rbeast" package [45] is suggested for further research and implementation on automated FDD processes. BEAST (Bayesian Estimator of Abrupt change, Seasonality, and Trend) is tailored for time series analysis and decomposition, helping to solve common issues such as trend detection, seasonal adjustment, and identifying change points. Originally developed to analyze land dynamics from satellite data, BEAST decomposes time series into trend, seasonal, and residual components using Bayesian models. This decomposition is crucial for uncovering underlying patterns, such as consistent seasonal effects or long-term trends. The package efficiently handles various types of seasonality, trends, and noise, making it highly adaptable to different time series structures. A key feature of BEAST is its capability to detect abrupt changes, known as level shift patterns, and identify outliers. BEAST employs the Bayesian averaging technique in an ensemble method to robustly estimate model parameters, accommodating uncertainty and variability in time series analysis and detecting the outlier points that deviate significantly from the fitted model.

While this study employed manual segmentation and pattern recognition for the FDD analysis, the proposed application of the BEAST methodology offers a valuable tool for future research. This approach would be particularly beneficial for analyzing larger datasets, as it automates the segmentation and pattern recognition tasks, improving scalability. In this article, it is demonstrated BEAST on two building cases serving as a proof of concept for its effectiveness in identifying trends, outliers, and change points within DH data. The decomposition process in this analysis did not consider seasonal components, focusing only on the trend, outlier detection, and abrupt changes.

2.6. Self-organizing maps and clustering

Further on, it is utilized SOM [47] to visualize the features identified in the dataset during the manual analysis in section 2.4. This method is used primarily for dimensionality reduction and is particularly well-suited for visualizing high-dimensional data. SOM helps to simplify complex, nonlinear statistical structures into simple geometric relationships in a low-dimensional space.

In this study, SOM is employed in R [48], opting for a hexagonal grid configuration for its nodes. This particular layout was chosen for its ability to maintain more uniform distances between nodes, which is essential for the accurate representation of the input data's topological features. When determining the size of the grid (number of nodes), it was aligned with the intricacy of the data at hand. A more extensive grid captures a higher level of detail, which simultaneously requires increased computational efforts and more extensive data for effective model training. Thus, a 4x4 grid was ultimately selected, which proved to be a balanced decision that reduced quantization error and boosted the explained variance, ensuring that almost every node was populated with data points.

The input data consisted of 17 variables retrieved from section 2.4, which encompassed both one-hot encoded and numeric variables. The one-hot encoded variables were utilized to represent categorical data, which were converted into binary form, encoding the presence or absence of certain characteristics. One can see in Table 3, each variable used in SOM.

Table 3: SOM input variables.

Input category	Input	Type
Pattern – Return temp.	Stroke	One-hot encoding
	Level shift	
	Degradation	

	Level shift with degradation	
Pattern – Volume flow	Stroke	One-hot encoding
	Level shift	
	Degradation	
	Level shift with degradation	
	Not observed	
Pattern – Energy	Stroke	One-hot encoding
	Level shift	
	Degradation	
	Level shift with degradation	
	Not observed	
Volatility – Return temp.	Calculated with Equations 1 and 2	Numeric
Volatility – Volume flow	Calculated with Equations 1 and 2	Numeric
Volatility – Energy	Calculated with Equations 1 and 2	Numeric
Season (fault period)	Values ranging from 0 to 5, where the lower numbers represent colder months and higher numbers represent warmer months	Numeric

Upon the implementation of the SOM on the dataset, it is integrated k-means clustering [49] to further segment the data. The k-means algorithm was calibrated to form 5 distinct clusters, this number having been identified as optimal due to its association with the highest silhouette score, an indicator of cluster cohesion and separation. The clustering results underwent a comparison against the ground-truth data derived from fault reports, providing meaningful and validated insight into the underlying structure of the fault characteristics.

3. Results and Discussion

This section presents the results of the analysis and discusses the challenges of identifying and diagnosing faults within the DH systems at the consumer level. Through the manual analysis of the return temperature, volume flow, and energy measurements from the SHM, it was identified distinct patterns indicative of system irregularities. These patterns – degradation, level shifts, and unexpected fluctuations serve as indications of potential faults. This investigation does not merely catalog these occurrences but also examines their durations, frequencies, and volatility within the data. Supplementing the manual inspection, it is suggested a detection algorithm that employs statistical methods to pinpoint changes in the measurement profiles and quantify these deviations automatically. The combination of manual and algorithmic approaches provides a comprehensive overview of the system's performance, enhancing our ability to rapidly detect, identify, and mitigate faults.

3.1. Manual detection and analysis of the fault patterns

This section focuses on the initial manual examination of the DH substation operational data to highlight anomalies and trends that may indicate faults. By plotting the number of cases per fault label and mapping them onto the temporal axis, it was discerned patterns and irregularities specific to the time of occurrence. Such visual insights not only aid in the identification of immediate issues but also contribute to a deeper understanding of the system's behavior over time. One can see in Table 4 the distribution of fault type labels for the considered 50 cases.

Table 4: Number of cases with a specific fault label.

Fault label	Number of cases
DHW HEX – Defective	3
DHW HEX – High settings	2
DHW HEX – No details	4
DHW Tank – High settings	2
DHW Tank – No details	4
Radiator – Defective	1
Radiator – High settings	14
Radiator – No details	6

Towel dryer – High settings	4
UFH – Defective	2
UFH – High settings	1
UFH – No details	7
Total	**50**

The categorization of faults is closely tied to the system component and the nature of the fault – 'Defective' implies a broken component, 'High settings' are generally ascribed to user preferences, and 'No details' signify insufficient information in the fault reports. All recordings in the dataset are consistent across the faults related to DHW heat exchangers (HEX), DHW storage tanks, radiators, towel dryers, and UFH systems. From Table 4, it is apparent that 'Radiator – High settings' represent the most prevalent issue within this dataset. Faults in 'Radiator' and 'DHW HEX' span across all subcategories, reflecting a diversity in the types of issues encountered. Conversely, 'UFH – High settings' and 'Radiator – Defective' are the least common faults noted.

A significant observation is the substantial number of reports marked with 'No details,' which substantially impedes comprehensive analysis and illustrates a broader issue with the reporting process. The prevalence of under-detailed reports underscores the necessity for a more structured and explicit reporting procedure.

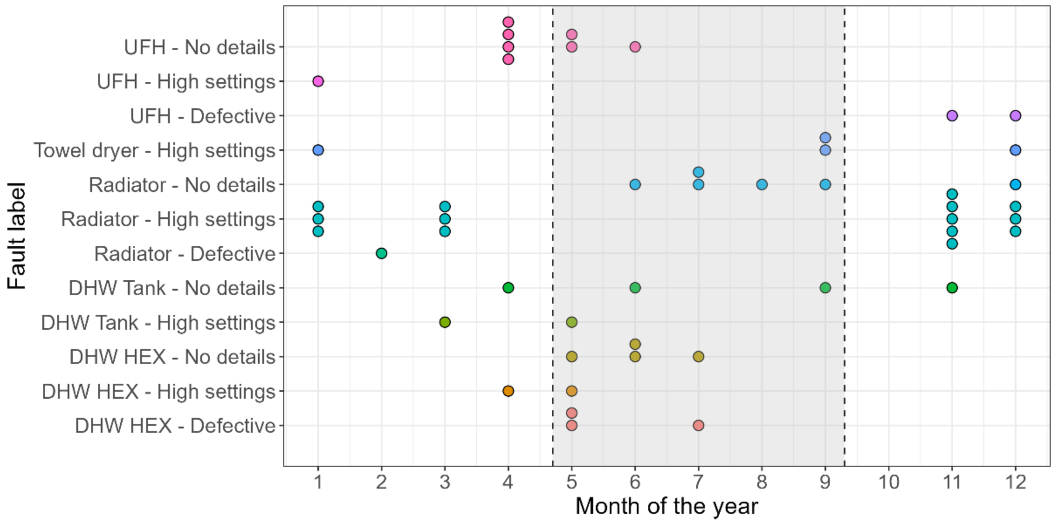


Figure 2: Representation of the month where each fault occurred per building. The grey shading is the months from May to September which is established as the summer season (no SH usage expected).

Figure 2 presents a time-based overview of fault occurrences across different buildings, overlaid with a grey shading area that denotes the warmer months from May to September – a period traditionally understood as having a small or no share of SH usage. One can observe that faults related to DHW production are more frequent during the summer season, as highlighted by the grey area. In contrast, issues with SH systems predominantly arise during the colder months, outside of the shaded area. This seasonal pattern suggests a correlation between the demand for specific heating services and the emergence of faults.

Despite the clear seasonal trend, some anomalies deviate from the expected patterns. These exceptions can be attributed to several factors. First, faults are not confined to any season and can occur due to random component failures, independent of the time of year. Additionally, human behavior introduces unpredictability. Instances of radiator usage, and consequently, faults, during summer months highlight this. This contradicts the common assumption that radiators remain dormant in warmer weather. Similarly, UFH system usage in summer, driven by personal preferences like wanting warm bathroom floors, can lead to unexpected faults.

Furthermore, seasonal variations in DHW and SH usage further complicate fault detection through system monitoring data. Because the proportion of SH and DHW contributions to the SHM readings changes throughout the year, faults in one system might be masked by the dominant operation of the other. During winter, the higher demand for SH can mask DHW-related faults. Conversely, during summer with minimal or no SH demand, SH component faults might go unnoticed.

During this phase of the analysis, a key characteristic that emerged was a few consistent patterns in the measurements (return temperature, volume flow, and energy) recorded during the fault period. In Figure 3, one can see the distribution of the different pattern types observed in each measured variable for each type of fault.

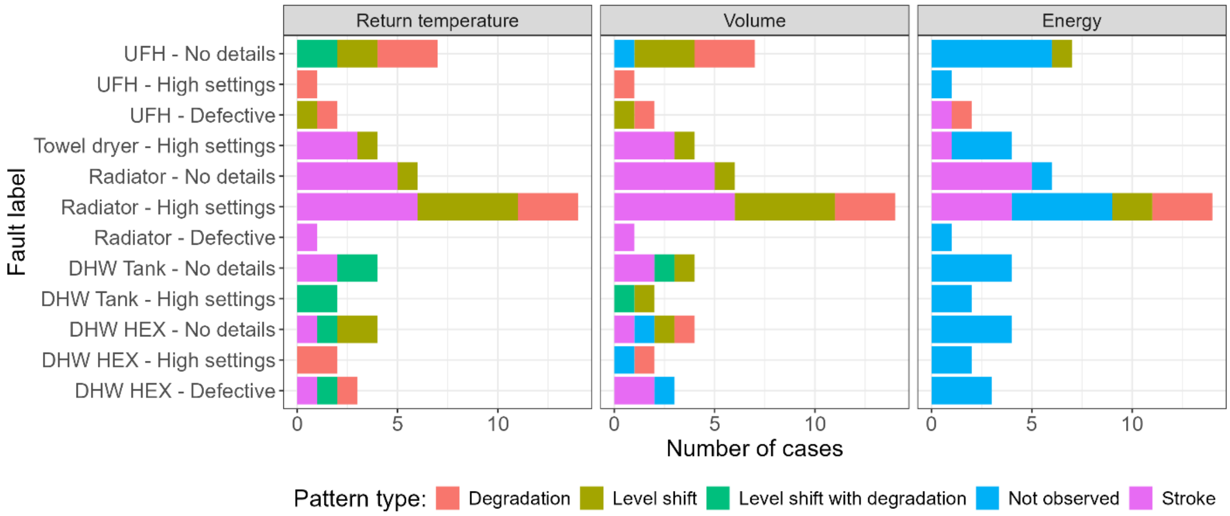


Figure 3: Horizontal bar chart of the distribution of the pattern types for the different fault labels and measurement variables within the heating system of the different monitored households. The y-axis lists various fault labels – while the x-axis tracks the number of occurrences for each fault label.

When analyzing the return temperature, it is evident that "Stroke," "Degradation," and "Level shift" patterns are prevalent, with "Level shift with degradation" being less common. Faults in radiators primarily show stroke and level shift patterns, whereas DHW production faults often exhibit patterns indicative of degradation (degradation and level shift with degradation). The volume flow variable displays patterns that mirror those found in return temperature measurements but includes a small number of "Not observed" instances across various fault labels. Energy measurements are distinct in that they do not present any cases of "Level shift with degradation," and a considerable amount of faults are categorized as "Not observed." Most of the detectable faults related to energy are associated with SH systems, particularly with radiator faults.

