

Heterogeneity-Aware Building Stock Modelling for Urban Energy Transitions

Santhan Reddy Penaka



UMEÅ UNIVERSITY

This work is protected by the Swedish Copyright Act (Act 1960:729)
ISBN: 978-91-8070-961-3 (print)
ISBN: 978-91-8070-962-0 (pdf)
Cover image generated using GPT-5.3
Electronic version available at: <http://umu.diva-portal.org/>
Printed by: Scandinavian Print Group SE, 2026

"Dedicated to my Parents (Subhash & Anusha) and Brother (Sarath)"

Contents

- Contents.....i**
- Abstract iii**
- Sammanfattning v**
- Preface vii**
- List of papersviii**
- Author contributions ix**
- Additional Papers not included in the thesis x**

- 1 Introduction 1**
 - 1.1 Background..... 4
 - 1.1.1 *Stock-level energy retrofitting* 4
 - 1.1.2 *Occupant behaviour and heterogeneous modelling* 5
 - 1.2 Research objectives 6

- 2 Method9**
 - 2.1 Building performance datasets 9
 - 2.2 Data fusion..... 11
 - 2.2.1 *Probabilistic record linkage*12
 - 2.2.2 *Inverse modelling*13
 - 2.3 Integrated data-driven modelling.....15
 - 2.3.2 *Ensemble machine learning*17
 - 2.3.3 *SHAP (explainable AI)*.....18
 - 2.3.4 *Model convergence*18
 - 2.4 Diverse occupant behaviour and heterogeneous urban building stock modelling 20
 - 2.4.1 *Representative sampling*21
 - 2.4.2 *Automated simulation workflow*21
 - 2.4.3 *Impact assessment* 25
 - 2.5 Benchmarking..... 25
 - 2.6 Urban interactive visualization 26
 - 2.6.1 *Fuzzy matching*28
 - 2.6.2 *Spatial matching* 29
 - 2.6.3 *3D visualization platform*.....30

3 Results and discussion	31
3.1 Applied data fusion.....	32
3.1.1 <i>Heterogeneous building performance data</i>	32
3.1.2 <i>Thermal performance of residential building stocks</i>	34
3.2 Tailored and impactful retrofits.....	36
3.2.1 <i>Insights across heterogenous clusters</i>	36
3.2.2 <i>Stock-level post-retrofit assessments</i>	40
3.2.3 <i>Implications</i>	42
3.3 Behavioural response	43
3.3.1 <i>Seasonal impacts of excessive window-opening</i>	43
3.3.2 <i>Behavioural response under 2027 regulation</i>	45
3.4 Public engagement	48
3.4.1 <i>Benchmarking</i>	48
3.4.2 <i>Interactive visual analytics (IVA) platform</i>	49
4 Implications	52
5 Limitations and future work	53
6 Conclusion	54
Acknowledgements	56
References	59

Abstract

Bottom-up building stock modelling (BBSM) approach is widely used to assess the energy performance of urban building stocks by modelling individual buildings in detail and aggregating them to larger spatial scales. It plays an important role in supporting urban energy transition planning and policymaking. However, existing BBSM studies are constrained by several limitations, including incomplete building performance datasets, reliance on archetype-based averages in retrofit prioritization, simplified representations of occupant behaviour and building properties, and limited integration of modelling into public engagement tools. These limitations obscure inherent heterogeneity that determines which buildings benefit most from specific measures, leading to one-size-fits-all retrofit strategies and biased estimation of energy-saving potentials and policy effectiveness of energy transition initiatives.

This thesis advances heterogeneity-aware BBSM through an integrated and cumulative methodological pipeline that fuses incomplete building-performance datasets, enables localized retrofit prioritisation, and supports evaluation and communication of demand-side behavioural impacts. First, a data-fusion framework combines multiple incomplete datasets using probabilistic record linkage and inverse modelling, filling data gaps by transferring information across sources rather than relying on archetype-level averages. This improves stock representation by capturing building-to-building variation within the same urban context. Second, the thesis integrates data fusion with ensemble machine learning and explainable AI (SHAP) to identify impactful envelope retrofit measures in a local, data-driven manner. Across 81 building-stock clusters in three Swedish municipalities, the results demonstrate substantial variation in the most influential thermal components across municipalities and climate zones, underscoring the need for local-specific retrofit prioritisation. Third, the thesis incorporates occupant-behaviour diversity alongside building heterogeneity via an enhanced DOB-HUBS framework based on representative clustering, automated physics-based simulations, and surrogate machine learning. Empirical results of the Umeå building stock demonstrate that oversimplified behavioural and homogeneous assumptions can bias energy outcomes by up to 15% (standard deviation 3.06%). The framework is further used to assess Sweden's forthcoming 2027 capacity-based electricity tariff, indicating that behavioural adaptations could reduce peak electricity demand by 6–17%, with heterogeneous impacts across clusters.

Building on these modelling advances, the thesis extends toward public engagement by developing data-driven benchmarking and interactive visual analytics platform that translate bottom-up modelling outputs into user-facing insights, including peer comparison and ‘what-if’ exploration of retrofit and behavioural scenarios. Collectively, the thesis contributes methods and empirical evidence for more credible, locally tailored, and publicly actionable building-stock analytics. This supports in designing targeted retrofit strategies, effective behavioural measures, and informed public participation in urban energy transition planning.

Keywords: Urban energy transition, building stock, bottom-up, heterogeneity, tailored retrofitting, public engagement, machine learning

Sammanfattning

Bottom-up-baserad modellering av byggnadsbestånd (BBSM) används i stor utsträckning för att bedöma energiprestandan hos urbana byggnadsbestånd genom att modellera enskilda byggnader i detalj och därefter aggregera resultaten till större spatial skala. Metoden spelar en viktig roll för planering och upprättandet av policyer för urbana energiomställningar. Befintliga BBSM-studier begränsas av flera faktorer, däribland ofullständiga dataset gällande byggnaders energiprestanda, beroende av arketypbaserade genomsnitt vid prioritering av energieffektiviserande renoveringar, förenklade representationer av boendebeteende och byggnadsegenskaper samt begränsad integrering av modellering i verktyg avsedd för offentligt nyttjande. Dessa begränsningar döljer den inneboende heterogenitet som avgör vilka byggnader som har störst nytta av specifika åtgärder, vilket leder till generella renoveringsstrategier och snedvridna uppskattningar av potentialen för energibesparings och policyers effektivitet i energiomställningsinitiativ.

Denna avhandling utvecklar heterogenitetsmedveten BBSM genom ett integrerat och kumulativt metodologiskt flöde som kombinerar ofullständiga dataset över byggnaders energiprestanda, möjliggör lokalprioritering av renoveringsåtgärder och stödjer utvärdering samt kommunikation av beteendebaserade effekter på efterfrågesidan. För det första kombinerar ett datafusionsramverk flera ofullständiga dataset genom probabilistisk postkoppling (probabilistic record linkage) och inverterad modellering, vilket fyller dataluckor genom att överföra information mellan datakällor i stället för att förlita sig på arketypbaserade genomsnitt. Detta förbättrar representationen av byggnadsbeståndet genom att fånga variation mellan enskilda byggnader inom samma urbana kontext. För det andra integrerar avhandlingen datafusion med ensemblebaserad maskininlärning och förklarbar AI (SHAP) för att identifiera effektiva klimatskalåtgärder på ett lokalt och datadrivet sätt. I analyser av 81 kluster av byggnadsbestånd i tre svenska kommuner visar resultaten betydande variation i vilka termiska komponenter som är mest inflytelserika mellan kommuner och klimatzoner, vilket understryker behovet av lokalt anpassad prioritering av renoveringsåtgärder. För det tredje integrerar avhandlingen variation i boendebeteende tillsammans med byggnadsheterogenitet genom ett utökat DOB-HUBS-ramverk baserat på representativ klustring, automatiserade fysikbaserade simuleringar och surrogat-

maskininlärning. Empiriska resultat från byggnadsbeståndet i Umeå visar att förenklade beteendeantaganden och homogena antaganden kan snedvrider energiberäkningar med upp till 15 % (standardavvikelse 3,06 %). Ramverket används vidare för att analysera Sveriges kommande kapacitetsbaserade eltariff från 2027, där resultaten indikerar att beteendeanpassningar kan minska toppefterfrågan på el med 6–17 %, med heterogena effekter mellan olika kluster.

Med utgångspunkt i dessa modelleringsframsteg utvidgar avhandlingen även perspektivet mot offentligt nyttjande genom utvecklingen av en datadriven, interaktiv benchmarking- och visualiseringsplattform som översätter resultat från bottom-up-modellering till användarorienterad information, inklusive jämförelser med liknande byggnader och ”what-if”-utforskning av renoverings- och beteendescenarier. Sammantaget bidrar avhandlingen med metoder och empiriska resultat som möjliggör mer tillförlitlig, lokalt anpassad och praktiskt användbar analys av byggnadsbestånd. Detta stödjer utformningen av riktade renoveringsstrategier, effektiva beteendebaserade åtgärder och välgrundad offentlig medverkan i planeringen av urbana energiomställningar.

Preface

The building sector consumes around 30% of global energy use and 26% of global CO₂ emissions. Improving the energy performance of existing building stocks is therefore essential for reducing energy demand, lowering CO₂ emissions, and supporting national and municipal energy and climate targets. In this context, bottom-up building stock modelling (BBSM) has emerged as an important tool for evaluating energy performance, identifying retrofit opportunities, and supporting urban energy planning. This thesis investigates how BBSM can be enhanced to better represent the inherent heterogeneity of urban building stocks and support more informed decision-making in the urban energy transition.

This doctoral thesis was carried out at the Intelligent Human–Buildings Interactions (IHBI) Lab, Department of Applied Physics and Electronics, Umeå University, Sweden. This research was funded by multiple projects including European Union Horizon 2020 project ‘AURORAL’ (grant no. 101016854), Formas (2020-02085 & 2022-01475), and Swedish Energy Agency (P2022-00141 & 52686-1). My PhD journey was guided by my principal supervisor Weizhuo Lu, and co-supervisors Kailun Feng, Thomas Olofsson, and Anders Rebbling at Umeå University. This scope and direction of this research have been shaped by project objectives, regional stakeholders’ insights, and consortium meetings especially under the AURORAL project. The primary data required for research was provided by Boverket, the Swedish National Board of Housing, Building and Planning.

List of papers

This thesis is based on the following papers, which are included in the thesis:

- I. **S.R. Penaka**, K. Feng, T. Olofsson, A. Rebbling, W. Lu, ‘Improved energy retrofit decision making through enhanced bottom-up building stock modelling’, *Energy and Buildings* (2024) 114492. <https://doi.org/10.1016/j.enbuild.2024.114492>.
- II. **S.R. Penaka**, K. Feng, T. Olofsson, A. Rebbling, E. Eklund, W. Lu, ‘An integrated framework for tailored building envelope retrofits in Swedish municipalities: from heterogeneous big data to explainable AI’, *Engineering, Construction and Architectural Management* (2025). (Under Review)
- III. **S.R. Penaka**, K. Feng, T. Olofsson, W. Lu, ‘Diverse occupant behaviour and urban building heterogeneity to enhance urban building energy modelling’, *Energy and Buildings* (2025) 116721. <https://doi.org/10.1016/j.enbuild.2025.116721>.
- IV. B. Liu, **S.R. Penaka**, W. Lu, K. Feng, A. Rebbling, T. Olofsson, ‘Data-driven quantitative analysis of an integrated open digital ecosystems platform for user-centric energy retrofits: A case study in northern Sweden’, *Technology in Society* 75 (2023) 102347. <https://doi.org/10.1016/j.techsoc.2023.102347>.
- V. K. Feng, **S.R. Penaka**, W. Lu, ‘Interactive visual analytics platform for public engagement in energy-efficient building transitions: Evidence from a Nordic city empirical study’, *manuscript*.

Author contributions

Paper I, Paper II & Paper III:

As first author, I led the conceptualization and development of methodological frameworks presented in these studies. I carried out the modelling, simulations, data analysis, and wrote the original manuscript drafts and revised them during review stages.

Paper IV:

As second author, I developed the building energy benchmarking framework presented as an application in the study. In addition, I contributed to writing parts of the original draft across application-related sections of the paper.

Paper V:

As second author, I developed the IVA platform for the Umeå empirical study. I contributed to conducting physical experiments with residents. I conducted simulations of energy retrofitting related results in the empirical study. In addition, I contributed by writing Section 2 of the original manuscript draft.