From Figure 3, one can see that radiator faults typically manifest stroke or level shift patterns – indicating an immediate fault response – across all three variables measured. Faults in towel dryers are similar to those in radiators, but no energy patterns are typically observed, possibly because towel dryers have a much lower energy output compared to radiators. This could also explain why some radiator faults exhibit no discernible pattern in energy consumption, likely because these are smaller units found in less critical areas like basements or bathrooms, where their faults have a negligible impact on energy output. Additionally, degradation and level shift with degradation are patterns more frequently observed in DHW and UFH systems. In these instances, there is typically no clear pattern in energy output, possibly due to the minimal impact of these smaller systems on overall energy usage. This is also compounded by the fact that these readings have daily resolution and in the summer months, when these types of faults occur more, the overall energy demand is reduced.

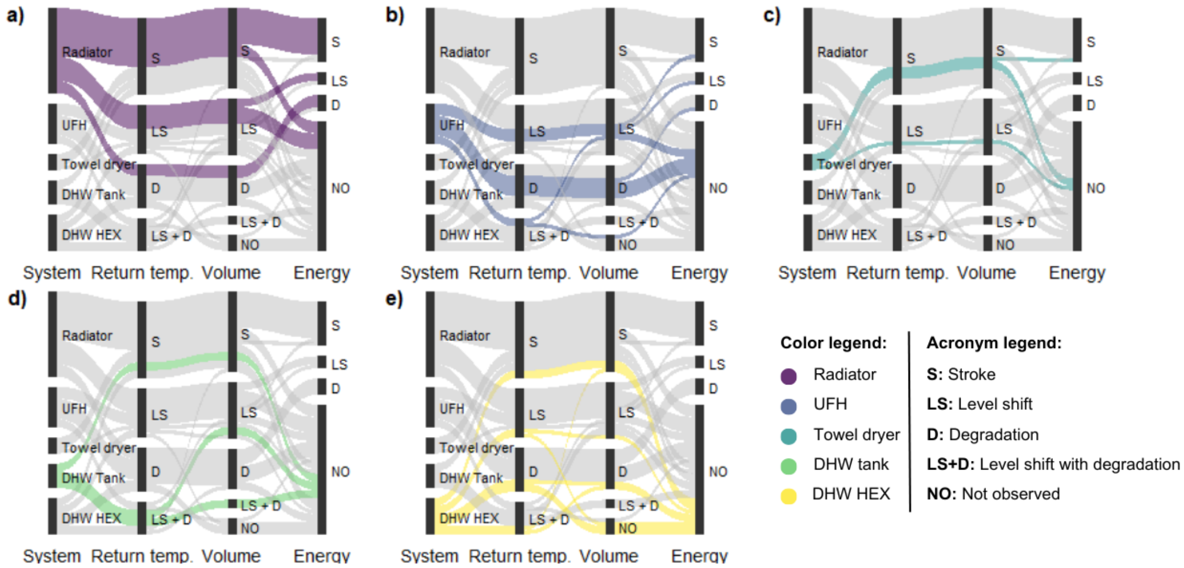


Figure 4: Sankey diagrams presenting the distribution of the patterns observed in the time series of the different measurement variables for the different systems where the fault occurred.

In Figure 4, one can see how these patterns are associated with each of the measured variables (return temperature, volume flow, and energy) across different systems in which the fault was encountered and identified by the technicians. This series of Sankey diagrams depicts some sort of fingerprint of typical patterns in the recordings of faulty subsystems like radiators, UFH, towel dryers, DHW tanks, and DHW HEX. The different patterns are categorized as follows: Strokes (S), Level shift (LS), Level shift with degradation (LD + D), Degradation (D), and Not observed (NO). The lines in the Sankey diagrams have thicknesses proportional to the frequency of observed patterns between measurement variables, highlighting where faults occurred in each system. In radiators, faults typically appear as strokes and level shifts, particularly affecting energy demand. Towel dryers show similar patterns to radiators, like strokes and level shifts in return temperature and volume flow, but without impacting energy usage. Unlike radiators, most other systems do not show significant patterns in energy usage, possibly due to their lower energy output. It can be noted that return temperature and volume flow often display consistent patterns across all systems. For instance, a stroke in return temperature is mirrored by volume flow. In UFH and DHW HEX, degradation is a common pattern, whereas the DHW tank system frequently shows a level shift combined with degradation.

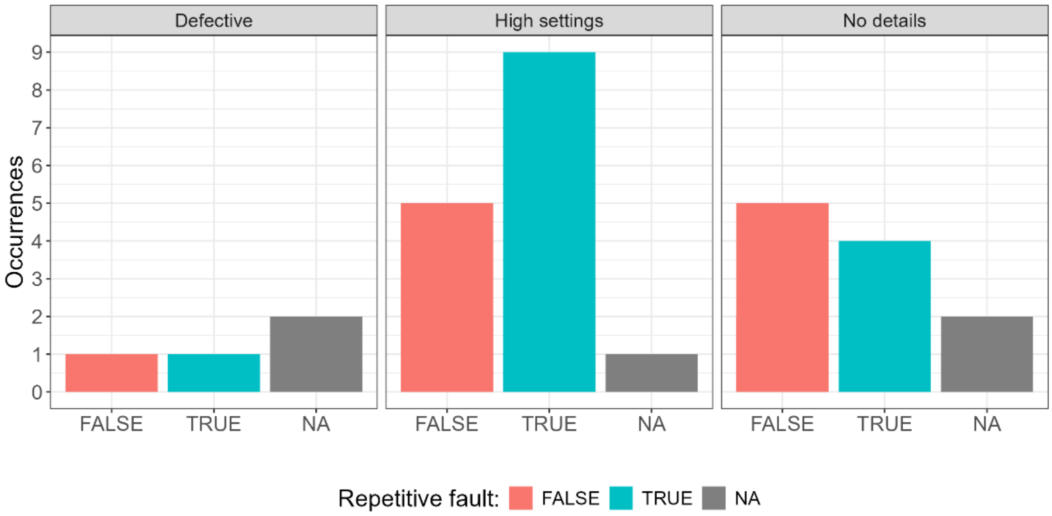


Figure 5: Visualization of whether or not a fault pattern is observed more than once in the data.

Another feature to highlight is the understanding of whether of not these patterns are repeated throughout the dataset, i.e., if the faulty patterns identified in our analysis have more than one occurrence. In Figure 5, one can see if there is more than one faulty pattern occurrence in the same building (TRUE) or only the one identified (FALSE). The few buildings with not enough data points are categorized as NA. This is usually due to the repetition of faults occurring during similar seasons. E.g., if the fault was identified during summer, it is possible that this fault might have occurred in the summer season of the previous year. This analysis was made by distinguishing faults originating from high settings and defective components. Unfortunately, as explained above, there are cases where it is not possible to know the origin of the fault: these were labeled as “No details”.

This chart can be used to understand patterns in the performance or issues related to a heating system's components, particularly focusing on whether certain conditions like being defective or having high settings are associated with repetitive patterns throughout the measurements. One can see that there are fewer instances (TRUE, FALSE, and NA) in the "Defective" category, suggesting fewer occurrences of component defects. However, the limited number of cases makes it difficult to conclude if faults caused by defective components are repetitive or not. The "High settings" category contains the majority of cases, primarily marked as TRUE, indicating a high frequency of repetitive faults when a component is set too high. This suggests a possible correlation between high settings and the occurrence of repetitive faults. This aligns with observations that these faults, often unnoticed by building occupants, are detected through data analysis by utilities. The "No details" category shows a similar frequency of TRUE and FALSE labels, indicating an equal distribution of cases with and without repeated faults. This category likely includes faults related to both defective components and high settings.

In theory, it is expected that high settings caused fault should be more repetitive throughout the dataset unless the occupants changed radically their heating settings due to tenants moving out or more impactful energy-saving measures. While the faults caused by a broken component are expected to be much less repetitive because it is the type of failure that occurs once and their impact is much visible to the occupants who will take measures to solve it.

the difference between the measurements during the start and end of the fault was determined. It was also assessed, how much these values increased or decreased per day, by measuring these differences and dividing them by the number of days of the segmented period of the identified faults. In Figure 6 and Figure 7, are presented two sets of plots displaying data points for return temperature, volume flow, and energy against their difference in values, and difference per number of days for the given fault labels, respectively.

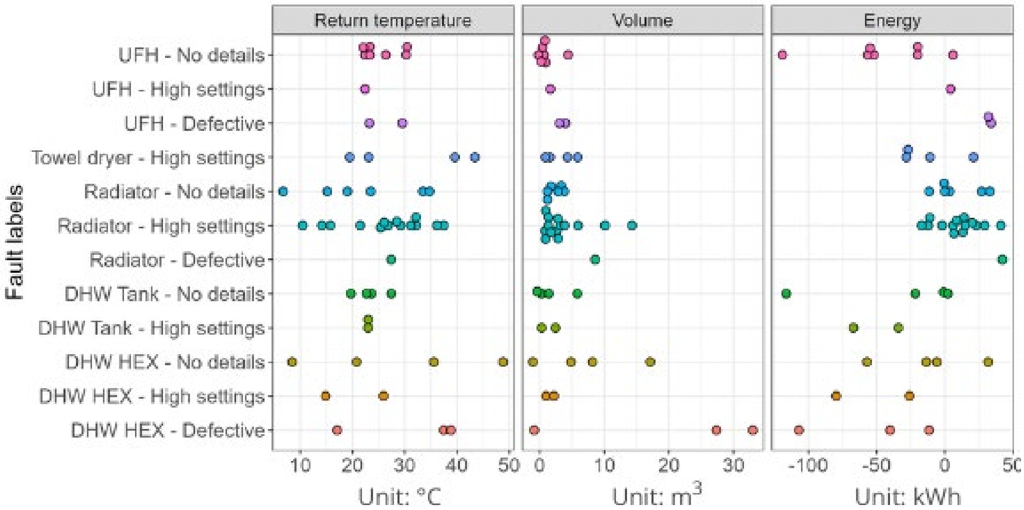


Figure 6: Difference of the measurements at the start and end period of the identified fault.

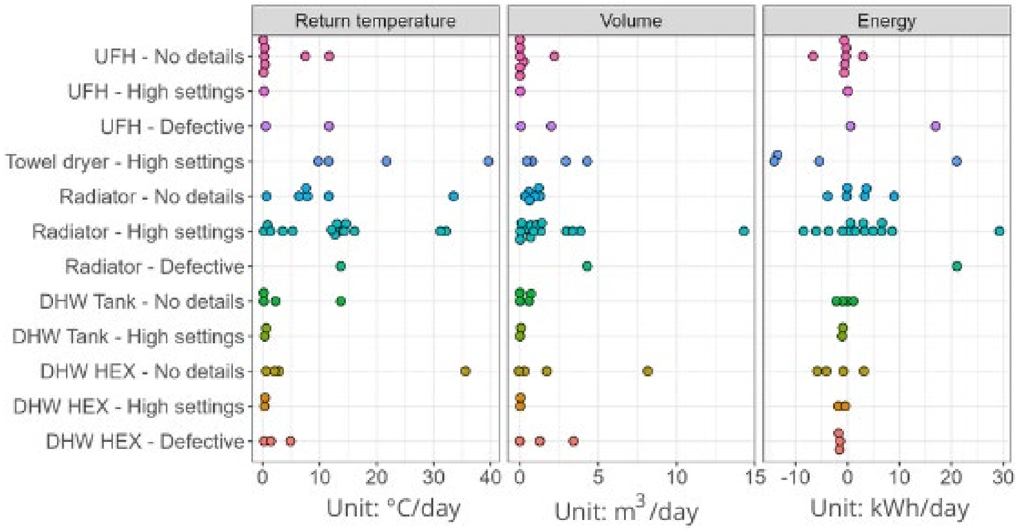


Figure 7: Difference of the measurements divided by the number of days of the segmented period.

Figure 6 presents a multifaceted view of the measurement changes during faulty operation periods. In Figure 6, the scatter of data points reflects the extent of measurement variations. The return temperature exhibits a notably higher variation than volume flow and energy measurements. Interestingly, energy readings include negative values, which predominantly occur during the transition from colder to warmer seasons, signifying a drop in the energy demand while the fault does not show a significant impact. When examining volume flow, the deviations related to the DHW production system are more pronounced than those for other components, hinting at a significant impact of faults in this system. In Figure 7, the normalization by the number of days results in a denser clustering of data points along the x-axis, indicating that the variations are less pronounced when accounting for time. However, SH systems, like radiators and towel dryers, present larger differences due to sharp increases over brief periods. Conversely, the UFH systems and DHW production, which exhibit more gradual degradation patterns over extended periods, display smaller normalized differences.

One last feature considered in this assessment is the volatility. In Figure 8, one can observe the volatility calculated for each measured parameter. The volatility is proposed as an indicator of the daily variation of the measurements during a degradation period (upward trend) or after an abrupt changepoint (variability after a level shift pattern). All stroke patterns were quantified as zero, due to the lack of data points after this sudden increase.

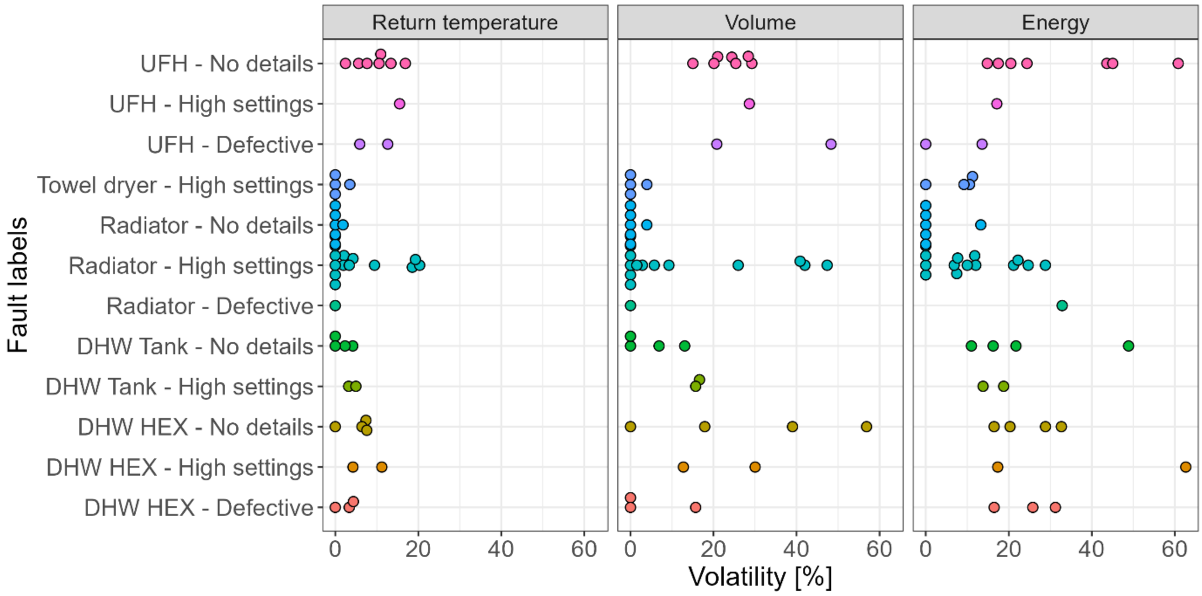


Figure 8: Volatility per fault label.

In Figure 8, one can see that for return temperature measurements, UFH systems and, to some extent, radiators demonstrate a larger volatility, which reflects the broader range of changes in the system's return temperature. DHW systems, in contrast, exhibit a smaller spread in these values, suggesting less variability in return temperature due to faults. When examining the volume flow and energy measurements, there is an apparent similarity in variability across all systems. Notable exceptions include the towel dryer and a few instances with radiators, where the observed volatility is lower. Additionally, the scale of calculated volatility differs significantly across the parameters. Return temperature presents a much smaller range of volatility when compared to volume flow and energy. This suggests that the impact of faults on the heating system is more noticeably reflected in variations in volume flow and energy usage rather than return temperature.

3.2. Automated detection of the fault patterns (time series segmentation)

As observed in Figure 3, the return temperature in all cases is the measured parameter where a fault always displays an anomalous pattern. Therefore, it is proposed that any FDD analysis for DH substations starts with a return temperature anomaly detection. From this, a time series segmentation must performed around the return temperature data that display an anomalous pattern. This segmented data is then used for further analysis regarding the other variables: volume flow and energy. This section presents the application of the BEAST method for detecting automatedly anomalous behavior in the time series by analyzing two different cases. The purpose is to exemplify the application of the method for enabling automated FDD.

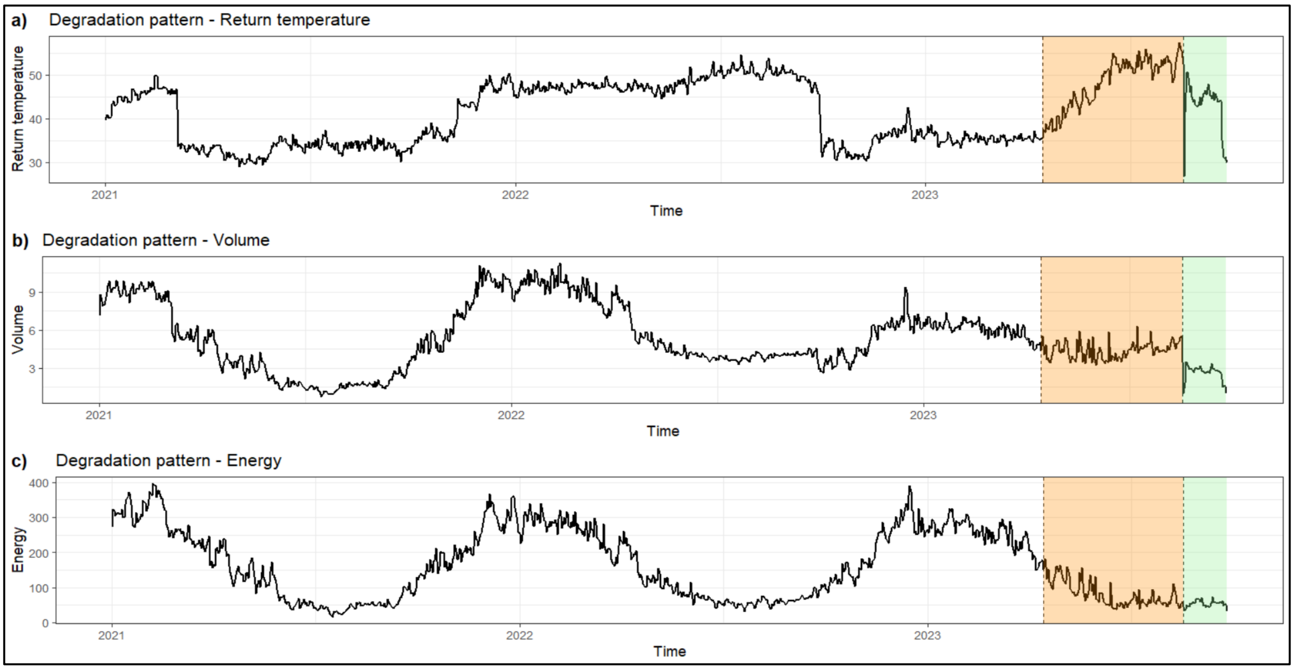


Figure 9: Segment with degradation pattern. a) Return temperature; b) Volume flow; c) Energy. The orange region identifies the fault segment analyzed in this study. The green region represents the period after the technician's visit.

The measurements in Figure 9 are from 2021-01-01 until 2023-09-26. The orange region displayed in the plot is the fault segment used in the analysis of this study, while the green region is the period following the technician's visit. The identified pattern is a degradation of the return temperature, while volume flow and energy do not show any singular pattern that might indicate a problem. However, it is observed that after the intervention, there is a large but short drop in the return temperature and water volume flow followed by a small longer-lasting reduction.

This specific fault report shows that this failure was due to a defective component in the heat exchanger of the DHW production. Specifically, the technician reported: "*A lack of cooling with high consumption. The domestic hot water is really warm. It is estimated to be a regulator valve (TPV) for controlling the hot water. The heat exchanger was closed and consumption fell significantly. The customer was advised to contact a plumber*". This confirms the large drop in return temperature and volume flow (by the technician closing fully the defective valve) and is followed by a small reduction of these variables (changing valve settings until the plumber replaces the component). From this building case, one can also observe a high return temperature during the majority of 2022 (winter season included), which most likely implies another failure in the system that was fixed before the technician's visit. In Table 5, one can see the list of obtained outputs from the BEAST methodology.

Table 5: BEAST output list.

Output variable	Description
Y	The original time series – Return temperature measurements.
Trend	The time series fitted trend component.
Pr(tcp)	The probability of a data point being an abrupt changepoint in the trend component.
SlpSign	Likelihood of the trend slope being upward (indicated by the red area), flat (represented in green), or downward (shown in blue).
Outlier	The detected outliers present in the time series.
Pr(ocp)	The probability of a data point being an outlier in the time series.
Error	The model residuals.

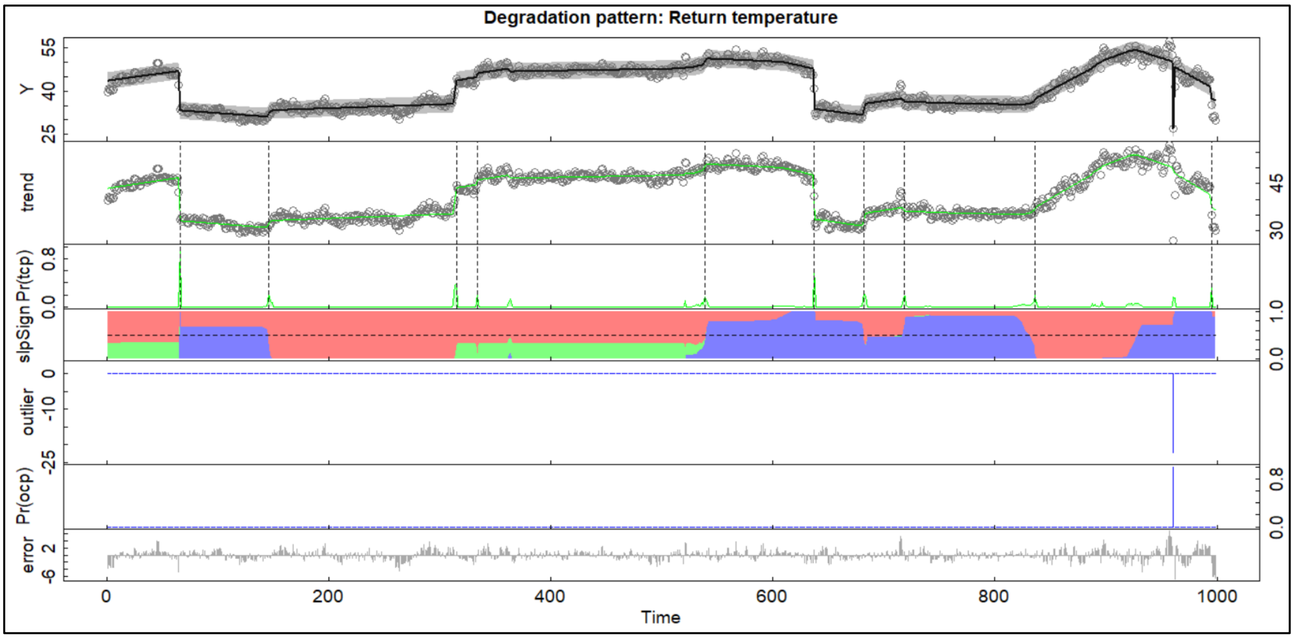


Figure 10: BEAST application – Degradation pattern detection.

In Figure 10, one can observe the BEAST algorithm output of the return temperature time series of the building case mentioned in Figure 9. Interpreting the plot in the context of return temperature and building heating system efficiency reveals that the applied methodology is accurate at identifying all the major change points within the temperature measurements. This capability extends to recognizing various types of anomalies, which is crucial for thorough monitoring. The analysis does not solely pinpoint abrupt shifts; it also encompasses the detection of gradual trend reversals, capturing instances where the trend shifts from rising to falling or stabilizes. Such nuanced detection is instrumental in identifying degradation patterns that emerge without sudden changes, an important aspect of preventive maintenance. Moreover, the outlier detection method employed here is refined enough to distinguish stroke patterns from level shifts or typical fluctuations. This refinement enhances the precision of the diagnostic tools and can be particularly effective in flagging irregular patterns that may indicate inefficiencies or faults within the heating system. Overall, these insights are invaluable for maintaining optimal operation, leading to potential improvements in the efficiency and reliability of building heating systems.