Additional Papers not included in the thesis

- I. **S.R. Penaka**, K. Feng, A. Rebbling, S. Azizi, W. Lu, T. Olofsson, ‘A data-driven framework for building energy benchmarking and renovation decision-making support in Sweden’. *IOP Conference Series: Earth and Environmental Science* (2023) 012005. <https://doi.org/10.1088/1755-1315/1196/1/012005>.
- II. **S.R. Penaka**, K. Feng, W. Lu, ‘Impact of thermal properties on building stock energy use using explainable Artificial Intelligence’, *International Conference on Construction and Real Estate Management 2024* (2025) 870–878. <https://doi.org/10.1061/9780784485910.084>.
- III. K. Feng, **S.R. Penaka**, H. Yu, S. Chen, W. Lu, ‘Projecting climate resilience of urban building stocks: A data-augmented archetype approach for future Nordic climates’, *Energy and Buildings* (2026) 117260. <https://doi.org/10.1016/j.enbuild.2026.117260>.
- IV. K. Feng, W. Lu, **S.R. Penaka**, E. Eklund, S. Andersson, T. Olofsson, ‘Energy-efficient retrofitting with incomplete building information: a data-driven approach’, *ROOMVENT 2022*. <https://doi.org/10.1051/e3sconf/202235601003>.
- V. C. Lu, S. Li, **S.R. Penaka**, T. Olofsson, ‘Automated machine learning-based framework of heating and cooling load prediction for quick residential building design’, *Energy* (2023) 127334. <https://doi.org/10.1016/j.energy.2023.127334>.
- VI. H. Yu, K. Feng, **S.R. Penaka**, Q. Man, W. Lu, T. Olofsson, ‘Data-driven modelling of building retrofitting with incomplete physics: A generative design and machine learning approach’, *13th Nordic Symposium on Building Physics (NSB-2023)*. 10.1088/1742-6596/2654/1/012053.

1 Introduction

Rising global energy demand and growing urgency of climate change have made the energy-efficient retrofitting more important than ever to reduce energy use in buildings. Recent spikes in energy prices across Europe further highlights the importance of energy efficiency for lowering energy demand and household costs. At the same time, several policy initiatives [1,2] have been introduced to reduce energy demand and CO₂ emissions from the buildings sector, which remains responsible for nearly 40% of Europe's final energy use and approximately 36% of related CO₂ emissions. This challenge is compounded by ageing building stock, more than 30% of buildings are now older than 50 years and nearly 75% are considered energy inefficient. Consequently, prioritizing the retrofitting of existing building stocks has become essential for achieving energy and climate targets [3].

This situation is particularly visible in Sweden. Around 1.4 million buildings were built between 1946 and 1975 [4], and a large share of them will be over 50 years old by 2030. These buildings often have minimal building envelope thermal insulation and lack heat-recovery ventilation systems [5]. Consequently, Swedish municipalities have set local energy-efficiency targets as part of their wider carbon-neutrality ambitions, and many have even signed Climate Contract 2030 to accelerate their transition efforts [6]. To meet these goals, municipal stakeholders require localized building stock-level decision-making knowledge, particularly in identifying the effective retrofitting strategies to design or policymaking stage [7]. In particular, the ability to determine which tailored retrofit measures are most impactful for each building stock cluster, considering the inherent heterogeneity (e.g., building age, size, energy use) within the overall municipal stock.

Also, occupants/public play a crucial role in this urban energy transition. Policies such as the capacity-based electricity tariff, an upcoming regulation in 2027 aims to improve electricity grid stability by incentivizing consumers/users to distribute their energy usage more evenly throughout the day and reduce peak demand. Such initiatives also emphasize the importance of public engagement in this energy transition. However, engagement would require accessible knowledge, such as understanding their building's energy performance (e.g., benchmarking against similar building clusters within the stock) [8], retrofit measures to prioritize at their subjective cluster level (e.g.,

changing window, adding wall insulation), and potential impacts of behavioural changes (e.g., window-opening).

BBSM is widely used to assess the energy performance of building stocks. In this approach, individual buildings are modelled in detail and then aggregated to the stock level, enabling performance analysis from neighbourhood to national scales. Based on previous literature, BBSM uses a representative sample stock, housing surveys, and modeller assumptions, combined with building energy calculation methods, to estimate energy use. While such models can simulate various combinations of technological measures and provide valuable insights for policymakers, they require a vast amount of bottom-up detailed data [9]. BBSM can support the design of tailored and localized energy-retrofit strategies, and the evaluation of different occupant behavioural patterns. However, several limitations have been identified in previous studies, and this PhD research aims to enhance BBSM capabilities by addressing the following key limitations:

1) Homogeneous representation of heterogeneous building stocks: Naturally, BBSM requires detailed information on building characteristics, including construction year, size, envelope thermal properties (U-values), and energy use etc. Although recent studies increasingly utilise energy performance certificate (EPC) data [10–12] and national surveys such as BETSI [13], no single data source provides comprehensive information on all relevant parameters. To address these challenges, many studies relied on average or statistical values for missing data.

While such approaches facilitate large-scale modelling, they homogenise inherently diverse building stocks and obscure substantial variation in construction quality, renovation history, and envelope performance. Buildings with similar age, type, or location may exhibit different thermal characteristics and energy performance, which are overlooked when represented by average parameters. As a result, stock-level energy estimates and retrofit potentials may be misrepresented. Therefore, there is a need for more approaches that integrate multiple incomplete datasets into a single comprehensive dataset while preserving inherent heterogeneity. This is detailed in Paper 1.

2) Generic retrofit prioritization: Among various building characteristics, envelope components including walls, roofs, windows, and floors are primary source of heat loss and gain, making their U-values (rate of heat transfer) crucial to overall energy performance [14]. Retrofitting existing building envelopes offers a significant opportunity

to reduce the energy use. However, the effectiveness of individual retrofit measures varies across buildings due to differences in architectural design, construction materials, building codes, and maintenance practices or retrofitting [15,16]. Consequently, the relative importance of individual envelope components could vary significantly between building clusters.

Despite this, retrofit prioritization in many BBSM studies is based on generalized or simplified approaches, including comparing envelope U-values against regulatory thresholds or targeting components with poorest nominal thermal performance. These approaches implicitly assume that upgrading the poorly performing component will yield the largest energy savings. However, in practice, the energy impact of retrofitting a specific envelope component depends on its relative contribution to total energy demand and its interaction with other building characteristics, climatic conditions, and operational patterns. Components with high U-values may have limited impact on overall performance if other features dominate heat losses, while moderately performing components may yield substantial savings when considered within a system-wide context. Without systematic relative assessment, retrofit strategies may result in misrepresenting energy savings and misallocate investment resources. This is detailed in Paper 2.

3) Modelling uniform occupant behaviour and homogeneous building properties: Occupant behaviour significantly influences building energy use through factors such as occupancy patterns (ways that people arrive, depart, move, and use spaces in a building) and interactions (e.g., window-opening, appliance usage). At the urban scale, evaluating these behavioural impacts is essential for assessing energy performance and policy effectiveness. However, behavioural patterns vary substantially across buildings due to demographic characteristics, seasonal preferences, and lifestyle differences [17]. Directly assigning unique occupancy and interaction schedules to every individual building in urban stock is impractical, as it requires detailed occupant-level data that are rarely available and would be computationally infeasible to manage at large scale. Urban populations include working, non-working, and retired groups etc., each with distinct occupancy and interaction patterns [17]. This diversity becomes even more pronounced when considering behaviours such as window opening, which vary significantly across seasons. For example, occupants tend to open windows more in summer, less in winter, and moderately in spring and autumn [18,19]. Moreover, occupant interactions and building properties are strongly interdependent. For instance, interactions such

as window opening and use of heating systems directly affect heating and cooling loads. However, the magnitude of these impacts is highly dependent on envelope U-values. Given the inherent heterogeneity of properties across the urban stock, even identical OB interactions can lead to varied energy impacts [20]. However, most existing studies assume uniform occupant behaviour and homogeneous building properties due to modelling limitations to incorporate diversity and heterogeneity, which is computationally infeasible at large scales [20,21]. This results in misrepresenting energy potentials and policy effectiveness. This is detailed in Paper 3.

4) Public engagement: While occupants/public play a critical role in achieving urban energy transition goals, most BBSM studies remain predominantly focused on technical energy modelling, with limited emphasis on public engagement and participation. This gap is particularly relevant in the context of emerging policies, such as capacity-based electricity tariffs, which rely on behavioural change. Interactive visual analytics (IVA) platforms have been one of the effective means to engage public awareness and participation through accessible decision-making knowledge [22–25]. Several previous attempts have demonstrated the potential of IVA in building energy and residential decision-making contexts [22–24]. In parallel, behavioural energy research has shown that peer influence and social comparison play a significant role in shaping household energy behaviours [8]. This suggests that IVA platforms integrating benchmarking feedback can effectively support informed public engagement. Such engagement includes understanding of a building’s energy performance, peer comparison, and exploring potential impacts of retrofit measures and behavioural change. This is detailed in Paper 4 and Paper 5.

To address these limitations, this thesis includes five complementary papers presenting enhanced frameworks demonstrated through empirical studies in Sweden, as illustrated in section 1.2.

1.1 Background

1.1.1 Stock-level energy retrofitting

Several previous studies have focused on building stock level retrofitting. Liu et al., [26] conducted a detailed study focused on energy savings and environmental impact assessments for eleven multi-family buildings in Sweden. By evaluating various retrofitting measures, they discovered the potential to reduce half of energy use and 48% of CO₂ emissions.

Similarly, Martin et al., [27] assessed the before and after retrofitting of multifamily building stock in Gävle, where renovations including adding roof insulation, external wall insulation, installing new windows, ground flooring, decentralized hot water piping, replacing the exhaust ventilation system with a bi-directional air supply and exhaust. Wang et al., [28] proposed an ensemble ML learning approach for building energy prediction. This approach employs bagging trees to achieve robust predictions and enhanced prediction accuracy. Feng et al., [29] proposed a data-driven method leveraging Swedish building performance datasets for retrofit decision-making under incomplete data conditions, and it is demonstrated using an empirical study which achieved over 90% model performance. Alrobaie and Krarti [30] conducted a comprehensive review of data-driven approaches for building stock retrofitting, emphasizing the advantages of ML techniques over traditional deterministic modelling approaches, driven by the increasing availability of building performance data. Recent studies [31–33] emphasize the shift from generalized retrofitting strategies to more tailored approaches. These tailored approaches require heterogeneous building performance data (e.g., age, size, energy use, U-values) to consider the inherent heterogeneity within the building stocks.

1.1.2 Occupant behaviour and heterogeneous modelling

OB is one of the important aspects influencing building's energy performance, alongside building geometry, weather, HVAC systems, and envelope thermal properties. Similarly, the inherent heterogeneity of building properties such as thermal transmittance (U-values) plays in crucial role in evaluating OB impacts [20,21]. While incorporating these aspects is well established at the single building level [34], but at large stock level UBEM modelling remains complex and computational complexity. Many existing UBEM studies have addressed this complexity through simplified assumptions such as uniform OB schedules [35,36], and average U-values based on building type or archetypes [31,37]. This assigns the same OB schedule and U-values for all the buildings of the same cluster. Sanam et al., [38] and Gabriel et al., [39] reviewed this is mainly due to lack of detailed occupant data to capture the realistic patterns and the limitations of current modelling approaches to incorporate diverse occupant schedules at stock level. While some studies have explored different approaches like probabilistic profiles to address this data challenge [40], the challenge of integrating diverse OB schedules into UBEM modelling remains. Aya and Mohamed [41] highlights that commonly used UBEM tools are limited in assigning static or standardized schedules (e.g., ASHRAE standards). Some tools

facilitate custom scripting capabilities to automate complex and repetitive tasks within physics-based simulation environments. For instance, OpenStudio's software development kit (SDK) [42] integrated with the EnergyPlus simulation engine, supports programmatic scripting using Ruby and Python. Wesley et al. [43] used OpenStudio to automate the EnergyPlus simulations to create reduced-order models for real-time predictive control in residential buildings. Exploring such complex capabilities in the context of OB modelling makes it possible to create and assign unique OB schedules for each building in a large stock, moving beyond simplified assumptions and allowing more representative evaluation of diverse OB simulations. However, such capabilities require significant computational efforts for simulation all unique buildings. To address such computational complexity, clustering techniques are increasingly used. For instance, K-Means clustering has been widely applied in building energy applications such as archetype selection [44], energy use patterns analysis [45], benchmarking [46]. By clustering building samples based on input features, a small set of representative buildings can be selected for bottom-up physics-based simulations. The results from these representative simulations can then be used to train surrogate machine learning models which can rapidly predict the energy performance of the entire building stock [28,47]. Therefore, these advancements in other contexts along with the identified gaps in current practices provide the foundational hypothesis for the Paper 3.

1.2 Research objectives

Based on the identified four limitations, this thesis defined the following four research objectives, as also illustrated in Figure 1:

RO1: To integrate multiple incomplete building performance datasets by capturing inherent heterogeneity within the building stock.

RO2: To identify tailored and impactful retrofits using data-driven techniques & heterogeneous building performance data.

RO3: To incorporate diverse occupant behaviour & urban building heterogeneity into building stock modelling.

RO4: To support public engagement through benchmarking & IVA based platform using open-source datasets and data-driven techniques.