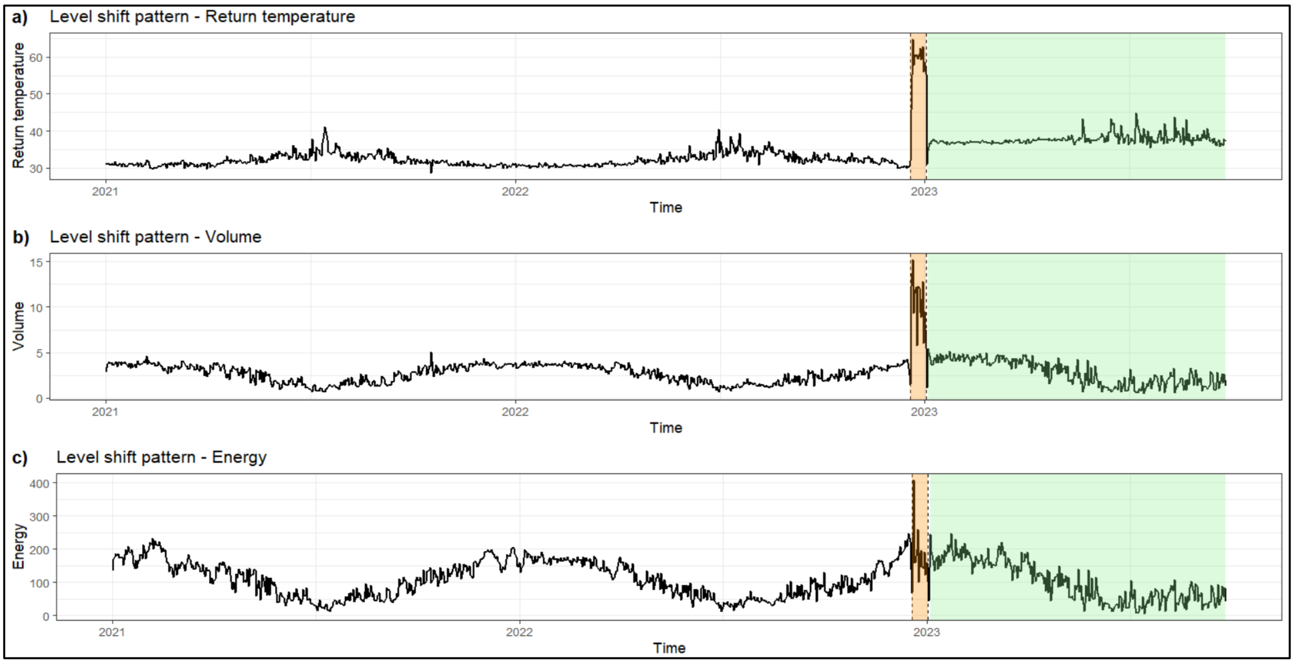


Figure 11: Segment with level shift pattern. a) Return temperature; b) Volume flow; c) Energy. The orange region identifies the fault segment analyzed in this study. The green region represents the period after the technician's visit.

Lastly, it is presented another building case in Figure 11 and Figure 12, however with a level shift pattern in the return temperature. The measurements in this substation are from 2021-01-01 until 2023-09-26 (similar to the one above). The identified pattern is clearly a level shift in the return temperature and volume flow, while the energy displays an initial stroke followed by lower values. After the intervention, there is a large reduction in the return temperature and water volume flow followed by stabilization of these measurements, whereas the energy does not display any significant change before and after the intervention, besides the aforementioned stroke. This specific fault report showed that this failure was due to a defective component in the UFH system. However, there were no other documented details regarding the nature of this fault.

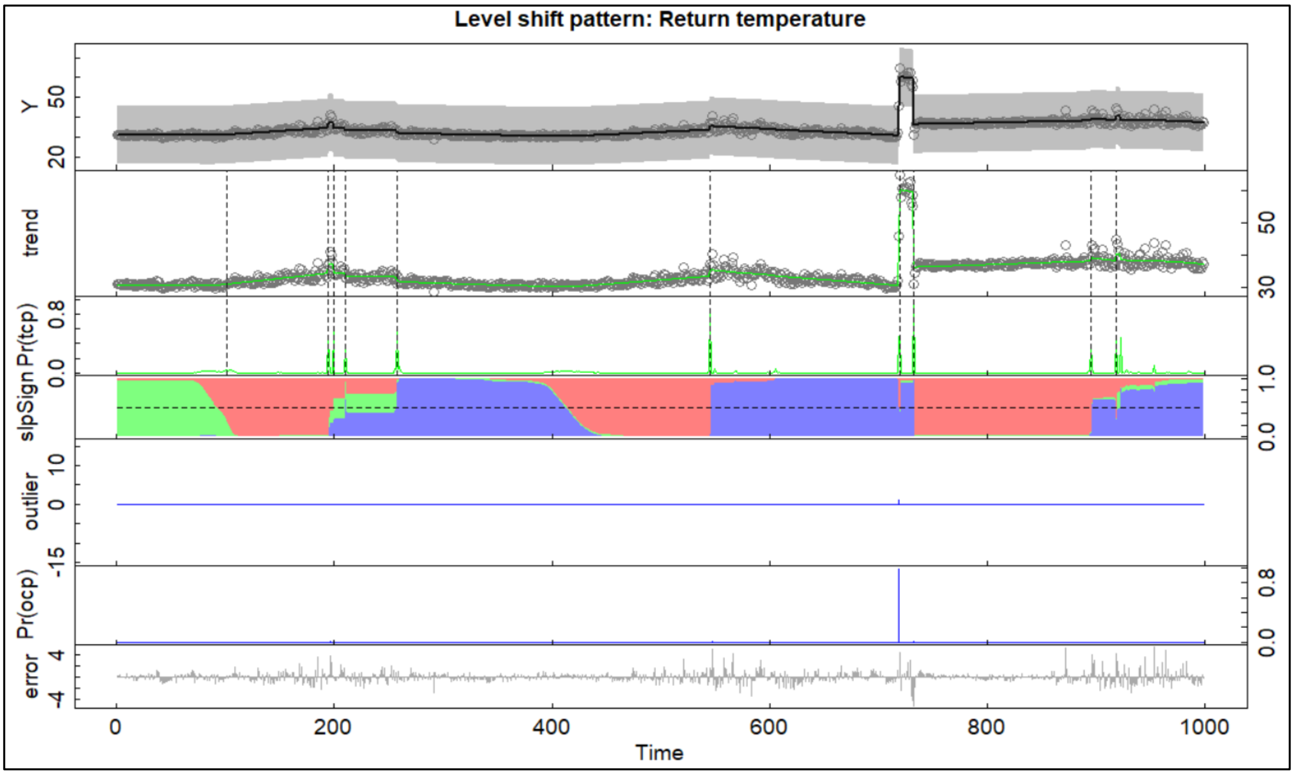


Figure 12: BEAST application – Level shift pattern detection.

In Figure 12, one can see the results of implementing the BEAST method into the return temperature of this case. Interpreting the plot about the return temperature provides valuable insights into the building's heating system efficiency. The method showcased that it is capable of precise detection and segmentation of anomalies, aligning closely with segments identified by expert assessment during the manual analysis. This includes the level shifts clearly demarcated within the dataset. The utility of the method extends beyond the identification of significant shifts to capturing smaller, abrupt changes, as evidenced around the 200-day mark on the timeline. Further scrutiny of the data post-intervention reveals that both the water volume flow and return temperature experience a decrease, thereafter stabilizing until the end of the observed period. This observation is in harmony with the findings in [50], which assert that well-functioning heating systems typically exhibit a constant and low return temperature throughout the year. This steadiness is an indicator of system efficiency and is a critical factor in the assessment of the heating system's performance. Through such analysis, the method not only confirms previous conclusions about system behavior but also offers a reliable means of monitoring and evaluating system efficiency over time.

3.3. Self-organizing maps and clustering

In the analysis, SOM was combined with the k-means clustering algorithm, as this integrated approach facilitates a more nuanced visualization and categorization of the complex patterns within the dataset. And the focus was on assessing pattern types, volatility levels, and fault periods, as these aspects seemed most relevant to identify the type of fault. However, it was not considered the differences in measurements before and after a fault, as these are heavily dependent on the specific systems present in individual households (this information was not available). Additionally, it was disregarded the feature of fault repetitiveness due to insufficient data for a reliable analysis of this characteristic.

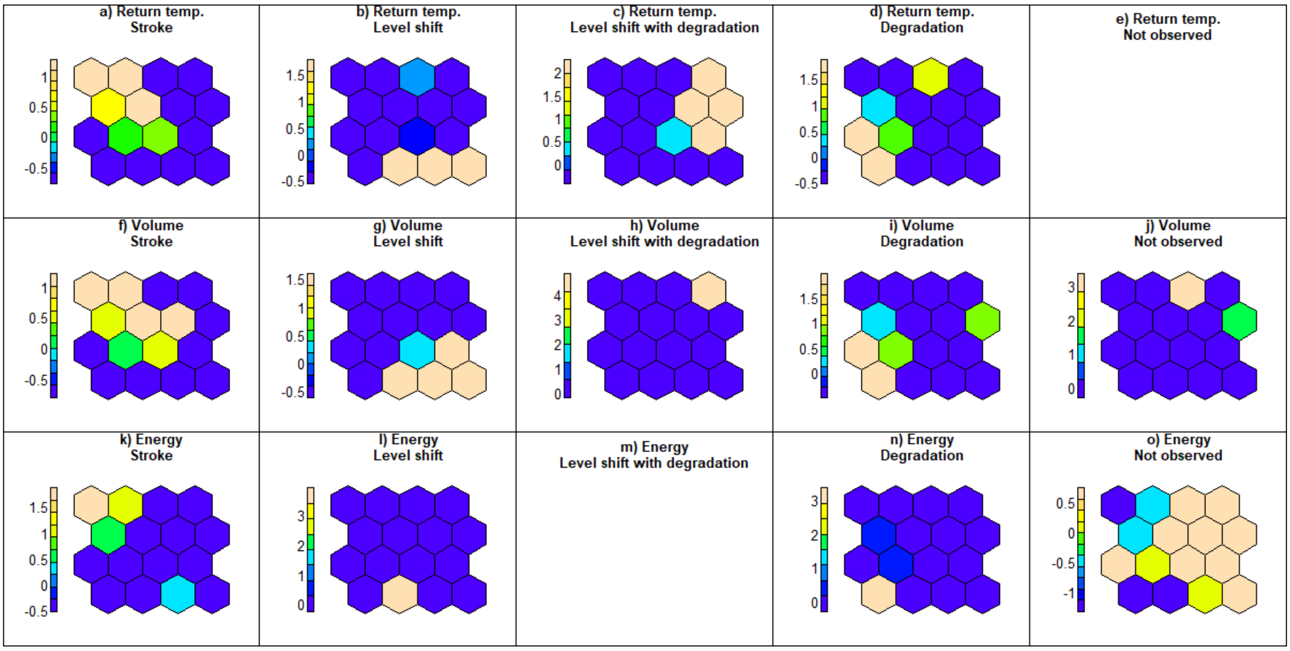


Figure 13: SOM property plots – Pattern features (one-hot encoded variables).

In Figure 13, one can see the SOM property plots that delineate the distribution of nodes according to the observed patterns in return temperature, volume flow, and energy. The analysis reveals that each plot is distinctively related to a specific measurement parameter and displays diverse patterns such as stroke, level shift, level shift with degradation, and degradation. The individual nodes of the SOM are structured with hexagonal cells, with a color-coding scheme that reflects the intensity of the measured variables according to the legend and its placement on the grid. Thus, data points with similar patterns are placed in nearby hexagons.

In terms of the observed conditions, the plots exhibit several grid layouts. The stroke condition, for instance, manifests itself with its cases being predominantly located towards the high left edge across all three variables, echoing a pattern of associated strokes identified in all variables as seen in earlier figures. The level shift pattern is primarily observed in the lower right corner. In contrast, the level shift with degradation pattern takes prominence in the top right area, with the return temperature showing a higher frequency of this pattern in comparison to volume flow; this particular pattern was not evident within the energy data. Conversely, degradation patterns are more apparent on the bottom left side of the map, with return temperature exhibiting a greater number of cases than the other variables. Lastly, the 'Not observed' designation signals the absence of recorded patterns, a scenario most prevalent in the energy data, whereas no instances were noted for return temperature.

Combined with the patterns' features, the SOM was developed using the calculated volatility of each measurement parameter and the season variable derived from the feature "fault month" (in Figure 2) where the fault occurred. The SOM of each of these features can be seen in Figure 14.