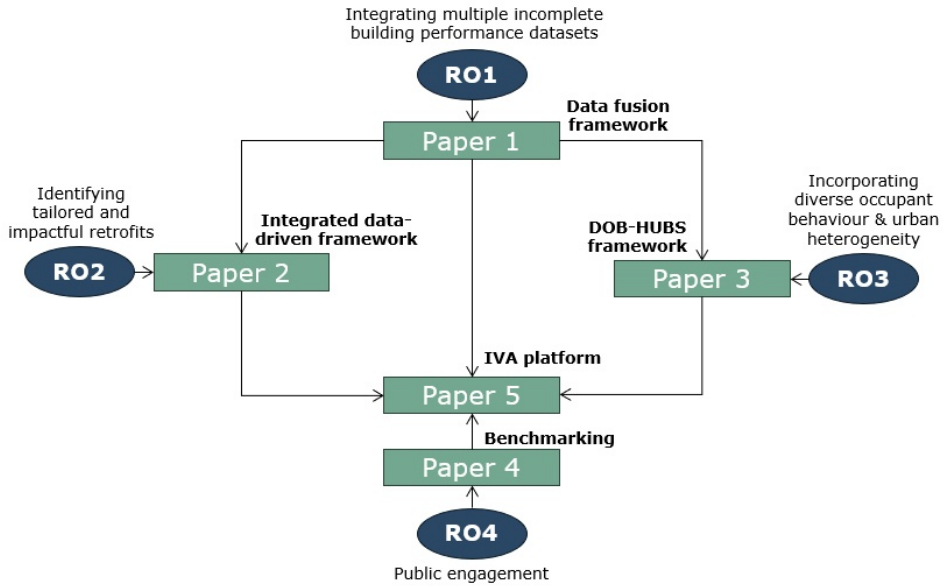


Figure 1. Overview of research flow.

Based on these research objectives, the five papers are connected and build on each other. Paper 1 addresses RO1 by addressing the incomplete data challenge by developing a data fusion framework that integrates multiple incomplete building performance datasets using probabilistic record linkage and inverse modelling. It demonstrates how BBSM abilities are enhanced by capturing inherent heterogeneity within building stock. Thus, establishes comprehensive datasets for the subsequent research. Leveraging this enhancement in Paper 1, Paper 2 focuses on RO2 to identify tailored and impactful building envelope retrofit strategies. Using explainable AI (SHAP) and ensemble machine learning (ML), it evaluates retrofit impacts across 81 heterogeneous building stock clusters in three Swedish municipalities. This demonstrates how data-driven techniques and heterogeneous data can support stakeholders in effective planning and prioritizing retrofit strategies. Paper 3 expands the thesis scope to address RO3 by incorporating diverse occupant behaviour (e.g. seasonal, demographic) alongside urban heterogeneity into building stock modelling. Building on data and ensemble modelling from Paper 1 and 2, Paper 3 introduces an enhanced framework that evaluates impacts of different behavioural changes across heterogeneous clusters. Building on these advancements, Papers 4 and 5 address RO4 by developing IVA based platform for public engagement, using standard open-source datasets and data-driven techniques. The allows homeowners to understand their current building

energy performance, peer comparison, and exploring impacts of ‘what-if’ retrofit and occupant behavioural scenarios at their tailored cluster level. By integrating IVA into the BBSM workflow, this work supports public awareness, participation, and informed decision-making in the urban energy transition.

2 Method

This section presents the frameworks introduced to address research objectives corresponding to appended papers. Section 2.1 introduces the primary datasets used in this thesis. Section 2.2 presents a data fusion framework proposed to address the incomplete data challenge related to RO1, corresponding to Paper 1. Section 2.3 presents an integrated data-driven modelling approach for designing tailored and impactful retrofitting, addressing homogeneous retrofit prioritization limitation related to RO2 and corresponding to Paper 2. Section 2.4 presents diverse occupant behaviour and heterogeneous urban building stocks (DOB-HUBS) framework which incorporates diverse occupant behaviour and heterogeneous building properties into building stock modelling connected to RO3 corresponding to Paper 3. Sections 2.5 and 2.6 presents methods for public engagement strategies connected to RO4 corresponding to Paper 4 and Paper 5.

2.1 Building performance datasets

The EPC was introduced as part of European Union Energy Performance of Buildings Directive [48]. This EPC database in Sweden is handled by the Swedish National Board of Housing, Building and Planning (Boverket), which consists of building energy performance certifications given by independent certified energy experts to property owners when they are selling the properties, for new buildings, and for buildings frequently visited by the public. An EPC is valid for ten years and available data is shown in Table 1. While EPC data consist of both measured and calculated energy use values depending on energy type and data availability [49], they provide representative standardized dataset suitable for empirical analysis across large sample of buildings. Another database, BETSI [13] is based on a nationwide survey conducted by Boverket in 2007-2008 to understand the technical status of the buildings across Sweden. The survey employed rigorous on-site measurement protocols and a standardized procedure involving inspection of 1800 buildings for determining the thermal performance of buildings components. This detailed information consisted in the dataset is shown in Table 2.

Table 1. Building information in a EPC (energy performance certificate) record sample [29].

No.	Building property
1	Location (province, city, postal code)
2	Building type (detached/attached)
3	Construction year
4	Heated floor area ($A_{temp}^{* (a)}$, m ²)
5	Number of floors
6	Number of stairs
7	Number of apartments
8	Number of people
9	Connected electrical power for heating and water (>10 W/m ²)
10	Energy use for space heating, DHW and households* (kWh)
11	Total energy use* (kWh)
12	Electricity use in total building energy use (kWh)
13	Normal year adjusted value (degree days)
14	Normal year adjusted value (Energy Index* ^(b))
15	Energy use per area (kWh/m ²)
16	Energy use per area of which electricity
17	Energy performance rating* ^(c)
18	Requirement for regular ventilation control in the building (Yes/No)
19	Ventilation systems (Yes/No)
20	Available air-conditioning systems with nominal cooling power greater than 12 kW (Yes/No)
21	Data of approval
22	EPC version

* According to the definition in regulations of Boverket (the Swedish National Board of Housing, Building, and Planning)

A_{temp} (heated floor area) is a standard measure representing building floor area that is heated above 10°C temperature, excluding storage areas. This parameter is generally used to assess the building's energy efficiency as it directly affects the energy required for heating [49].

Normal year adjusted value (energy index) is a normalized measure of building's energy performance, adjusted to standard climate conditions in a 'normal year'. This adjustment allows fair comparison between buildings and across different years that may have varying weather conditions.

Energy performance rating is typically measured by a scale from A to G, with A being the most energy-efficient and G the least. It is used to indicate the building energy performance which is based on energy use intensity (kWh/m²) and year.

Table 2. Building information in a BETSI record sample.

No.	Building property
1	Location (municipality)
2	Climate zone
3	Building type (detached/attached)
4	Construction year
5	Application* (villa/apartment/public)
6	Heated floor area* (A_{temp}^{*} , m ²)
7	Number of floors above ground
8	Number of floors below the ground
9	Number of stairs

10	Number of apartments
11	U-value window (W/m ² . K)
12	U-value exterior walls (W/m ² . K)
13	U-value sloping roof (W/m ² . K)
14	U-value main floor (W/m ² . K)
15	U-value attic floor (W/m ² . K)
16	U-value basement wall underground (W/m ² . K)
17	U-value basement wall above ground (W/m ² . K)
18	Overall U-value average (W/m ² . K)
19	U-value average including cold bridges (W/m ² . K)
20	Construction materials
21	Renovation history (if available)
22	Heating supply systems (Yes/No)
23	Date of inspection

* According to the definition in regulations of Boverket (the Swedish National Board of Housing, Building, and Planning)

Both these datasets consist of useful data for building stock modelling and share variables such as location, archetype, construction year, and floor area. The EPC dataset lacks U-values, while the BETSI database includes U-values but is missing several crucial data that is available in EPC.

2.2 Data fusion

This framework, as shown in Figure 2, involves a two-step process for merging and refining multiple building performance datasets. The first stage, probabilistic record linkage (PRL) links individual building records from incomplete datasets into a single comprehensive dataset. This linkage is based on match probability scores calculated for each common variable shared between the datasets. The second stage, inverse modelling, further refines any unrealistic parameters within the newly linked records using known parameters and specific constraints. PRL utilizes these shared variables to incorporate the missing parameters from one dataset into another dataset, thus achieving a merged comprehensive dataset. These scores are computed for all identified common variables between the two datasets. Through inverse modelling, the linked parameters are refined using the metropolis algorithm, which utilizes known input parameters and known output to minimize the difference between predicted output and actual output, while parameters not subjected to refinement retain their initial values. This framework is implemented in a Python environment using multiple toolkits, including record linkage, scikit-learn, pandas and NumPy.

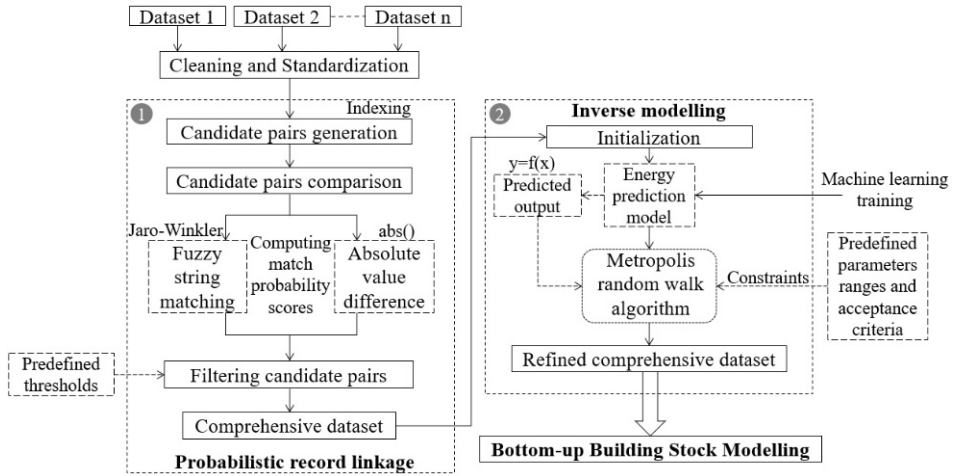


Figure 2. Overall Data Fusion Framework (from Paper 1).

2.2.1 Probabilistic record linkage

This section provides an overview of the PRL process. While PRL is complex and requires user knowledge, with a proper understanding, it can be easily implemented using established toolkits and algorithms detailed in this section. However, this implementation is categorized into three steps:

1. *Candidate pairs generation*: This step involves generating possible candidate pairs of records by pairing each building record in one dataset with every record in another dataset, assuming they share common variables. This pairing creates an exploratory space for PRL to identify potential matches for missing parameters. Since it is a computationally intensive process, this is implemented using a standard indexing algorithm [50] in Python record linkage toolkit [51].
2. *Candidate pairs comparison*: For each possible candidate pair of records, the common variables from both distinct datasets are compared, and their individual match probability score is computed. For string (textual) common variables (e.g. architecture type, location) matching comparison, the Jaro-Winkler fuzzy string matching [52] is employed, which is a widely used [53] technique to compare string values and compute match probability scores ranging between 0 and 1. This matching score is given based on the number of matching characters between two strings and their transpositions. If the quality of subjected records is high with no data discrepancies such as spelling errors, nicknames,

abbreviations other inconsistencies, then the matching score is 1. For numerical common variables (e.g. building construction year, floor area) matching comparison, the absolute value difference [54] is employed which is a built-in function in Python. This function compares two numerical values and computes binary output: if the difference between the values is within a pre-defined threshold, the match probability score is 1 and otherwise 0. For instance, if the threshold distance is defined as 10 for a building constructed in 1990, the building records from 1981 to 2000 are considered potential matches. Thus, this is performed for all the identified common variables in each record pair. Therefore, the comprehensive dataset is formed by compiling these scores for each candidate pair.

3. *Filtering candidate pairs*: In this step, all the possible candidate pairs of records are filtered based on the predefined thresholds of matching scores. These thresholds are introduced individually for all the common variables to control the acceptance of candidate pairs with high match probability scores. The thresholds shall be defined based on data quality and research context i.e., sensitivity of data parameter. For instance, if the sum of the matching scores across all common variables is equal to the number of common variables, that means the subjected candidate pair is considered a perfectly matched record.

2.2.2 Inverse modelling

Inverse modelling (IM) is a process where known or observed data is used to estimate unknown or uncertain system parameters. While forward modelling (FM) typically predicts outputs based on known input parameters, IM focuses on adjusting the input parameters using the known outputs. The IM in this study is used to refine any unrealistic parameters within the building records from PRL stage. The IM concept is inspired by the novel inverse data-driven modelling approach for performance based building design in the early stages, as introduced by Roya Rezaee et al. [55].

To implement this, the metropolis algorithm is employed, which is a variant of the Markov Chain Monte Carlo (MCMC) method, MCMC methods are commonly used for creating new samples from complex probability distributions [56,57]. This computationally intensive algorithm is simplified and explained in following two step process:

2.2.2.1 Initialization

Firstly, the ‘linked parameters’, ‘known input parameters’ and ‘actual energy use’ from the comprehensive dataset are used to train a regression model and predict ‘predicted energy use’. This model shall be trained to learn the relationship between input parameters and output energy use. In addition, the absolute difference between ‘predicted energy use’ and ‘actual energy use’ is computed for all building records.

2.2.2.2 Sampling and acceptance

In this method, metropolis algorithm’s primary objective is to create new synthetic samples to ‘linked parameters’ and adjust them within a defined exploratory space, while minimising the residual between ‘predicted energy use’ and ‘actual energy use’. The algorithm operated on two foundational sub-stages: sampling and acceptance.