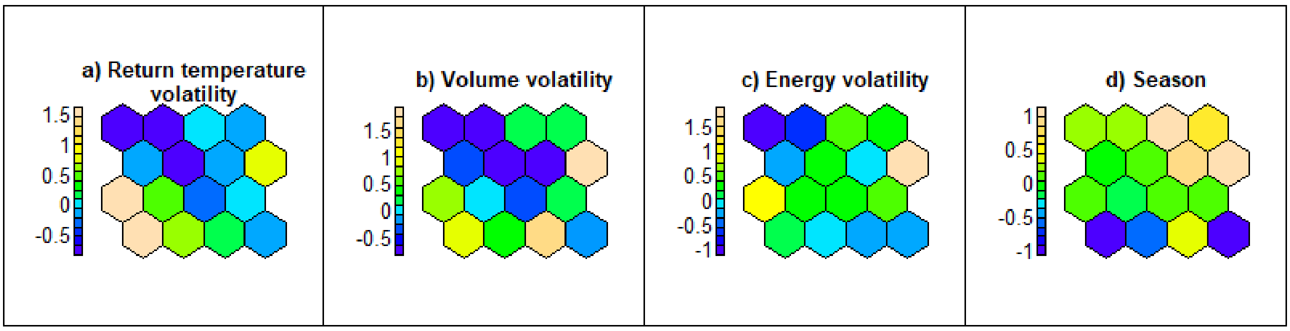


Figure 14: SOM property plot – Volatility and season features (numeric variables).

In Figure 14, one can conclude that data points indicating higher volatility, specifically those related to faults, tend to cluster differently per parameter of the map. In terms of return temperature and volume flow volatility, these values are similarly placed on the bottom left side of the map, with a few nodes also appearing in the top right corner. The spread of energy volatility across the map is more diffuse, with a notable concentration in the top right corner, corresponding to the "Not observed" pattern of Figure 13. The seasonal feature presents a less scattered pattern across the map. The colder months, marked in blue indicating lower values, are found in the lower level of the map. Conversely, the faults that occurred during warmer months are attributed to higher values of the scale, which are situated in the top right corner. Comparatively with Figure 13, the "Level shift" and "Degradation" patterns are associated with nodes representing colder months. On the other hand, the warmer months coincide with the "Level shift with degradation" pattern for return temperature and volume flow, and with the "Not observed" patterns for both volume flow and energy variables.

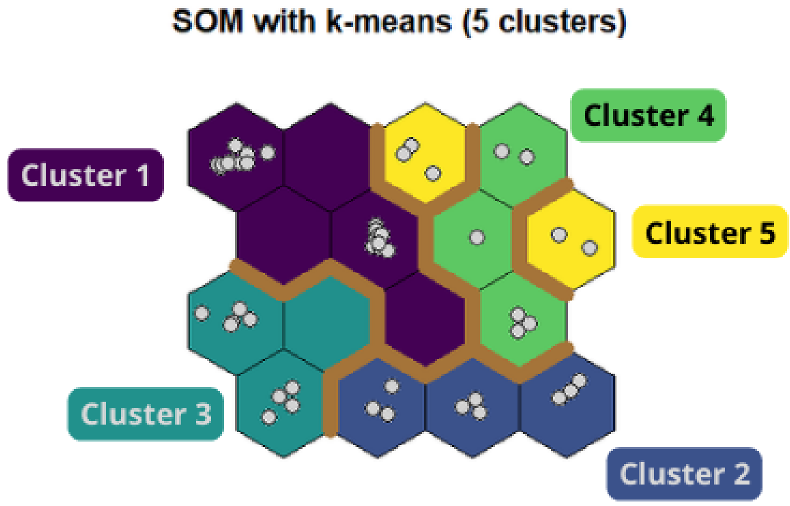


Figure 15: K-means cluster map.

The k-means clustering is then applied to the developed SOM with 5 clusters to combine the nodes with similar properties. In Figure 15, one can observe the cluster map where k-means was applied to SOM. The results from the clustering and the layout of each of its variables from Figure 13 and Figure 14 are summarized in Table 6.

Table 6: Summary of the k-means clusters implemented in SOM.

Cluster	Nr. of points	Nr. of nodes	Season	Return temperature		Volume flow		Energy	
				Pattern	Volatility	Pattern	Volatility	Pattern	Volatility
1	19	5	Mid	S	Low	S	Low	S NO	Low Medium
2	11	3	Mid Cold	LS	Medium Low	LS	High Low	LS NO	Low
3	9	3	Mid Cold	D	High Medium	D	Medium	NO D	High Medium
4	6	3	Warm Mid	LS + D	Low	LS + D S LS	Medium Low	NO	Medium
5	5	2	Warm	D LS + D	Low Medium	NO D	Medium High	NO	Medium High

To understand how the different features can lead to fault diagnostics, the different clusters are associated with ground truth obtained from the technician reports when inspecting faulty systems (see Figure 16).

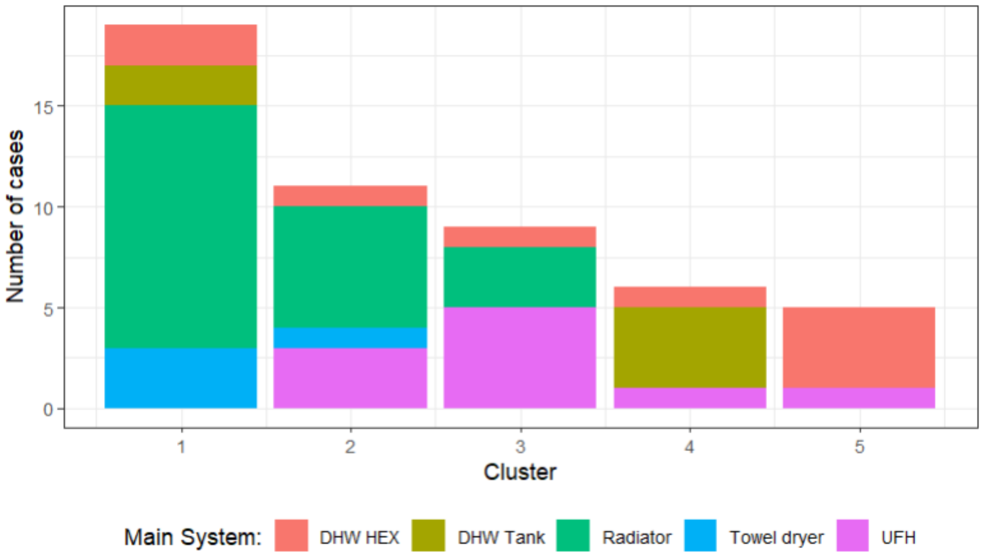


Figure 16: Association of the main system where the fault occurred with a given cluster.

In the study, five distinct clusters were identified based on fault occurrences in various heating systems, differentiated by seasonal timing, observed patterns, and volatility levels. In Figure 16, each main system was compared within the attributed cluster in order to observe if the different input variables could be used to diagnose the type of system where the fault occurred.

Cluster 1 – composes of 19 cases, primarily composed of SH system faults, with a noticeable prevalence of radiator and towel dryer issues. These faults predominantly occur during months with mid-range external temperatures. A consistent pattern observed across this cluster is the stroke pattern, with a few instances where the pattern was not observed in energy measurements. Volatility across all measured parameters is generally low, with occasional medium volatility noted in energy.

Cluster 2 – composes of 11 cases, also features faults in SH systems, notably radiators and UFH systems, with a significant number of radiator cases. The timing of these faults corresponds with colder to mid-external temperatures. The level shift is the consistent pattern in this cluster, with some energy patterns *not observed*. Volatility varies medium-low for return temperature, high-low for volume flow, and low for energy.

Cluster 3 – composes of 9 cases, once again shows faults primarily in SH systems, including radiators and UFH systems, with a significant number in the latter and occurring during the cold months. The degradation pattern is common here, with some cases of *not observed* patterns in energy measurements. Volatility is noted as high to medium across all parameters.

Cluster 4 – composes of 6 cases and is characterized by faults in the DHW system, particularly in storage tanks. These faults tend to happen in months with warmer to mid-external temperatures. This cluster exhibits predominantly the level shift with degradation pattern for return temperature and no observable pattern for energy, suggesting diverse fault characteristics. The levels of volatility are recorded, as medium and low for all three parameters.

Cluster 5 – composes of 5 cases, mostly comprises DHW system faults in the heat exchanger and occurs during warmer months. The return temperature often shows a level shift with degradation and degradation alone patterns, while no patterns are observed for volume flow and energy measurements. Each parameter exhibits varied volatility: low-medium for return temperature and medium-high for volume flow and energy.

From the clustering and the intervention reports, one can conclude the following:

- Radiator and towel dryer faults have similar patterns as stroke or level shift. Radiators, however, might display a similar pattern in the energy measurement while the towel dryers do not. This might be due to the maximum energy output being generally expected smaller for towel dryers.
- A fault in the UFH system may sometimes display a similar pattern as a radiator (abrupt change) while other times it might display a slower pattern (degradation). However, both these faults cause higher levels of volatility in the measurements.
- The faults occurring in the SH systems seem to show often in the mid-cold external temperatures season. While faults in the DHW production system usually occur in the summer.
- A fault in DHW production does not seem to cause any impact on the energy usage. Nevertheless, the return temperature and water volume flow are negatively affected by it. The results also show that DHW tank faults

might display the level shift with a degradation pattern, while a fault in the heat exchanger might cause a degradation pattern.

3.4. Lessons learned and suggestions for further work

This research underscores the necessity for implementing automated FDD methods within DH systems. As automated FDD methods are crucial for the advancement towards the 4th generation of DH systems, leveraging SHM data that is already available. Following such need, the article closes with some remarks regarding the lessons learned from this research but also proposes several suggestions for further work.

- Expert assessment and dataset limitations:

Due to the small dataset of 50 cases with some ambiguous fault labels, this study heavily relied on expert assessments, analyzing each case individually. This approach, while effective for the study, is impractical for real-world DH networks, which comprise numerous connected buildings. Consequently, the study proposes the use of the BEAST methodology to detect and segment data points exhibiting anomalous behavior, addressing the scalability issue.

- Effective fault measurement:

It was observed that return temperature is the most effective measurement for detecting faults. In all studied buildings, return temperature exhibited abnormal readings during the expected fault periods mentioned in the reports. Conversely, energy usage, despite being a more accessible metric, proved less reliable in detecting all faults. This is because it only identifies faults when the power output is substantially high, such as with radiators.

- Need for high-quality ground truth data:

Integrating and enhancing ML models for fault diagnosis in DH substations requires a larger quantity of high-quality ground truth data. The current study, constrained by a dataset of only 50 cases, was insufficient for training robust ML models. Traditional classification algorithms like Random Forest or XGBoost, as well as deep learning models, demand a substantial amount of well-labeled data for effective performance.

- Data compilation and sharing:

DH companies should compile, anonymize, and share data to expand the field and develop standardized datasets. This collaboration would facilitate the validation and comparison of different models using a uniform dataset, accelerating advancements in fault detection technologies.

- Handling SHM data issues:

A notable issue identified is the handling of hourly SHM data in Danish utilities, where rounding errors from truncation (rounding down measurements to the nearest integer) processes lead to asynchronous measurements. Although this was mitigated by aggregating data into daily granularity, it is imperative to eliminate these truncation processes to ensure data accuracy. As it is also envisioned that certain faults might only be detected with high-resolution measurements.

- Diverse data types for improved diagnostics:

Incorporating a broader range of data types can significantly enhance diagnostic capabilities. Indoor temperature data and radiator heat allocators, for example, could identify faults in specific rooms, while water meter data could highlight usage patterns and potential faults in DHW systems. Despite being obtainable with more sensors, these data are not consistently accessible to all customers. Additionally, pressure readings, though underutilized by DH companies, could aid in diagnosing faulty components by analyzing pressure differences.

- Utilizing public household information:

Integrating household area data, such as from the Danish Building and Residence Register (BBR) in Denmark, could help standardize SHM data and enable comparisons of faults across different buildings. This holistic data collection and analysis approach would significantly improve the accuracy and efficiency of fault diagnosis in DH substations.

4. Conclusion

The study initiates with the intricacies of gathering ground-truth data for the enhancement of fault diagnosis models. A significant step forward has been made with the second iteration of reports from Aalborg DH company, moving past the ambiguous nature of initial collections. However, the investigation reveals that issues persist, such as the mention of a single fault in instances of multiple concurrent faults, the premature timing of interventions obscuring full fault pattern development, and a bias towards faults unnoticed by residents.

Despite these hurdles, from an initial batch of 127 reports, 50 cases of household substation faults were analyzed. Due to the scarcity of comprehensive datasets in the literature and the mentioned challenges, the study concentrated on a detailed case-by-case examination. Analysis hinged on segmenting the fault period recorded in maintenance reports, identifying the month of occurrence, observing the fault patterns, their association with measured parameters, their repeatability, and the extent of parameter value changes due to faults. Findings indicate that while return temperature consistently displays a fault pattern and volume flow typically mirrors this pattern, energy usage data does not. Seasonal trends and volatility in fault occurrence were identified as meaningful for diagnosing fault nature, although data was insufficient to thoroughly assess the impact of fault repeatability and measurement variances before and after the fault.

Furthermore, the research advocates for the need for an automated time series decomposition method to segment SHM data to scale this analysis for all DH customers, as case-by-case analysis is unachievable. The segmentation is proposed using the return temperature because this parameter was always affected by the observed faults. The article refers to the method BEAST [45] as a suitable algorithm for further application, as the methodology proves capable of detecting extreme, short-duration, level shifts, and degradation anomalies.

To categorize symptoms and patterns of faults extracted by the segmentation, the study utilized the observed patterns, volatility, and seasonality to apply SOM with k-means clustering. This analysis produces clusters of faults with similar characteristics, revealing that different heating systems might exhibit similar features. The paper endorses SOM as a method not only to group symptoms for fault diagnosis but also to interpret high-dimensional data, as demonstrated in the study. By segmenting data indicative of operational faults and clustering into similar groups, the method solidifies its utility in advancing toward more effective DH-automated FDD processes. Thus, substantially improving the efficiency and reliability of the DH grid.

Credit author statement

Daniel Leiria: Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Data curation; Writing - original draft; Visualization. **Hicham Johra:** Conceptualization; Methodology; Writing - review & editing; Supervision. **Justus Anoruo:** Conceptualization; Data curation; Formal analysis; Writing - review & editing. **Imants Praulins:** Conceptualization; Data curation; Formal analysis; Writing - review & editing. **Marco Savino Piscitelli:** Conceptualization; Writing - review & editing; Supervision. **Alfonso Capozzoli:** Conceptualization; Writing - review & editing; Supervision. **Anna Marszal-Pomianowska:** Conceptualization; Resources; Writing - review & editing; Supervision. **Michal Zbigniew Pomianowski:** Conceptualization; Writing - review & editing; Supervision; Project administration; Funding acquisition.

All authors have read and agreed to the published version of the manuscript.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 958345 (PRELUDE).

Declaration of competing interest

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

Acknowledgments

The authors would like to acknowledge all partners involved in the E-DYCE and PRELUDE projects. Also, the first author would like to express his gratitude to his colleagues (Kamilla H. Andersen, Markus Schaffer, Martin Frandsen, and Simon P. Melgaard) in the Ph.D. “Corner Office” and everyone at BAEDA Lab in Politecnico di Torino for the valuable discussions and support.

References

[1] J. Volt et al., “District heat and cold management in the European Union – Status report on technology development, trends, value chains and markets – 2022”, European Commission, Joint Research Centre, Publications Office of the European Union, 2022, https://data.europa.eu/doi/10.2760/168004

[2] B. V. Mathiesen et al., “Towards a decarbonised heating and cooling sector in Europe: Unlocking the potential of energy efficiency and district energy,” Aalborg University, 2019, Accessed: July. 04, 2024. [Online]. Available: https://vbn.aau.dk/en/publications/towards-a-decarbonised-heating-and-cooling-sector-in-europe-unloc

[3] A. M. Jodeiri et al., “Role of sustainable heat sources in transition towards fourth generation district heating – A review”, *Renewable and Sustainable Energy Reviews*, Volume 158, 2022, 112156, doi: https://doi.org/10.1016/j.rser.2022.112156.

[4] H. Lund et al., “4th Generation District Heating (4GDH): Integrating smart thermal grids into future sustainable energy systems”, *Energy*, Volume 68, 2014, Pages 1-11, doi: https://doi.org/10.1016/j.energy.2014.02.089.

[5] H. Gadd, S. Werner, “Achieving low return temperatures from district heating substations”, *Applied Energy*, Volume 136, 2014, Pages 59-67, doi: https://doi.org/10.1016/j.apenergy.2014.09.022.

[6] D. S. Østergaard, S. Svendsen, “Costs and benefits of preparing existing Danish buildings for low-temperature district heating”, *Energy*, Volume 176, 2019, Pages 718-727, doi: https://doi.org/10.1016/j.energy.2019.03.186.

[7] K. Lygnerud et al., “A study on how efficient measures for secondary district heating system performance can be encouraged by motivational tariffs,” *Energy, Sustainability and Society*, vol. 13, no. 1, 2023, doi: https://doi.org/10.1186/s13705-023-00417-0.

[8] S. Månsson et al., "Faults in district heating customer installations and ways to approach them: Experiences from Swedish utilities," *Energy*, vol. 180, pp. 163–174, 2019, doi: https://doi.org/10.1016/j.energy.2019.04.220.
[9] K. Lygnerud, "Challenges for business change in district heating," *Energy, Sustainability and Society*, vol. 8, no. 1, 2018, doi: https://doi.org/10.1186/s13705-018-0161-4.
[10] D. Schmidt, "Digitalization of district heating and cooling systems," *Energy Reports*, vol. 7, pp. 458–464, 2021, doi: https://doi.org/10.1016/j.egyr.2021.08.082.
[11] D. Schmidt, Guidebook for the Digitalisation of District Heating: Transforming Heat Networks for a Sustainable Future, Final Report of DHC Annex TS4, 2023.
[12] A. Nissen, H. R. Shaker, and B. N. Jørgensen, "Automated and real-time anomaly indexing for district heating maintenance decision support system," *Applied Thermal Engineering*, vol. 233, pp. 120964–120964, 2023, doi: https://doi.org/10.1016/j.applthermaleng.2023.120964.
[13] M. Kozlovska et al., "Enhancing Energy Efficiency and Building Performance through BEMS-BIM Integration," *Energies*, vol. 16, no. 17, p. 6327, 2023, doi: https://doi.org/10.3390/en16176327.
[14] M. Pozzi et al., "District heating network maintenance planning optimization," *Energy Reports*, vol. 7, pp. 184–192, 2021, doi: https://doi.org/10.1016/j.egyr.2021.08.156.
[15] N. Es-sakali et al., "Review of predictive maintenance algorithms applied to HVAC systems," *Energy Reports*, vol. 8, pp. 1003–1012, 2022, doi: https://doi.org/10.1016/j.egyr.2022.07.130.
[16] J. van Dreven et al., "Intelligent Approaches to Fault Detection and Diagnosis in District Heating: Current Trends, Challenges, and Opportunities," *Electronics*, vol. 12, no. 6, p. 1448, 2023, doi: https://doi.org/10.3390/electronics12061448.
[17] S. Buffa et al., "Advanced Control and Fault Detection Strategies for District Heating and Cooling Systems—A Review," *Applied Sciences*, vol. 11, no. 1, p. 455, 2021, doi: https://doi.org/10.3390/app11010455.
[18] H. Gadd and S. Werner, "Fault detection in district heating substations," *Applied Energy*, vol. 157, pp. 51–59, 2015, doi: https://doi.org/10.1016/j.apenergy.2015.07.061.
[19] Ece Calikus et al., "Ranking Abnormal Substations by Power Signature Dispersion," *Energy Procedia*, vol. 149, pp. 345–353, 2018, doi: https://doi.org/10.1016/j.egypro.2018.08.198.
[20] S. Farouq et al., "Large-scale monitoring of operationally diverse district heating substations: A reference-group based approach," *Engineering Applications of Artificial Intelligence*, vol. 90, p. 103492, 2020, doi: https://doi.org/10.1016/j.engappai.2020.103492.
[21] M. Pozzi, et al., "Digitalisation in District Heating and Cooling systems: A tangible perspective to upgrade performance," Euroheat & Power, 2023, Accessed: July. 04, 2024. [Online]. Available: https://www.euroheat.org/data-insights/reports/dhc-report-on-digitalisation
[22] S. Månsson et al., "A machine learning approach to fault detection in district heating substations," *Energy Procedia*, vol. 149, pp. 226–235, 2018, doi: https://doi.org/10.1016/j.egypro.2018.08.187.
[23] F. Theusch et al., "Fault Detection and Condition Monitoring in District Heating Using Smart Meter Data," *Proceedings of European Conference of the Prognostics and Health Management Society*, vol. 6, no. 1, pp. 11–11, 2021, doi: https://doi.org/10.36001/phme.2021.v6i1.2786.
[24] C. Johansson and F. Wernstedt, "N-dimensional fault detection and operational analysis with performance metrics," presented at the *The 13th International Symposium on District Heating and Cooling*, 2012.
[25] M. Vallee et al., "Generation and evaluation of a synthetic dataset to improve fault detection in district heating and cooling systems," *Energy*, vol. 283, pp. 128387–128387, 2023, doi: https://doi.org/10.1016/j.energy.2023.128387.
[26] E. Guelpa and V. Verda, "Automatic fouling detection in district heating substations: Methodology and tests," *Applied Energy*, vol. 258, p. 114059, 2020, doi: https://doi.org/10.1016/j.apenergy.2019.114059.
[27] G. Bode et al., "Real-world application of machine-learning-based fault detection trained with experimental data," *Energy*, vol. 198, p. 117323, 2020, doi: https://doi.org/10.1016/j.energy.2020.117323.
[28] E. Calikus et al., "A data-driven approach for discovering heat load patterns in district heating," *Applied Energy*, vol. 252, p. 113409, 2019, doi: https://doi.org/10.1016/j.apenergy.2019.113409.
[29] P. Xue et al., "Fault detection and operation optimization in district heating substations based on data mining techniques," *Applied Energy*, vol. 205, pp. 926–940, 2017, doi: https://doi.org/10.1016/j.apenergy.2017.08.035.
[30] C. Wang et al., "New methods for clustering district heating users based on consumption patterns," *Applied Energy*, vol. 251, p. 113373, 2019, doi: https://doi.org/10.1016/j.apenergy.2019.113373.
[31] Z. Ma et al., "Building energy performance assessment using volatility change based symbolic transformation and hierarchical clustering," *Energy and Buildings*, vol. 166, pp. 284–295, 2018, doi: https://doi.org/10.1016/j.enbuild.2018.02.015.
[32] A. Capozzoli, M. S. Piscitelli, and S. Brandi, "Mining typical load profiles in buildings to support energy management in the smart city context," *Energy Procedia*, vol. 134, pp. 865–874, 2017, doi: https://doi.org/10.1016/j.egypro.2017.09.545.