At the sampling stage, the algorithm proposes a new synthetic set of parameters based on the posterior distribution of the ‘linked parameters’ thereby creating an exploratory space (Figure 3, box 2). Further, two constraints are introduced to define the boundaries of this space: 1) Parametric ranges are determined for the ‘linked parameters’ to avoid unrealistic synthetic samples (Figure 3, box 1). 2) The model parameters i.e., number of steps and step size are introduced. The number of steps or iterations represents the volume of newly created synthetic samples, and step size represents the magnitude of deviation from the old parameter value in each step. These model parameters should be optimized for specific research context.

At the acceptance stage, the new synthetic set of parameters is reviewed by introducing the acceptance criteria, ensuring that only realistic adjustments are accepted (Figure 3, box 4). The criteria are: 1) The residual between ‘predicted energy use’ and ‘actual energy use’ should be nearly zero for each building record for the algorithm to continue its assessment (Figure 3, box 3). Otherwise, the specific new set of parameters is rejected. 2) The acceptance probability is determined by taking the exponential of the difference in residuals between the previously accepted parameters and newly proposed parameters. This is helpful when there are multiple potential adjustments, the set with the least residual is identified as the optimized adjustment for that particular record.

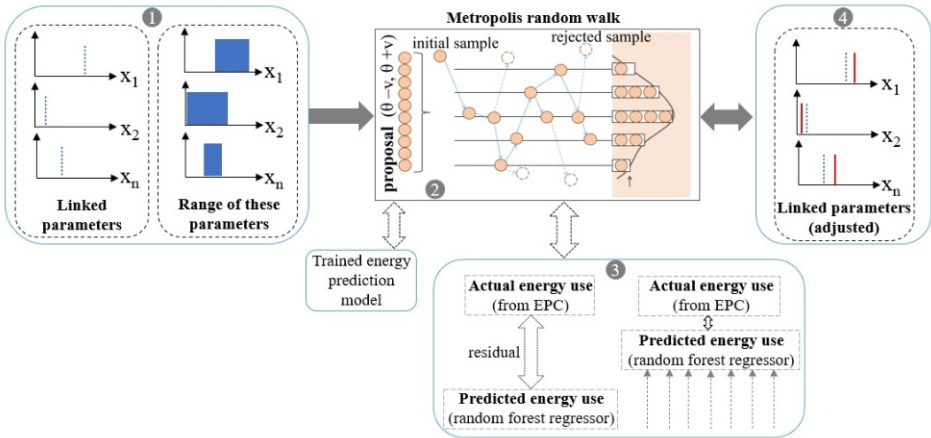


Figure 3. Workflow of inverse modelling using metropolis random walk (from Paper 1).

Essentially, for each newly created synthetic sample within exploratory space, it makes a probabilistic assessment to either accept or reject the move, depending on the residual between ‘predicted energy use’ and ‘actual energy use’. When new samples exceed the introduced constraints and acceptance criteria, the model reverts to default samples.

2.3 Integrated data-driven modelling

The integrated data-driven framework, illustrated in Figure 4, is developed to identify tailored and impactful building envelope retrofit strategies for diverse residential building stocks across three Swedish municipalities.

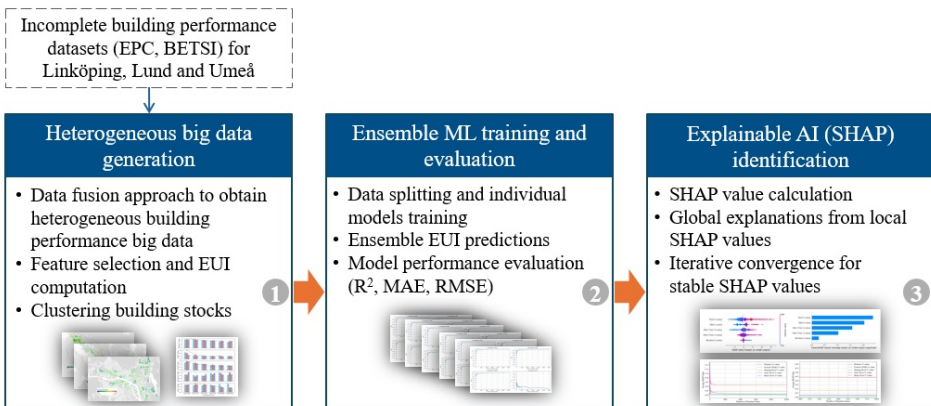


Figure 4. Integrated data-driven framework (from Paper 2).

The heterogeneous building performance big data is obtained using the data fusion approach presented in Paper 1. An ensemble machine learning (ML) model is trained to predict energy use intensity (EUI) for each building stock individually using multiple random data splits to ensure robust and unbiased predictions. Explainable AI (XAI) based SHAP (SHapley Additive exPlanation) is used to quantify the contribution of building features to each prediction. SHAP values are computed for each iteration and then averaged to ensure stable and generalized interpretations. The framework is implemented in a Python environment using scikit-learn, pandas, NumPy, SHAP and matplotlib.

Sweden is a country with diverse climate zones, therefore, three municipalities i.e., Linköping, Lund and Umeå are chosen from different climate zones and also those signed Climate Contract 2030 [58]. As a result, the final obtained datasets contain 14,846 (Linköping); 13,372 (Lund); and 13,954 (Umeå) residential building samples, including both single-family and multi-family buildings.

2.3.1.1 Clustering building stocks

The building stocks from three municipalities are clustered based on characteristics: construction year, application type, and energy performance class. Construction years are categorized into six clusters following the TABULA classification [59]. Application types are categorized as single-family or multi-family buildings. Energy performance is defined using EPC ratings in accordance with Boverket's building regulations (BBR, BFS 2011:6) code [60]. These classes, ranging from A to G, are consolidated into three clusters: A-C, D-E, and F-G, to ensure adequate sample sizes. Although 108 clusters are theoretically possible, considering adequate sample sizes (no. of samples > 50), this is reduced to 81 clusters. Figure 5 shows distribution of average U-values across these stocks, compared against national reference values [59].

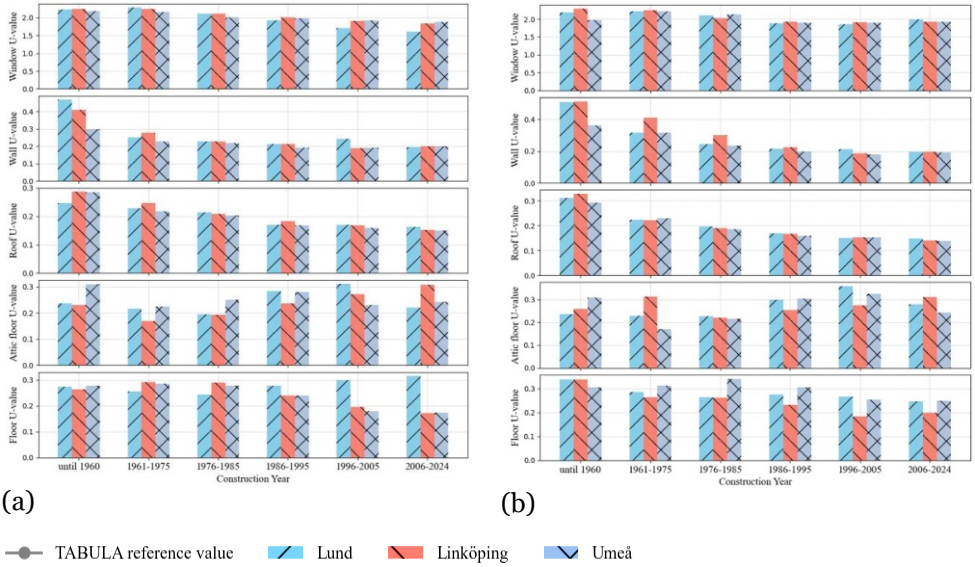


Figure 5. Distribution of U-values of Lund (sky blue), Linköping (salmon) and Umeå (light steel blue) residential building stocks. a) single-family, b) multi-family, (from Paper 2).

2.3.2 Ensemble machine learning

An ensemble ML model is developed to predict EUI and compute associated SHAP values. The model integrates three widely used decision-tree based algorithms i.e., Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting (LightGBM) to enhance generalization by leveraging complementary training mechanisms. RF is a bagging approach that trains multiple trees in parallel on bootstrap samples to reduce variance. In contrast, XGBoost and LightGBM are boosting-based algorithms that sequentially build trees to reduce bias. XGBoost is used for its ability to control model complexity, while LightGBM is known for effective model generalization with computational efficiency. Each algorithm is trained and evaluated independently, and their predictions are averaged to obtain ensemble prediction, balancing the bias-variance trade-off as demonstrated in previous studies [61,62]. For each random iteration, the dataset of each building stock is split into 80% training and 20% testing sets. Hyperparameters are optimized using GridSearchCV, and the model performance is evaluated using coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) [63].

2.3.3 SHAP (explainable AI)

SHAP is an explainable AI technique from cooperative game theory, used to interpret and quantify the contribution of input features to the output predictions of ML models, that are often regarded as black-box models [64]. In this study, SHAP values represents the impact of input features on the predicted heating EUI, considering the complex interactions between the features. The SHAP value (ϕ_i) for the i -th feature is calculated using the following equation [65],

$$\phi_i = \sum_{S \subseteq Z_{\{i\}}} \frac{|S|!(|Z| - |S| - 1)!}{|Z|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (5)$$

Where Z is the set of all input features, S is a subset of Z excluding the i -th feature. $\frac{|S|!(|Z| - |S| - 1)!}{|Z|!}$ represents the contribution weight of each subset contributes to prediction. The terms $f_{S \cup \{i\}}(x_{S \cup \{i\}})$ and $f_S(x_S)$ denote predictions with and without the i -th feature, respectively.

Figure 6 shows a sample SHAP summary plot of building U-values (mean impact values) on Linköping municipality and relationships between each feature and the EUI predictions. Each data point in the plot corresponds to one sample, with colour legend representing the feature value. The summary plot show that roof and wall U-values have high impact compared to attic floor, main floor and window U-values.

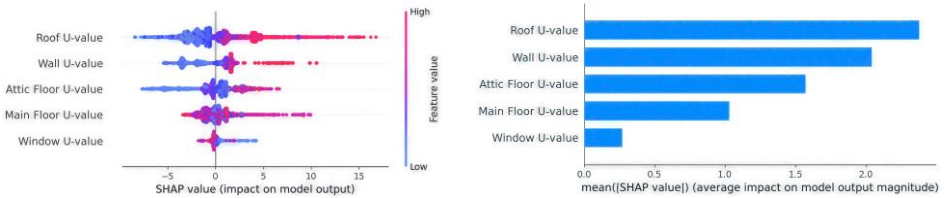


Figure 6. Sample SHAP summary plot of building U-values on Linköping municipality building stock (global interpretation) (from Paper 2).

Furthermore, to ensure stability and generalizability, SHAP values are averaged over many random iterations.

2.3.4 Model convergence

This section highlights the main strategies of the overall framework to ensure generalized SHAP values. The ensemble ML framework is iterated over many random data splits (e.g., 10,000) to avoid bias from any specific dataset segmentation. SHAP values are computed for each

iteration and averaged, resulting in stable feature attributions for each of the 81 building stocks. A sample convergence plot is shown in Figure 7.

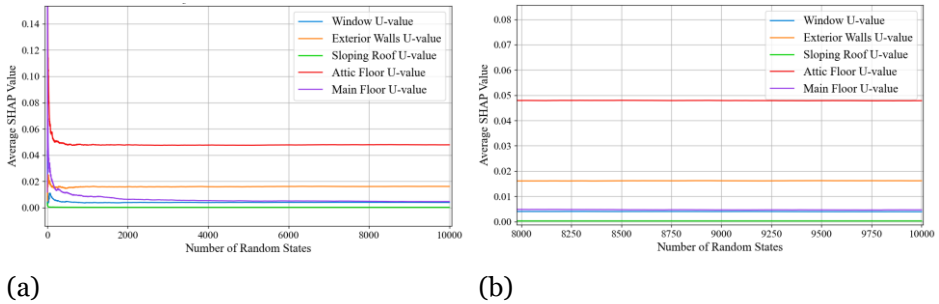


Figure 7. Sample convergence plot over random data splits. a) full convergence, b) magnified convergence, (from Paper 2).

Model performance metrics R^2 , MAE, RMSE are evaluated for each iteration on both training and testing datasets. Convergence of these metrics for a sample building stock is shown in Figure 8. The results indicate high accuracy and consistency, which further supports the reliability of the associated SHAP values.

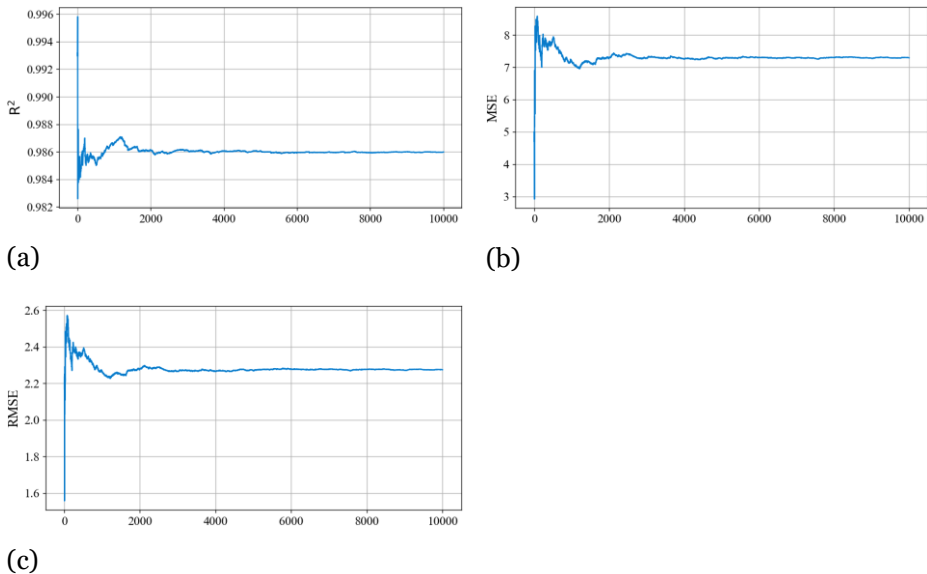


Figure 8. Sample convergence plot of model performance metrics. a) coefficient of determination (R^2), b) mean absolute error, c) root mean square error, (from Paper 2).

Finally, the employed ensemble approach of XGBoost, LightGBM, and RF algorithms, combines the strengths of each algorithm [61]. SHAP values averaged across these models, thereby obtaining more generalized and stable interpretations of SHAP impacts on EUI predictions.

2.4 Diverse occupant behaviour and heterogeneous urban building stock modelling

As illustrated in Figure 9, the DOB-HUBS (Diverse Occupant Behaviour and Heterogeneous Urban Building Stocks) framework includes five main steps. 1) Acquiring heterogeneous urban building stock data. 2) Representative sampling to select set of samples from the urban big dataset while preserving behaviour diversity and urban stock heterogeneity. This involved unsupervised K-Means clustering to group similar buildings, and proportional stratified sampling to allocate samples from each cluster. 3) Automated bottom-up physics-based simulations. 4) Large-scale performance prediction using an ensemble machine learning model, detailed in 2.3.2. 5) Impact assessment using performance indicators including EUI, peak load reduction, and CO₂ emissions.

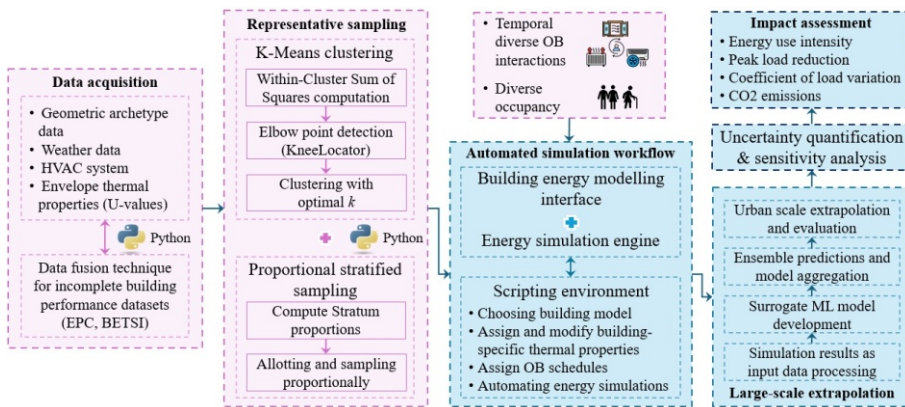


Figure 9. Overview of DOB-HUBS framework (from Paper 3).

This framework is demonstrated through an empirical study of single-family residential building stock in Umeå, Sweden. The study evaluates: (1) seasonal impacts of excessive window opening (winter, spring, summer, autumn) and (2) behavioural responses to the upcoming 2027 capacity-based electricity tariff under three distinct behavioural change scenarios (minor, moderate, significant). A

A dataset of 4,014 single-family buildings obtained through a data fusion approach is used. A typical meteorological year (TMY) from Meteonorm, representative hourly data are used as weather input. The building archetypes and geometry are obtained from the TABULA building typologies [59].

2.4.1 Representative sampling

Since the urban stock dataset contains 4014 building samples, running physics-based simulations for the entire dataset would be computationally intensive. To address this, simulations are conducted for a representative subset while preserving behavioural diversity and urban heterogeneity. This is achieved using a two-stage approach combining K-Means clustering and proportional stratified sampling [66]. In the first stage, K-Means clustering is used to cluster all buildings based on influential parameters including construction year, size, envelope U-values, HVAC system type, and occupancy patterns. In the second stage, proportional stratified sampling selects representative buildings from each cluster in proportion to its size in the full dataset. This ensures the relative proportion of building types is preserved [67]. This is implemented in Python using *scikit-learn* and *pandas* libraries [68]. For these 24 clusters, a representative subset of 1,500 buildings is selected by allocating samples according to cluster-specific proportions [69].

2.4.2 Automated simulation workflow

This section details the automated workflow for conducting physics-based energy simulations for selected 1500 representative buildings. The workflow is designed to dynamically assign building-specific parameters, including occupancy, OB interaction schedules, and envelope thermal properties (U-values of windows, walls, roofs and floors). Simulations are performed using EnergyPlus v23.2.0, accessed through the OpenStudio v3.7.0 software development kit (SDK). A custom Ruby script is developed to manage the entire process, including model development, parameter modification, simulation execution, and output processing, leveraging the OpenStudio SDK's Ruby bindings and API [70,71]. Ruby is used as the primary programming language due to its deep integration with the OpenStudio environment.

Each representative building is simulated under all defined scenarios, with building-specific inputs provided through a scenario file. A predefined set of archetype building models are developed in OpenStudio to represent single-family building stocks in Umeå. These

are defined based on construction period, floor area, and HVAC system configurations. A total of 18 archetype OpenStudio models are developed incorporating geometry, thermal zoning, construction materials and HVAC system characteristics for Swedish conditions following TABULA reference database.

2.4.2.1 Occupancy and interactions modelling

To incorporate diverse occupancy and OB interactions, a comprehensive set of custom *ScheduleRuleset* objects are modelled within a master OpenStudio model. These represent occupancy, electric equipment usage, and window opening behaviours for single-family building stock in Umeå. These input schedules are developed and implementation for two objectives of this study.

To assess the impact of 2027 capacity-based electricity tariff regulation during summer season, four different scenarios are modelled i.e., a baseline scenario representing before new regulation and three behavioural change scenarios (minor, moderate, and significant), as shown in Figure 10. The baseline is derived based on typical usage in Sweden from previous literature [72]. The behavioural change scenarios are developed based on assumptions reflecting increasing levels of occupant awareness and demand-shifting potential to maintain stable energy demand daily profile. These scenarios are implemented by modifying electric equipment usage schedules in OpenStudio.

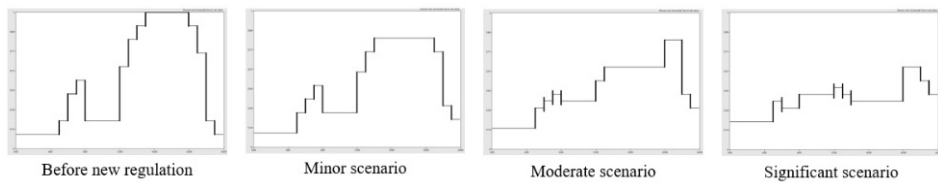


Figure 10. Electric equipment schedules across behavioural change scenarios [72] (from Paper 3).

Obtaining detailed real occupant-related data such as occupancy and electric equipment usage schedules, setpoint temperatures for each building at city-scale is challenging. In this study, these parameters are incorporated using municipal open census data [73] i.e., population distributions of working, retired etc., at the neighbourhood level. Five representative occupancy profiles in summer are defined based on information from multiple sources [74–77] as shown in Figure 11. Due to lack of building-specific occupancy data, the schedules are assigned

using population age distributions across different neighbourhoods in Umeå [73]. For occupancy profiles who are away from home, no electric equipment usage is modelled, and windows are always closed. The desired indoor temperature for summer is set to 26°C, which is recommended design temperature in Sweden [78].

Occupancy profiles	Population proportion	Windows operation
Working day	10%	Open
Holiday and stay at home	30%	Open
Holiday and away from home	40%	Closed
Retired and stay at home	10%	Open
Retired and away from home	10%	Closed



Figure 11. Occupancy fractional schedules [72] (from Paper 3).

To quantify the impacts of excessive window opening behaviour under four seasonal scenarios (winter, spring, summer, autumn), a baseline set of window opening schedules are estimated based on knowledge from previous literature on window operations in Sweden [79,80] complemented by assumption to address data gaps, as shown in Figure 12. Hourly opening fractions are assigned using the *ZoneVentilation:WindandStackOpenArea* object in OpenStudio. During seasonal simulations, the excessive opening schedule is applied only to the target season, while baseline schedules are used for the remaining periods. The defined seasonal period ranges are shown in Figure 12. The desired indoor temperatures are set to 18°C during non-summer periods, and for summer period, 26°C is defined.

Season	Range
1 Winter	November 16 - March 15
2 Spring	March 16 - May 31
3 Summer	June 1 - August 31
4 Autumn	September 1 - November 15

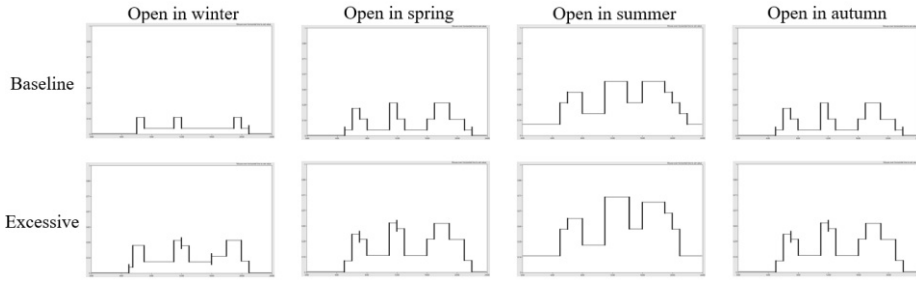


Figure 12. Window interaction schedules across seasons (from Paper 3).

All schedules (occupancy, electric equipment, and window operations) defined in the master OpenStudio model are programmatically copied to the full set of archetype building models using a batch Ruby script. The `schedule.clone()` method from OpenStudio SDK is used to replicate each `ScheduleRuleset` object across the models.

2.4.2.2 Thermal properties customization

In order to programmatically modify the material properties of walls, roofs, and floors with custom values, the Ruby script accesses `set.defaultExteriorSurfaceConstructions`, `set.defaultInteriorSurfaceConstructions`, and `set.defaultGroundContactSurfaceConstructions` within the respective OpenStudio model's `DefaultConstructionSet` object [81]. The thermal conductivity (k) for a material is recalculated based on building's U-values (U) and material thickness (d) using the standard equation,

$$k = U \times d \quad (1)$$

This is elaborated in detail in Paper 3 methodology section.

2.4.2.3 Simulation execution, Output processing and aggregation

The customized OpenStudio model for each building is translated to IDF file using a `ForwardTranslator` to perform EnergyPlus simulations for defined time periods and timesteps. Parallel processing is enabled for improve computational efficiency, especially for handling a large number of buildings with sub-hourly timesteps. The simulation results are extracted at both hourly and annual temporal scales for each building. A batch data aggregation Ruby script is used to programmatically

aggregate the results to the overall building stock level at every timestep [82].

2.4.3 Impact assessment

To estimate the environmental impact, the CO₂ emissions associated with the energy use of urban building stock is calculated. Due to lack of detailed energy source data for individual buildings, an aggregated estimation approach based on residential sector data in Umeå is used. The total energy use for the entire stock is obtained from the current energy balance of Umeå municipality [83]. This data provides total energy use and the corresponding shares of district heating, grid electricity, biofuels, and oil. Emission factors for each energy source are obtained from local authority data, the Swedish Environmental Protection Agency, and previous studies [84–86]. To estimate the total CO₂ emissions of the building stock, the total energy use is multiplied by the weighted average emission factor $EF_{weighted}$ which is calculated using the following equation,

$$EF_{weighted} = \sum_{i=0}^n (S_i \times EF_i) \quad (5)$$

where S_i is share of energy source in the energy mix, EF_i is emission factor for energy source i , and n is the number of energy sources.

The peak load reduction at different temporal scales is calculated using the standard formula [87],

$$\begin{aligned} & \text{Peak load reduction (\%)} \\ & = \left(\frac{P_{baseline} - P_{scenario}}{P_{baseline}} \right) \times 100 \end{aligned} \quad (6)$$

where $P_{baseline}$ and $P_{scenario}$ represent the peak power demand in the baseline and subjective scenario respectively.

2.5 Benchmarking

This method aims to provide benchmarking advice on user's building energy performance by comparing it with similar buildings in the surrounding neighbourhood. The specific question is: *How does the subjected user building performance compared to similar neighbourhood buildings?*

The peer groups are classified based on the influential factors: building location municipality, construction year of the building, number of families living in the building and total floor area of the building excluding the basement area. The influential factors are chosen based on previous literature [88–90], EPCs data availability and the assumed users' knowledge to input information.