[33] W. Sun, D. Cheng, W. Peng, "Anomaly detection analysis for district heating apartments," *Journal of Applied Science and Engineering*, 21 (1), pp. 33-44, 2018, doi: https://doi.org/10.6180/jase.201803_21(1).0005.
[34] A. M. Tureczek et al., "Clustering district heat exchange stations using smart meter consumption data," *Energy and Buildings*, vol. 182, pp. 144–158, 2019, doi: https://doi.org/10.1016/j.enbuild.2018.10.009.
[35] S. Kiluk, "Algorithmic acquisition of diagnostic patterns in district heating billing system," *Applied Energy*, vol. 91, no. 1, pp. 146–155, 2012, doi: https://doi.org/10.1016/j.apenergy.2011.09.023.
[36] Y. Choi and S. Yoon, "Autoencoder-driven fault detection and diagnosis in building automation systems: Residual-based and latent space-based approaches," *Building and Environment*, vol. 203, p. 108066, 2021, doi: https://doi.org/10.1016/j.buildenv.2021.108066.
[37] R. Kim et al., "System-level fouling detection of district heating substations using virtual-sensor-assisted building automation system," *Energy*, vol. 227, p. 120515, 2021, doi: https://doi.org/10.1016/j.energy.2021.120515.
[38] Y. Wang, C. Yang and W. Shen, "A Deep Learning Approach for Heating and Cooling Equipment Monitoring," *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, Vancouver, BC, Canada, 2019, pp. 228-234, doi:10.1109/COASE.2019.8843058.
[39] F. Zhang and H. Fleyeh, "Anomaly Detection of Heat Energy Usage in District Heating Substations Using LSTM based Variational Autoencoder Combined with Physical Model," *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, Kristiansand, Norway, 2020, pp. 153-158, doi:10.1109/ICIEA48937.2020.9248108.
[40] M. Li et al., "A Data-Driven Method for Fault Detection and Isolation of the Integrated Energy-Based District Heating System," in *IEEE Access*, vol. 8, pp. 23787-23801, 2020, doi:10.1109/ACCESS.2020.2970273.
[41] S. Månsson et al., "A taxonomy for labeling deviations in district heating customer data," *Smart Energy*, vol. 2, p. 100020, 2021, doi: https://doi.org/10.1016/j.segy.2021.100020.
[42] D. Leiria et al., "Towards automated fault detection and diagnosis in district heating customers: generation and analysis of a labeled dataset with ground truth," In *Proceedings of Building Simulation 2023: 18th Conference of International Building Performance Simulation Association*, Shanghai, China, 2023, Vol. 18, pp. 3620-3628, doi: https://doi.org/10.26868/25222708.2023.1576
[43] S. Månsson et al., "A Fault Handling Process for Faults in District Heating Customer Installations," *Energies*, vol. 14, no. 11, pp. 3169–3169, 2021, doi: https://doi.org/10.3390/en14113169.
[44] J. van Dreven et al., "A Data Generation Approach for Intelligent Fault Detection and Diagnosis in District Heating," *The 35th Swedish Artificial Intelligence Society (SAIS'23) annual workshop*, Karlskrona, Sweden, 2023.
[45] K. Zhao et al., "Detecting change-point, trend, and seasonality in satellite time series data to track abrupt changes and nonlinear dynamics: A Bayesian ensemble algorithm," *Remote Sensing of Environment*, vol. 232, 2019, doi: https://doi.org/10.1016/j.rse.2019.04.034.
[46] M. Schaffer, "Increasing the accuracy of low-resolution commercial smart heat meter data and analysing its error," In *EC3 Conference 2023 – European Council on Computing in Construction*, Crete, Greece, 2023.
[47] Teuvo Kohonen, *Self-Organizing Maps*. Springer Science & Business Media, 2012.
[48] R. Wehrens and L. M. C. Buydens, "Self- and Super-organizing Maps in R: The kohonen Package," *Journal of Statistical Software*, vol. 21, no. 5, 2007, doi: https://doi.org/10.18637/jss.v021.i05.
[49] A. M. Ikotun et al., "K-means Clustering Algorithms: a Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data," *Information Sciences*, vol. 622, no. 622, 2022, doi: https://doi.org/10.1016/j.ins.2022.11.139.
[50] D. S. Østergaard, M. Tunzi, and S. Svendsen, "What does a well-functioning heating system look like? Investigation of ten Danish buildings that utilize district heating efficiently," *Energy*, vol. 227, p. 120250, 2021, doi: https://doi.org/10.1016/j.energy.2021.120250.

5.4 Further discussion

This research underscores the crucial role and multifaceted benefits of implementing automated FDD methods within the DH systems, specifically regarding the buildings connected to the grid. As we move towards the improvement of DH systems, leveraging SHM data becomes indispensable. This section delves into the challenges as well as the advantages of establishing a robust FDD process within DH grids.

Challenges:

- Data quality and availability:

One of the primary challenges is the availability and quality of data. This study, constrained by small datasets in both articles with some ambiguous fault labels, highlighted the difficulty in training robust machine learning models. Thus high-quality, well-labeled ground truth data is mandatory for effective fault diagnosis, necessitating substantial efforts in data collection and annotation.

- Scalability:

The reliance on expert assessment for fault detection, while effective in this study, is impossible for large-scale DH networks with numerous connected buildings. Automated, accurate, and well-trained methods, such for anomaly detection and classification algorithms, are required if any DH company wants to fully implement FDD in their operation.

- Handling SHM data issues:

SHM data often suffers from rounding errors and asynchronous measurements due to hourly data truncation. This research mitigated these issues by aggregating data to daily granularity but it must be emphasized the need for eliminating truncation processes to ensure data accuracy. High-resolution measurements are crucial for detecting certain types of faults.

Advantages:

- Enhanced efficiency and reliability:

Implementing automated FDD processes enhances the overall efficiency and reliability of DH systems. By accurately identifying and diagnosing faults, maintenance can be performed proactively, reducing downtime and improving service continuity for connected buildings.

- Data standardization and sharing:

Encouraging DH companies to compile, anonymize, and share data can foster collaboration and innovation. Standardized datasets allow for the validation and comparison of different fault detection models, accelerating technological advancements towards fault diagnosis.

- Cost savings:

Effective FDD can lead to significant cost savings by preventing extensive damage, optimizing maintenance schedules, and extending the lifespan of DH infrastructure. Early detection of faults reduces the need for emergency repairs and lowers operational costs.

- Sustainability:

Efficient DH systems contribute to environmental sustainability by optimizing energy use and reducing waste. Accurate fault detection helps maintain optimal performance, supporting the transition towards greener, more sustainable heating solutions within the DH sector.

Overall, the implementation of automated FDD methods in DH systems still presents several challenges, nevertheless, its advantages are substantial. High-quality data, scalable methods, and the integration of diverse data types are critical for success. The benefits, including improved diagnostic accuracy, enhanced system efficiency, cost savings, and sustainability, underscore the importance of advancing FDD processes within DH grids.

Chapter 6. Conclusions

This dissertation has explored the innovative use of smart heat meter (SHM) data within district heating (DH) systems, highlighting significant advancements across various aspects of energy efficiency and fault detection and diagnosis (FDD). The research is structured around several key chapters, each contributing unique insights and methodologies to enhance the efficiency and sustainability of buildings connected to the DH systems. Each Chapter from 2 to 5 attempted to answer the research questions stated in Chapter 1, section 1.2.

6.1 Key insights: Addressing the research questions

> *RQ1. How can weather data, energy performance certificates (EPC), and visualization tools be used to aid in decision-making for energy management within district heating?*

Chapter 2 validated the applicability of SHM in assessing the connections between buildings and the district heating (DH) grid. This foundational work confirmed that SHM is not only effective in monitoring energy usage but also crucial in understanding the performance of the buildings connected to the grid. In this work, SHM data was combined with the information present in buildings' energy performance certificates (EPC) and weather data.

By integrating weather data, EPC information, and a visualization tool (developed in Shiny) with SHM data, energy management within DH systems can be significantly enhanced. Weather data provides a critical context for interpreting energy usage patterns, as external temperatures and other weather conditions directly influence heating demand. By correlating SHM data with weather information, DH operators can better comprehend and predict fluctuations in energy usage, enabling more efficient management of heating supply. The EPC, on the other hand, offers valuable insights into the energy efficiency and characteristics of buildings connected to the DH network. EPC contains information about the thermal performance of buildings, including insulation quality, heating system efficiency, and overall estimated energy usage ratings. By having access to the EPC data, DH operators and building owners can identify potential buildings with poor energy performance that may require targeted interventions or upgrades to improve their efficiency.

This allows for a more tailored approach to energy management, focusing efforts on buildings that will benefit most from efficiency improvements.

Visualization tools play a crucial role in making complex data more accessible and actionable. By visualizing SHM data alongside weather conditions and EPC information, DH operators can gain a clearer, more comprehensive view of the factors influencing energy usage. These tools can highlight trends, identify anomalies, and facilitate easier interpretation of data, supporting more informed decision-making.

> *RQ2. Can machine learning (ML) models based on SHM data accurately predict the energy demand for space heating and domestic hot water?*

Chapter 3 developed and tested a new methodology for disaggregating energy demand for space heating (SH) and domestic hot water (DHW) using the SHM data from residential buildings. The results from this chapter indicate that ML models based on SHM data can indeed be used to predict the energy demand for both SH and DHW. By leveraging the weather data and starting from the assumption that high daily energy values are the combined shares of SH and DHW, the developed method can predict such values.

The ML models were trained using historical SHM data, which included energy usage patterns and were supplemented with weather measurements. The key strength of the methodology lies in its ability to distinguish between SH and DHW energy usage from the low-resolution data from the SHM.

Further in this topic, it was observed that truncated measurements are a current procedure in the data collection strategy done by the DH utility companies. Therefore, it was investigated the impact of this rounding down process in the disaggregation method and to mitigate this issue.

> *RQ3. Can smart heat meters data, supplemented with information about indoor environmental conditions, enhance our understanding of the energy use and heating practices of building occupants?*

The role of occupant behavior in shaping energy usage patterns is a complex variable that is often difficult to quantify without detailed data or prior knowledge of their routines. As variations in individual heating practices within different rooms can significantly influence the overall energy usage of a building. Chapter 4 underscores that while SHM data alone provide valuable insights, they are insufficient to fully capture the nuances of occupant behavior.

To address this, detailed indoor condition data, such as temperature, CO2 levels, humidity, and room-specific heat measurements, were incorporated. These data sources, combined with interviews with tenants, provided a more comprehensive understanding of how occupant behavior affects energy usage. This multifaceted approach revealed that occupant behavior can lead to significant variations in energy consumption patterns, even in buildings with similar thermal properties.

Moreover, the investigation into adjusting the current ES model from its linear nature to a sigmoid function further highlighted the potential for more accurate modeling of energy usage patterns. This adjustment improved the explainability of the relationship between building heating usage and outdoor temperature, particularly in colder conditions, paving the way for more targeted and effective energy management strategies.

RQ4. How effective are smart heat meters at identifying specific types of faults within the district heating system customers?

Chapter 5 advanced the field of automated FDD in DH systems. This chapter showcased the importance of refined data analysis and innovative algorithms, proving that enhanced FDD can lead to substantial improvements in operational efficiency and fault management. The research highlighted that SHM play a crucial role in this advancement by continuously collecting data, which is essential for detecting specific types of faults within the DH system infrastructure.

The SHM can monitor key parameters such as flow rates, temperatures, and energy usage in real-time. This continuous data stream allows for the application of sophisticated analytical methods to identify anomalies that may indicate faults. For instance, deviations in return temperatures, which in the research presented as the best fault presence indicator, can signal issues like improper or blockages in the heating systems.

However, it is important to emphasize the need and importance of ground truth data about faults. Ground truth data provides a reliable baseline that is crucial for training and validating the fault detection algorithms. Without accurate ground truth data, it is challenging to assess the effectiveness of the fault detection system and to improve its precision over time.

From these data patterns measured by the SHM, a future proactive approach will help mitigate the fault impact before it escalates, thus maintaining DH operational efficiency. By leveraging high-dimensional data clustering and self-organizing maps, SHM can accurately categorize fault symptoms and patterns. This precision in fault diagnosis ensures timely and appropriate interventions even before the technician's intervention. And in some cases, the intervention will be done by a mere phone call from the technician to the homeowner.

6.2. Implications for district heating management

The findings of this dissertation have several important implications for the management buildings connected within the DH systems. By leveraging their SHM, DH operators can achieve deeper knowledge of their customers, better operational efficiency, enhance energy sustainability, and indirectly improve customer satisfaction. The enhanced data-driven insights facilitate more responsive and proactive management of DH systems, allowing for predictive maintenance and quicker resolution of issues. This proactive approach not only reduces downtime and maintenance costs but also optimizes energy usage, leading in the future to greater environmental benefits.

For DH utility companies, it is recommended to implement and investigate further the findings of this research to optimize energy usage and reduce their carbon impact. This includes adopting the new methodologies for energy demand disaggregation, integrating SHM with additional sensor data (if available), and enhancing FDD capabilities through refined algorithms and better data collection.

The implementation of SHM has introduced new roles, responsibilities, and potential challenges for DH systems. The ability of SHM to continuously monitor and analyze data provides DH operators with a more comprehensive view of the entire system, including the buildings. This integration enhances the ability to manage energy use more effectively and address issues promptly. Reflections on the new role of the DH sector with SHM data reveal increased accountability and the need for more sophisticated operational strategies. The potential challenges include managing large volumes of data, and ensuring data accuracy while maintaining system security. However, the benefits of adopting SHM, such as enhanced fault detection and improved energy management, outweigh these challenges, presenting a compelling case for their widespread implementation.

6.3 Directions for future research

Despite the significant advancements presented, this research encountered several limitations. Data availability was a primary constraint, affecting the generalizability of the findings. Additionally, the constraint of current SHM being limited to one-hour granularity of truncated measurements was a large barrier. These limitations highlight the need for continued investment in more advanced and broader data collection efforts.

The study has identified several gaps that warrant further investigation, as follows:

- Future research should focus on a deeper analysis of energy usage patterns across different building typologies and demographic segments to understand the variability in energy usage.
- Further refinement of fault detection algorithms is necessary to enhance their sensitivity, ensuring more reliable fault diagnosis through better data collection and preprocessing.
- Finally, continued investigation into the impact of occupant behavior on energy usage is essential. Developing effective user feedback systems could drive behavioral changes that contribute to energy savings and enhanced system performance.