This focuses on one and two family buildings, which is a common building type in Sweden. After data cleaning, 12 624 EPCs dataset from the Västerbotten region is utilized. These samples are classified by municipality (15 municipalities), construction year (until 1960, 1961-1980, and after 1980), number of families (one or two), and total floor area (0-100 m², 101-200 m², and more than 200 m²). This results in 18 reference peer groups per municipality and a total of 270 reference peer groups in the region.

These reference peer groups of buildings along with the user's actual EUI value shall be used to answer the benchmark question. The average reference EUI value of each peer group is calculated using the following equation,

$$EUI_{avg} = \frac{\text{average energy use of the specific reference group (kWh)}}{\text{average floor area of specific reference group (m}^2\text{)}} \quad (1)$$

User inputs include building location municipality, construction year, number of families, total heated floor area, and annual energy use from electricity and other energy sources. These are used to compute the user's actual EUI and benchmark it against the corresponding peer group.

2.6 Urban interactive visualization

This section focuses on developing Interactive Visual Analytics (IVA) based 3D urban interactive visualization platform to enhance public engagement in urban energy transition. This is demonstrated in Umeå city focused on providing energy, environmental and economic insights for residential building users. The scenarios focused on current building performance, retrofit potentials of building envelope components aligning with latest building code requirements, and potential impacts of excessive window-opening behaviours. This implementation workflow is illustrated in Figure 13. This involves four key stages: 1) data collection and processing, 2) energy modelling and impact assessments, 3) fuzzy string matching, 4) spatial matching.

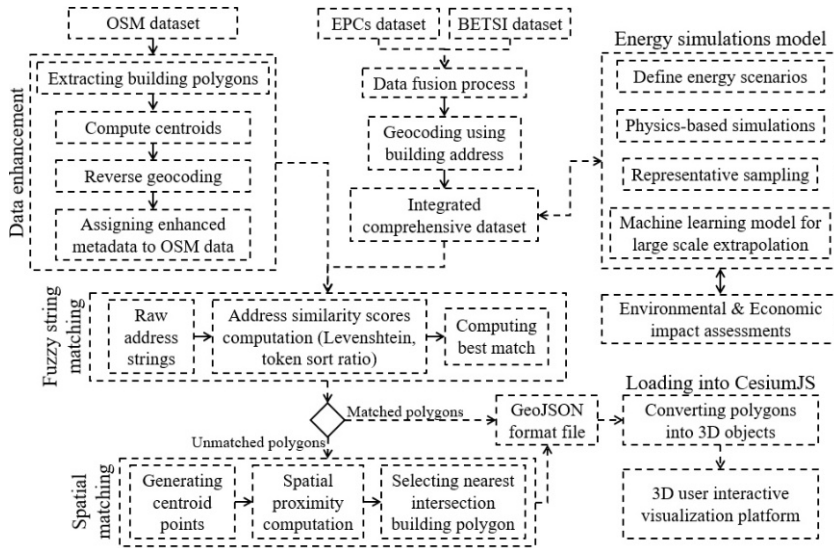


Figure 13. Interactive Visual Analytics (IVA) based platform development workflow (from Paper 5).

This study primarily relies on three open-source datasets: OpenStreetMap (OSM) dataset for 3D model extraction, EPCs and BETSI datasets for detailed building properties information. OSM is a widely used geospatial data source that provides spatial information such as building footprint polygons, land use, and associated metadata including building application and address information. In this study, OSM data is extracted using the Overpass Turbo, which is the data mining tool for accessing OSM [91]. OSM has been extensively adopted in urban scale modelling due to its global coverage and open-source. It is also used as main input data (e.g. building geometries) for energy modelling [92]. However, the data quality and availability of its attribute data, particularly building level metadata such as addresses, remain spatially inconsistent. Previous studies have shown that while OSM is often reliable, some key data such as building addresses might be incomplete, especially outside major urban areas including Umeå city in this study context [93]. This limitation can lead to misrepresentation of building level energy performance when used directly in applications. To avoid this, this study computes the centroids of all building footprint polygons extracted from OSM, and these centroid points are treated as the representative building coordinates (latitude and longitude) and subsequently used for standard reverse geocoding to retrieve complete address information. This is performed using the HERE Geocoding and Search API. The extracted full building addresses are replaced with

default OSM address metadata. In addition, EPC and BETSI are integrated into a comprehensive energy performance dataset using data fusion, as detailed in Paper 1. Using the addresses of all buildings in this dataset, the point coordinates (latitude and longitude) are computed using standard geocoding.

2.6.1 Fuzzy matching

This section involves linking the individual buildings from the integrated building performance dataset with corresponding building polygons extracted from OSM model. Since both datasets are independent sources with different data collection methodologies, direct matching using unique identifiers (e.g., building addresses) is often challenging due to variations such as spellings, abbreviations, or missing elements. For example, “*Gnejsvagen 12*” vs “*Gnejsvagen 12B*”, “*13 Gnejsvagen*” vs “*gnejsvagen 13*”. Due to these variations, exact string matching would fail for many valid buildings. Addressing these challenges would be imperative to achieve most number of valid building matches for visualization. To address these challenges, this study employs fuzzy string matching to compute the similarity between textual strings. This matching process is implemented using the *RapidFuzz* Python library [94], and employed two widely used complementary string similarity techniques i.e., Levenshtein ratio (*fuzzy.ratio*) and Token sort ratio (*fuzz.token_sort_ratio*).

The Levenshtein ratio technique measures the number of characters need to be changed, inserted, or deleted to transform one address into another [95,96]. For example, “*gnejsvagen 23*” and “*gnejsvagen 23a*” are very similar, “*gnejsvagen 23*” and “*gnejsvagen 8*” are very different. The similarity scores range from 0 to 100, where 100 indicate two identical strings. This technique is effective for detecting minor spelling variations or minor text differences. The Token sort ratio technique focuses on word order independence. It divides the address into individual words referred as tokens, sorts them alphabetically, and then compares the strings [96]. For example, the addresses “*32 Gnejsvagen*” and “*Gnejsvagen 32*” are treated as equivalent despite different word order. This technique is particularly useful when addresses appear in different sequences across the datasets.

For each building in the energy performance dataset: firstly, the address is compared against all possible address variants from the OST dataset. Secondly, both similarity techniques are applied to compute the individual similarity scores. And finally, the technique with highest

similarity score is selected as the best match. However, it is important to ensure to avoid any invalid matches when linking energy performance indicators (e.g., energy use, retrofit potential) to physical buildings. To ensure high reliability, only matches with perfect similarity score (100) are accepted. This strict threshold ensures that buildings are only linked if they are valid matches.

As a result, this process obtained 4839 of 7690 perfectly matched buildings in Umeå. These matched buildings are streamlined and compiled into a GeoJSON dataset file, which is a commonly used format for geospatial applications [97]. The unmatched buildings are possibly due to the data quality or errors with data collection process of EPC and BETSI, or also during reverse geocoding process.

2.6.2 Spatial matching

The unmatched buildings from fuzzy string matching section are handled in this section to further increase the number of matches by employing spatial matching approach as a secondary strategy. This approach focuses on geographical proximity rather than the textual similarity, and is useful when address attributes are inconsistent. This process utilizes the geographical coordinates (latitude and longitude) associated with each unmatched building in the integrated building performance dataset. These are derived during geocoding process and represent the approximate centroid location of the building. Similarly, building footprints from the OSM dataset are represented as building polygons. To ensure exact distance calculations, both datasets are reprojected into a common projected coordinate reference system (EPSG:32633, UTM Zone 33N) for Umeå city [98], which allows measuring distances in meters.

Considering any minor inaccuracies from geocoding or centroid approximation, each unmatched building point is buffered by 5 meters radius. A spatial intersection is used to determine whether each unmatched building point spatially overlaps with a building footprint polygon from OSM. This buffer enables a localized search area within which potential building polygon matches are identified. This spatial intersection is performed using *GeoPandas* Python library [99]. In cases where a buffered point intersects with multiple building polygons, this is addressed by computing the Euclidean distance between the centroid of each candidate polygon and the unmatched building point. The polygon with the shortest distance is selected as the most probable match. This approach assumes that the closest polygon represents the actual physical

building, and it is particularly helpful in high density urban built environments. Once a spatial match is identified, the corresponding building identifier from the OSM dataset is assigned to the building record from the building performance dataset. And all the associated energy, environmental and economic insights are then linked to the matched building polygon.

Through this process, it obtained additional 452 building matches were found that could not be linked in fuzzy string matching stage, which makes it 5291 matches in total from both stages. The results from both stages are compiled into a unified GeoJSON dataset containing building geometries with associated building performance insights across all defined scenarios.

2.6.3 3D visualization platform

The final GeoJSON dataset is used to generate 3D urban interactive visualization platform. The building polygons are converted into 3D objects by extruding building footprints based on geometric and attribute information for spatial representation. This is implemented using CesiumJS [100], an open-source JavaScript library [101] designed for creating high-performance, web-based 3D geospatial applications. CesiumJS is primarily used for visualization of large-scale geospatial datasets and enables real-time rendering of 3D objects directly within a web browser. It is widely used for digital twins, urban analytics and decision-support systems [100,102]. Within the CesiumJS environment, the obtained GeoJSON dataset is loaded and rendered as interactive 3D building models. Each building is linked to its corresponding energy, environmental, and economic attributes, and subsequent colours are assigned based on the performance. This allows users to interactively explore building-level information through visual inspection and scenario-based comparisons.

3 Results and discussion

This section presents the summary of main results of the thesis corresponding to the four research objectives and the appended papers, as detailed in Table 3. Subsection 3.1 presents the results of Paper 1 addressing RO1 by demonstrating data fusion framework through an empirical study in Umeå residential building stock. Subsection 3.2 addresses RO2 by demonstrating Paper 2’s framework, applied to 81 building stock clusters across three Swedish municipalities representing distinct climate zones. Subsection 3.3 focuses on addressing RO3 through DOB-HUBS framework introduced in Paper 3, which is demonstrated by assessing seasonal window-opening and behavioural response to upcoming 2027 tariff policy. Finally, subsection 3.4 addressed RO4 by summarizing results from Paper 4 and 5, focusing on public engagement through building performance benchmarking and 3D interactive visualization for Umeå residential stock.

Table 3. Structure of main results section.

Sub-section	Research objective	Appended paper	Focus	Method sub-section
3.1	RO1	Paper 1	Demonstrating data fusion framework using an empirical study in Umeå building stock.	2.2
3.2	RO2	Paper 2	Applied to 81 building stock clusters across three distinct Swedish municipalities from climate zones.	2.3
3.3	RO3	Paper 3	Applied to Umeå stock, assessing seasonal window-opening, and behavioural response to 2027 policy.	2.4
3.4	RO4	Paper 4 & 5	Public engagement through 1) benchmarking building performance, 2) interactive visualization.	2.5 & 2.6

3.1 Applied data fusion

To address the RO1, the proposed data fusion framework is applied to integrate two incomplete building performance datasets (EPC and BETSI) in Sweden for modelling the building stock of the Umeå city. Given the subarctic climate conditions and ageing buildings in Umeå, retrofiting importance is evident.

3.1.1 Heterogeneous building performance data

Firstly, probabilistic record linkage (PRL) is applied to integrate building records from the EPC and BETSI datasets into a single comprehensive dataset, as shown in Figure 14. These datasets consist of different useful information needed for building stock modelling and share some common variables data i.e., location, archetype, construction year, and floor areas, as detailed in section 2.1. However, the specific objective of probabilistic record linkage is to identify, for each EPC record, the most probable corresponding record in the BETSI dataset, using the common variables between them. The candidate pairs or matches are filtered based on the predefined thresholds for common variables, as shown in Table 4.

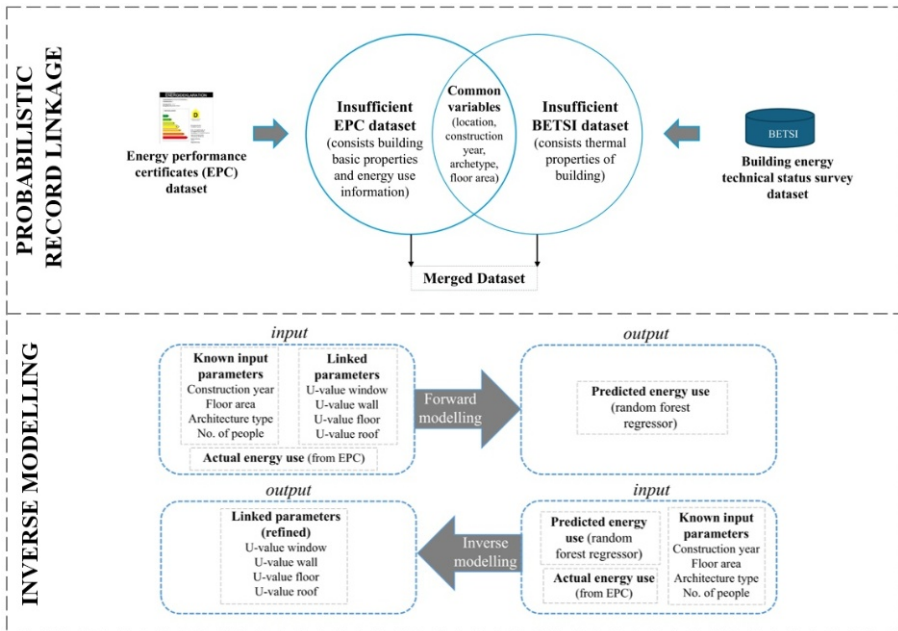


Figure 14. Applied data fusion to Umeå building stock (from Paper 1).

Table 4. Defined thresholds for common variables (from Paper 1).

Probabilistic record linkage	Match criteria
<ul style="list-style-type: none"> • Jaro-Winkler fuzzy string similarity (for textual attributes) • Absolute value difference function (for numerical attributes) 	<ul style="list-style-type: none"> • Building architecture type: threshold=0.85 • Construction year: threshold=10, Heated floor area: threshold=70

This comprehensive dataset, achieved from record linkage, includes ‘known input parameters’ such as construction year, floor area, architecture type, and occupancy, which already exist in the EPC dataset, as well as ‘linked parameters’ like U-values of building components. These values are predicted (matched from BETSI) through record linkage. Combined with actual energy use data from the EPC dataset, a Random Forest Regression model [104] has been trained to predict energy use for all records. The coefficient of determination (R^2) of this model for trained and testing datasets are 99.38% and 97.29% respectively. Random Forest is a supervised machine learning prediction model that is used for classification and regression trees (CART) problems. It combines several decision trees to improve the accuracy. This is a typical FM process. In contrast, IM adjusts/refines any unrealistic parameters in the ‘linked parameters’ using ‘known input parameters’, ‘actual energy use’, and ‘predicted energy use’, as explained in section 2.2. In sampling stage, new synthetic set of parameters for linked parameters from PRL are considered. These are proposed by metropolis random walk algorithm, which is based on posterior distribution [56,57] to create exploratory space for input parameters.

Figure 15 provides a comparison between actual and predicted energy use for the adjusted records. The blue dots represent the predictions made using ‘linked parameters’ before adjustments. The red dots represent predictions made using ‘linked parameters (adjusted)’ after adjustments. The black diagonal line represents perfect prediction where the actual and predicted energy use are equal, indicating the adjustments are accepted by the metropolis algorithm. However, the results clearly indicates that the proposed new set of parameters are only accepted at the acceptance stage, if the respective predicted energy use aligns with its actual energy use.

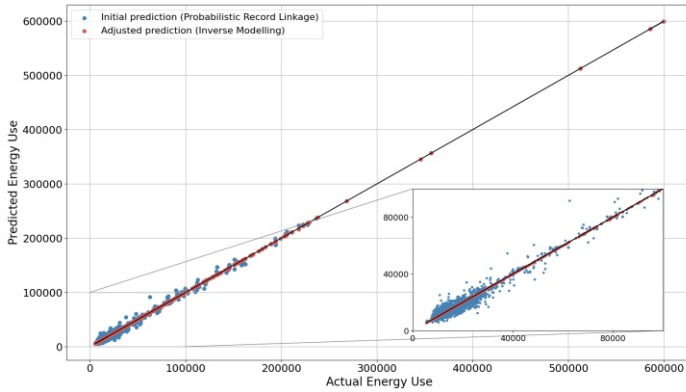


Figure 15. Actual energy use vs predicted energy use for adjusted records (from Paper 1).

3.1.2 Thermal performance of residential building stocks

As a result of data fusion, the integrated comprehensive dataset obtained matches for 7512 residential buildings including U-values of building envelope components. These U-values are compared against the Swedish national statistical U-values (TABULA reference values) [59], as illustrated in Figure 16. These stock level U-values align well with the overall trends observed in the reference data.

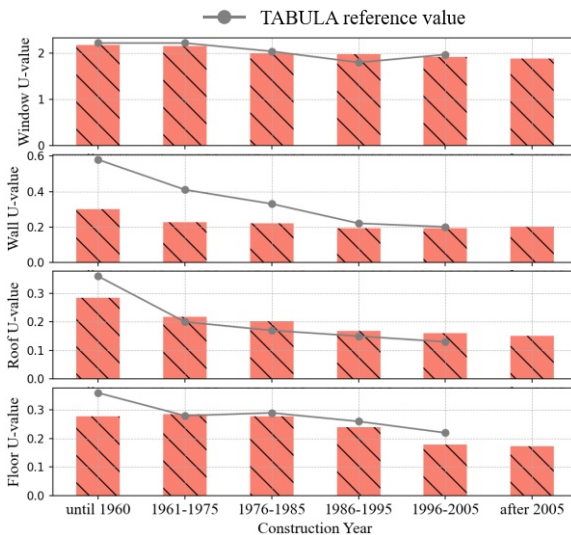
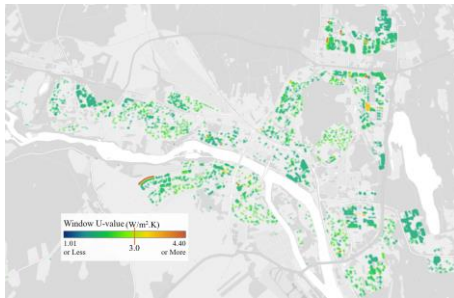
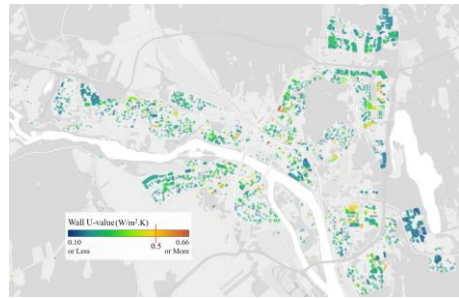


Figure 16. U-values ($W/m^2 \cdot K$) of single-family building stocks in Umeå.

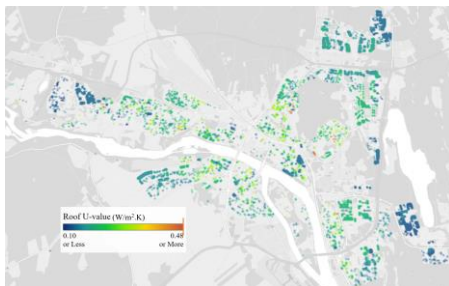
Figure 17 shows the U-value distributions of building envelope components across the Umeå building stock. High U-values, shown in red or orange, indicate poorly insulated areas, while low U-values, shown in blue or green, indicate well-insulated areas. This map allows comparisons at various levels, from small neighbourhoods to entire districts. The vertical lines in legend scales represent the maximum allowed U-values according to the Boverket building code SBN 1980 [105]. The Boverket building codes has been regarded as important guidelines for construction practices for ensuring energy efficiency and thermal performance of buildings.



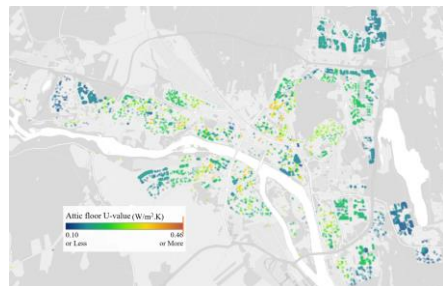
(a)



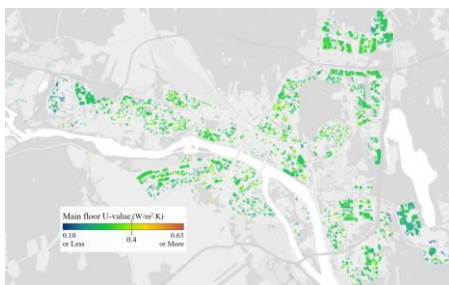
(b)



(c)



(d)



(e)

Figure 17. Distributions of U-values across various building components in Umeå building stock. a) Window, b) Wall, c) Roof, d) Attic floor, e) Main floor, (from Paper 1).

These results reveal significant spatial and component-level heterogeneity in envelope performance, both within and across neighbourhoods. Such variations are often overlooked in conventional bottom-up stock models that rely on statistical or homogeneous assumptions to address the missing data. In contrast, the proposed data fusion framework addresses this limitation by filling the data gaps in one dataset with information from another dataset. By capturing these variations, the framework allows the identification of specific building components and areas with the good and poor envelope thermal performance.

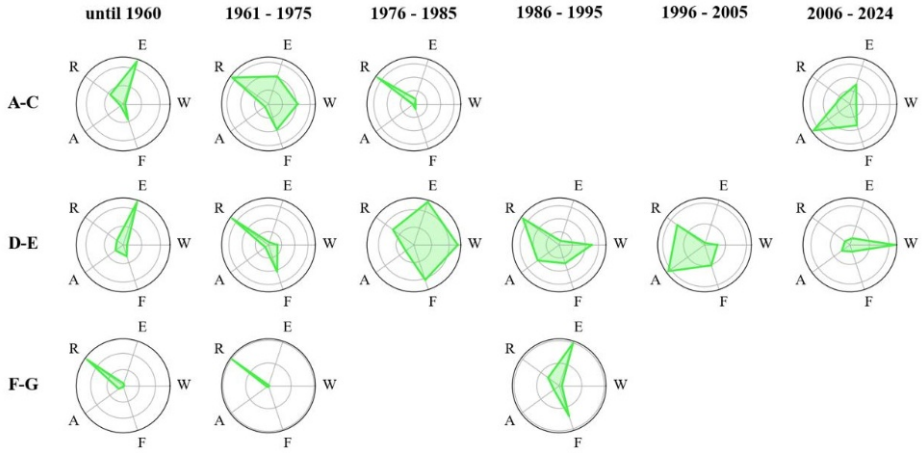
Overall, Paper 1 demonstrates how the proposed approach addresses the incomplete data challenge by capturing inherent heterogeneity within the building stock, aligning with RO1. This provides a foundation for subsequent research and enables future exploration of tailored retrofitting strategies to support effective energy and urban planning.

3.2 Tailored and impactful retrofits

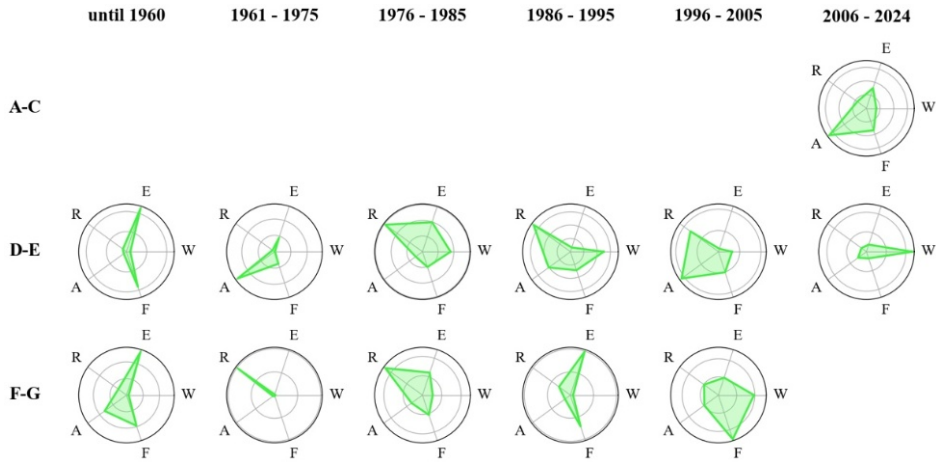
This section presents main results in correspondence to RO2, focused on identifying tailored and impactful building envelope retrofit strategies for 81 clustered building stocks using integrated data-driven framework of explainable AI (SHAP), ensemble machine learning (ML) and heterogenous data, as detailed in section 2.3. These are clustered based on building application type, construction year, and energy class, across three municipalities i.e., Linköping, Lund and Umeå. This study developed and evaluated model performance across 81 clusters individually, and the performance metrics are presented in Paper 2. Overall, the ensemble ML model has showed robust performance on EUI predictions with average R^2 value (0.87), MAE (5.57) and RMSE (8.54) across the clusters.

3.2.1 Insights across heterogenous clusters

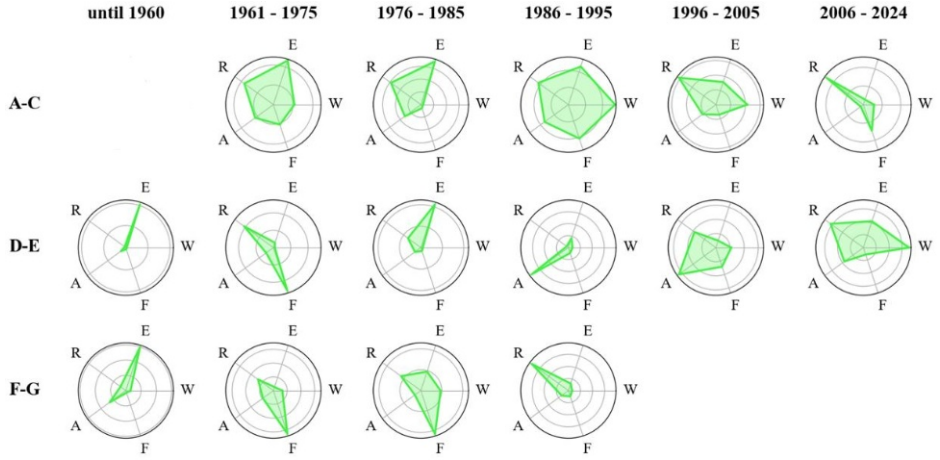
Figure 18 presents tailored and impactful building envelope retrofit insights for the 81 building stocks. Each radar plot represents a building stock, illustrating the most impactful building U-values (SHAP values of U-values) that contribute to model EUI predictions for that respective stock.