In conclusion, this dissertation has laid the groundwork for significant improvements in the DH system management of its connected buildings through the use of SHM data. By addressing the limitations and pursuing the suggested directions for future research, the potential for even greater advancements in energy efficiency and sustainability in DH systems can be realized.

References

[1] L. Rohde, T. S. Larsen, R. L. Jensen, and O. K. Larsen, “Framing holistic indoor environment: Definitions of comfort, health and well-being,” Indoor and Built Environment, vol. 29, no. 8, Sep. 2019, doi: https://doi.org/10.1177/1420326x19875795.
[2] EU Building Stock Observatory, “EU Building Stock Observatory – Factsheets” [Online], Available: https://building-stock-observatory.energy.ec.europa.eu/factsheets/ [Accessed: Jul. 5, 2024].
[3] European Commission, “Energy performance of buildings directive”, 2018 [Online], Available: https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en [Accessed: Jul. 5, 2024].
[4] European Commission, “Heating and cooling”, 2022 [Online], Available: https://energy.ec.europa.eu/topics/energy-efficiency/heating-and-cooling_en [Accessed: Jul. 5, 2024].
[5] IEA, “Heating”, 2023 [Online], Available: https://www.iea.org/energy-system/buildings/heating [Accessed: Jul. 5, 2024].
[6] Fetting, C. (2020). “The European Green Deal”, ESDN Report, December 2020, ESDN, Office, Vienna.
[7] Directive (EU) 2018/2001, “Directive (EU) 2018/2001 of the European Parliament and of the Council of 11 December 2018 on the promotion of the use of energy from renewable sources (recast),” European Parliament and Council, Dec. 11, 2018, OJ L328/82.
[8] Directive 2010/31/EU, “Directive 2010/31/EU of the European Parliament and of the Council of 19 May 2010 on the energy performance of buildings”, European Parliament and Council, May 19, 2010.
[9] European Commission, “Proposal for a Directive of the European Parliament and of the Council on the energy performance of buildings (recast),” (2021) 802 final, Dec. 15, 2021 [Online], Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0802&qid=1641802763889 [Accessed: Jul. 5, 2024].

[10] L. Minh Dang *et al.*, "Fifth generation district heating and cooling: A comprehensive survey," *Energy Reports*, vol. 11, pp. 1723–1741, Jun. 2024, doi: https://doi.org/10.1016/j.egyr.2024.01.037.
[11] Danish Energy Agency, "Technology Data Catalogue for Electricity and District Heating," Danish Energy Agency, Apr. 2022 [Online], Available: https://ens.dk/sites/ens.dk/files/Statistik/technology_data_catalogue_for_el_and_dh_-_0009.pdf. [Accessed: Jul. 5, 2024].
[12] F. H. Ouerghi, M. Omri, Kottakkaran Sooppy Nisar, Rasha, and A. I. Taloba, "Investigating the potential of geothermal energy as a sustainable replacement for fossil fuels in commercial buildings," *Alexandria Engineering Journal /Alexandria Engineering Journal*, vol. 97, pp. 215–229, Jun. 2024, doi: https://doi.org/10.1016/j.aej.2024.03.094.
[13] L. Bertelsen, "THE GENERATION BATTLE – 4th versus 5th generation of District Heating – DBDH," Jan. 25, 2023 [Online], Available: https://dbdh.dk/the-generation-battle-4th-versus-5th-generation-of-district-heating/#:~:text=Centralized%20heat%20generation%20enables%20economy [Accessed: Jul. 5, 2024].
[14] Ministry of Climate, Energy and Utilities, "Denmark's Integrated National Energy and Climate Plan (NECP)," European Commission, 2019 [Online], Available: https://energy.ec.europa.eu/system/files/2020-01/dk_final_necp_main_en_0.pdf [Accessed: Jul. 5, 2024].
[15] Danish Energy Agency, "District Energy: A Climate-Friendly Energy Solution," Danish Energy Agency, 2020 [Online], Available: https://ens.dk/sites/ens.dk/files/contents/material/file/district-energy.pdf [Accessed: Jul. 5, 2024].
[16] K. Johansen and S. Werner, "Something is sustainable in the state of Denmark: A review of the Danish district heating sector," *Renewable and Sustainable Energy Reviews*, vol. 158, p. 112117, Apr. 2022, doi: https://doi.org/10.1016/j.rser.2022.112117.
[17] D. B. Olawade, O. Z. Wada, Aanuoluwapo Clement David-Olawade, Oluwaseun Fapohunda, A. O. Ige, and J. Ling, "Artificial intelligence potential for net zero sustainability: Current evidence and prospects," Next sustainability, vol. 4, pp. 100041–100041, Jan. 2024, doi: https://doi.org/10.1016/j.nxsust.2024.100041.

[18] M. Schaffer, T. Tvedebrink, and A. Marszal-Pomianowska, "Three years of hourly data from 3021 smart heat meters installed in Danish residential buildings," *Scientific Data*, vol. 9, no. 1, Jul. 2022, doi: https://doi.org/10.1038/s41597-022-01502-3.
[19] D. Leiria, H. Johra, A. Marszal-Pomianowska, M. Z. Pomianowski, and P. Kvols Heiselberg, "Using data from smart energy meters to gain knowledge about households connected to the district heating network: A Danish case," *Smart Energy*, vol. 3, p. 100035, Aug. 2021, doi: https://doi.org/10.1016/j.segy.2021.100035.
[20] D. Leiria, H. Johra, A. Marszal-Pomianowska, and M. Z. Pomianowski, "A methodology to estimate space heating and domestic hot water energy demand profile in residential buildings from low-resolution heat meter data," *Energy*, vol. 263, p. 125705, Jan. 2023, doi: https://doi.org/10.1016/j.energy.2022.125705.
[21] D. Leiria *et al.*, "Validation of a new method to estimate energy use for space heating and hot water production from low-resolution heat meter data," *E3S Web of Conferences*, vol. 362, pp. 10001–10001, Jan. 2022, doi: https://doi.org/10.1051/e3sconf/202236210001.
[22] D Leiria, M. Schaffer, H Johra, A Marzsal-Pomianowska, and M. Z. Pomianowski, "Estimating residential space heating and domestic hot water from truncated smart heat data," *Journal of Physics. Conference Series*, vol. 2600, no. 2, pp. 022017–022017, Nov. 2023, doi: https://doi.org/10.1088/1742-6596/2600/2/022017.
[23] D. Leiria *et al.*, "From showers to heaters: Evaluating the different factors that play a role in buildings' energy signature," Submitted in *Energy and Buildings*, 2024.
[24] D. Leiria *et al.*, "Towards automated fault detection and diagnosis in district heating customers: generation and analysis of a labeled dataset with ground truth," *Building Simulation Conference Proceedings*, Sep. 2023, doi: https://doi.org/10.26868/25222708.2023.1576.
[25] D. Leiria *et al.*, "Is it returning too hot? Time series segmentation and feature clustering of end-user substation faults in district heating systems," Submitted in *Applied Energy*, 2024.
[26] RStudio, Inc., "Shiny: Web Application Framework for R," R package version 1.6.0, Rstudio, Inc., 2021 [Online], Available: https://shiny.rstudio.com/ [Accessed: Jul. 5, 2024].
[27] P. Pandiyan, S. Saravanan, K. Usha, R. Kannadasan, M. H. Alsharif, and M.-K. Kim, "Technological advancements toward smart energy

management in smart cities," Energy Reports, vol. 10, pp. 648–677, Nov. 2023, doi: https://doi.org/10.1016/j.egyr.2023.07.021.
[28] J. van Dreven, V. Boeva, S. Abghari, H. Grahn, J. Al Koussa, and E. Motoasca, "Intelligent Approaches to Fault Detection and Diagnosis in District Heating: Current Trends, Challenges, and Opportunities," *Electronics*, vol. 12, no. 6, p. 1448, Jan. 2023, doi: https://doi.org/10.3390/electronics12061448.
[29] S. Bager and L. Mundaca, "Making 'Smart Meters' smarter? Insights from a behavioural economics pilot field experiment in Copenhagen, Denmark," *Energy Research & Social Science*, vol. 28, pp. 68–76, Jun. 2017, doi: https://doi.org/10.1016/j.erss.2017.04.008.
[30] IEA DHC/CHP, "IEA DHC Annex TS4 Guidebook," 2023 [Online], Available: https://www.iea-dhc.org/fileadmin/documents/Annex_TS4/IEA_DHC_Annex_TS4_Guidebook_2023.pdf. [Accessed: 05-Jul-2024].
[31] H. Lund *et al.*, "4th Generation District Heating (4GDH)," *Energy*, vol. 68, pp. 1–11, Apr. 2014, doi: https://doi.org/10.1016/j.energy.2014.02.089.
[32] T. Knayer and N. Kryvinska, "An analysis of smart meter technologies for efficient energy management in households and organizations," *Energy Reports*, vol. 8, pp. 4022–4040, Nov. 2022, doi: https://doi.org/10.1016/j.egyr.2022.03.041.
[33] A. Nissen, Hamid Reza Shaker, and Bo Nørregaard Jørgensen, "Enabling Alarm-Based Fault Prediction for Smart Meters in District Heating Systems: A Danish Case Study," *Smart Cities*, vol. 7, no. 3, pp. 1126–1148, May 2024, doi: https://doi.org/10.3390/smartcities7030048.
[34] A. Hansen, D. Leiria, Hicham Johra, and A. Marszal-Pomianowska, "Who Produces the Peaks? Household Variation in Peak Energy Demand for Space Heating and Domestic Hot Water," *Energies*, vol. 15, no. 24, pp. 9505–9505, Dec. 2022, doi: https://doi.org/10.3390/en15249505.
[35] M. Schaffer, J. Eduardo Vera-Valdés, and A. Marszal-Pomianowska, "Exploring smart heat meter data: A co-clustering driven approach to analyse the energy use of single-family houses," *Applied Energy*, vol. 371, pp. 123586–123586, Oct. 2024, doi: https://doi.org/10.1016/j.apenergy.2024.123586.
[36] R. Gopinath, M. Kumar, C. Prakash Chandra Joshua, and K. Srinivas, "Energy management using non-intrusive load monitoring techniques –

State-of-the-art and future research directions,” *Sustainable Cities and Society*, vol. 62, p. 102411, Nov. 2020, doi: https://doi.org/10.1016/j.scs.2020.102411.

[37] M. Schaffer, D. Leiria, J. E. Vera-Valdés, and A. Marszal-Pomianowska, “Increasing the accuracy of low-resolution commercial smart heat meter data and analysing its error,” in *Proceedings of the 2023 European Conference on Computing in Construction and the 40th International CIB W78 Conference*, European Council on Computing in Construction, 2023, pp. 761-767. doi: 10.35490/EC3.2023.208.

[38] M. Schaffer, J. Widén, J. E. Vera-Valdés, A. Marszal-Pomianowska, and T. S. Larsen, “Disaggregation of total energy use into space heating and domestic hot water: A city-scale suited approach,” *Energy*, vol. 291, p. 130351, Mar. 2024, doi: https://doi.org/10.1016/j.energy.2024.130351.

[39] K. Ritosa, D. Saelens, and S. Roels, “Estimating the as-built thermal performance of dwellings using simulated on-board data: From ideal to limited monitoring,” Energy and buildings, vol. 312, pp. 114171–114171, Jun. 2024, doi: https://doi.org/10.1016/j.enbuild.2024.114171.

[40] H. Bahlawan *et al.*, “Detection and identification of faults in a District Heating Network,” *Energy Conversion and Management*, vol. 266, p. 115837, Aug. 2022, doi: https://doi.org/10.1016/j.enconman.2022.115837.

[41] S. Månsson, P.-O. Johansson Kallioniemi, M. Thern, T. Van Oevelen, and K. Sernhed, “Faults in district heating customer installations and ways to approach them: Experiences from Swedish utilities,” *Energy*, vol. 180, pp. 163–174, Aug. 2019, doi: https://doi.org/10.1016/j.energy.2019.04.220.

[42] H. Lund, J. E. Thorsen, S. S. Jensen, and F. P. Madsen, “Fourth-Generation District Heating and Motivation Tariffs,” *ASME Open Journal of Engineering*, vol. 1, Jan. 2022, doi: https://doi.org/10.1115/1.4053420.

[43] A. Nissen, Hamid Reza Shaker, and Bo Nørregaard Jørgensen, “Automated and real-time anomaly indexing for district heating maintenance decision support system,” *Applied Thermal Engineering*, vol. 233, pp. 120964–120964, Oct. 2023, doi: https://doi.org/10.1016/j.applthermaleng.2023.120964