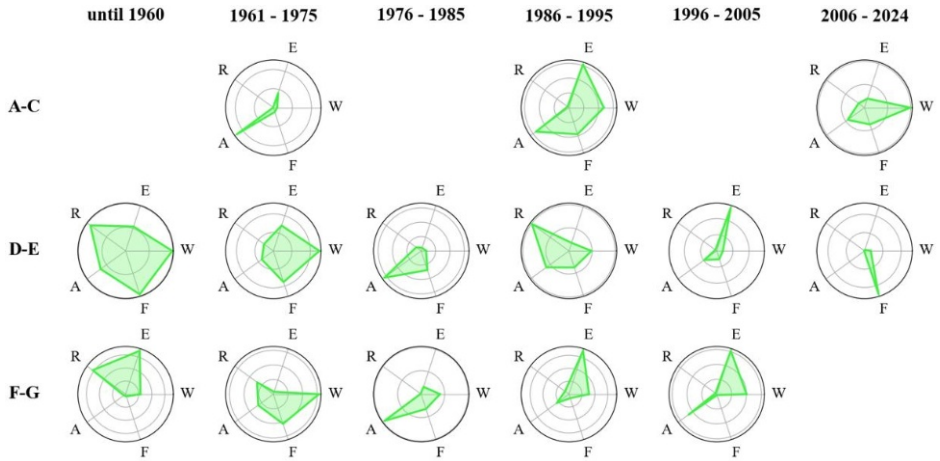
(a)



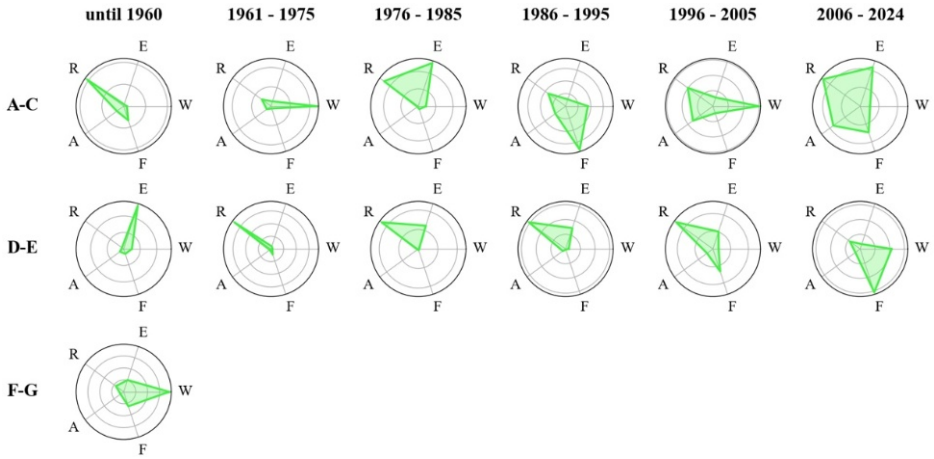
(b)



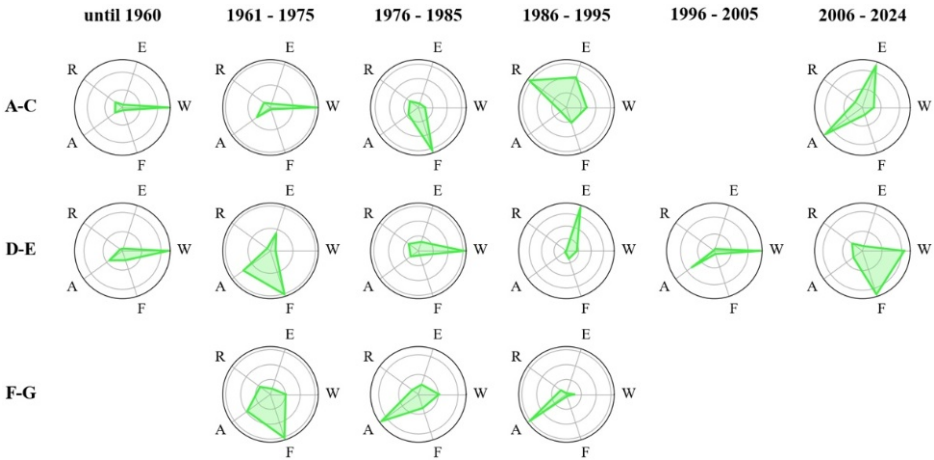
(c)



(d)



(e)



(f)

Figure 18. SHAP impact values of building U-values for 81 building stocks. W- SHAP impact value of window U-value, E- SHAP impact value of exterior wall U-value, R- SHAP impact of Roof U-value, A- SHAP impact value of attic floor U-value, F- SHAP impact value of main floor. a) Linköping, single-family stock, b) Linköping, multi-family stock, c) Lund, single-family stock, d) Lund, multi-family stock, e) Umeå, single-family stock, f) Umeå, multi-family stock, (from Paper 2).

In these radar plots, ‘W’, ‘E’, ‘R’, ‘A’, and ‘F’ represents the SHAP impacts of the window U-value, exterior wall U-value, roof U-value, attic floor U-value, and main floor U-value, respectively. For instance, a single-family stock in Linköping constructed before 1960 with an energy class of A-C,

shows that exterior walls have the higher impact, while attic floors have the lowest impact. Thus, it is recommended to prioritize retrofitting exterior walls to enhance the overall energy efficiency of this stock. These results reveal that the impact of building U-values on heating energy efficiency varies significantly across building stocks within the municipalities and also between climate zones. This variability reflects the contextual uniqueness of each building stock, driven by diversity in building characteristics, climate conditions, U-values, occupant behaviour, and the complex interactions among the building features [106]. Although multiple features contribute to overall heating energy performance, this study investigated the impact of building U-values, as they are direct retrofit measures and primarily affect the heat transfer through the building envelope. Consequently, the relative impact of individual building components (window, walls, roofs, floors) on heating EUI is highly context-specific and does not follow a uniform pattern across all building stocks. For instance, a well-insulated wall (low U-value) have high impact on energy efficiency in a cold climate but a much lower impact in a milder climate, as shown in Figure 18.

3.2.2 Stock-level post-retrofit assessments

In order to assess the stock-level energy performance outcomes following the retrofit insights, the post-retrofit heating energy use reductions are evaluated for six randomly selected building stocks across three municipalities, as illustrated in Figure 19. For each stock, retrofitting potential has been evaluated by comparing energy reductions for the most and least impactful U-values identified in the retrofit insights.

To predict stock-level energy performance under various retrofit scenarios, an XGBoost machine learning prediction model has been trained on 80/20 train-test data split. During training, the data samples from the same stock were excluded to ensure unbiased predictions. The specific retrofitting measures are designed based on Boverket’s building regulations (BBR, BFS 2017:5) code, shown in Table 5. For instance, the window retrofitting measure involves retrofitting all windows in the building stock, except buildings with already efficient windows (U-value=1.2 W/m². K).

Table 5. Stock-level retrofit measures based on Boverket building code (BBR, BFS 2017:5) [60].

Building feature	U-value (W/m ² . K)
------------------	--------------------------------

Window U-value	1.2
Wall U-value	0.18
Roof U-value	0.13
Floor U-value	0.15

Figure 19 presents the stock-level predicted heating energy use reductions for the six building stocks. In Linköping, pre-1960 single-family stocks with energy classes A-C shows that retrofitting wall has more energy impact than retrofitting window. However, for 1961-1975 single-family with energy classes A-C, retrofitting roof has more energy reduction than retrofitting wall. These trends align with the identified retrofit insights shown in Figure 18. Similarly, in Lund, for both pre-1960 (D-E) and 1961-1975 (A-C) single-family stocks, retrofitting wall yielded more energy reductions than retrofitting floor. In Umeå, for pre-1960 (A-C) single-family stock, retrofitting roof is more impactful than retrofitting window. Conversely, for 1961-1975 (A-C) stock, window retrofitting showed higher impact than retrofitting wall. On the other hand, for 1961-1975 (A-C) stocks, the impactful retrofitting U-values varies across three municipalities located in different climate zones. This variation is primarily due to significant longitudinal differences.

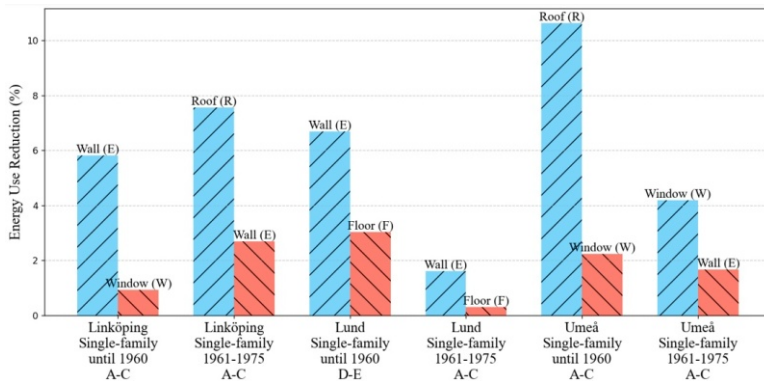


Figure 19. Stock-level heating energy use reductions of retrofitting high and low impactful building envelope U-values for chosen building stocks (from Paper 2).

Overall, these trends support the effectiveness of the identified retrofit insights and reinforce the importance of context-specific strategies. Given the inherent heterogeneity in each building stock, it is crucial to design tailored retrofitting strategies, especially when resources are limited and prioritized retrofitting is required. While retrofitting building envelope components solely based on high and low U-values is straight-forward approach, this alone may not guarantee significant

energy reduction. It is also important to assess the relative impact of different building components on overall energy performance, as demonstrated in this study.

3.2.3 Implications

Historically, Sweden's national building regulations have evolved significantly. As discussed in Paper 2, for instance, requirements shifted from individual building components limits (allowed U-values) before 1989 to performance-based limits on total energy use intensity [107]. This transition reflects shift from a one-size-fits-all insulation requirement towards an outcome-based approach, as long as the building's overall energy performance meets the target. In essence, post-1989 code enabled designers and building owners to choose flexible retrofitting strategies targeting most impactful measures for each building rather than uniform retrofitting strategies.

This flexibility is particularly important given the diversity of Sweden's building stock. Many buildings were constructed during Million Homes Program (1965-1975) and are now over 50 years old, presenting both challenges and opportunities for energy retrofitting [29]. For instance, one building may primarily lose heat through a poorly insulated attic, while another loss through old windows or uninsulated walls. Due to this heterogeneity, no single retrofit solution fits all buildings. Local governments recognize that meeting climate targets will require retrofit programs tailored to each building's specific conditions. However, identifying the most impactful retrofit measures for each building across an entire city is a complex task. With thousands of buildings of different ages, types and thermal properties, city planners face the challenges such as: which neighbourhoods or building types should be prioritized? What specific retrofits will yield the highest and fastest reductions in energy use?

This study aims to address these challenges by providing strategic data-driven insights. By understanding which retrofits are impactful for which buildings across a municipality building stock, stakeholders can optimise retrofit investments and improve overall energy efficiency.

From a practical governance perspective, XAI-based retrofit insights allow municipalities to allocate retrofit investments where they achieve the greatest impact. Rather than applying uniform measures, resources can be directed toward the most effective upgrades for specific buildings, supporting targeted policy outcomes under budget constraints. This is

what this research aims to support municipalities through identified strategic insights and introduced framework.

Previous building stock retrofit studies often oversimplified this heterogeneity by relying on homogeneous assumptions and strategies, limiting their ability to provide tailored retrofit insights. Such approach introduces significant uncertainties due to sensitive assumptions especially at larger stock levels [108], where small errors can be amplified. In contrast, this study's innovative integration of multiple data-driven techniques has enabled exploration while leveraging big data. As detailed in section 2.3.4, embedded reliability strategies ensure stable SHAP results and robust model performance, enabling effective and context-specific retrofit prioritization.

3.3 Behavioural response

To address RO3, this section presents main results of demonstrating introduced DOB-HUBS framework through empirical study in Umeå single-family building stock, as detailed in Paper 3. This study focuses on evaluating: (1) seasonal impacts of excessive window opening on energy and environmental performance (Winter, Spring, Summer, Autumn), and (2) behavioural responses to the upcoming 2027 capacity-based electricity tariff under three behavioural change scenarios (minor, moderate, significant).

3.3.1 Seasonal impacts of excessive window-opening

Figure 20 presents energy use intensity including both heating and cooling, with impacts ranging from very low (0-25 kWh/m²) to maximum (201-250 kWh/m²). Baseline scenario assumes standard window opening behaviour, shows most buildings performance is between medium and medium-high. The performance increased significantly in other excessive scenarios. Most significant impacts observed in winter scenario due to low outdoor temperatures during winters and increased heating needs to maintain indoor thermal comfort. Similarly, autumn and spring scenario show high increased intensity and less compared to winter scenario. This is despite the modelling low window operations in winter and high in autumn and spring. In contrast, summer scenario showed negligible increase. This is due to primary energy demand shifts from heating to cooling. Since high window operations are modelling in this scenario, the natural ventilation meets the increased cooling needs, leading to minimal increase in energy

use. However, the trend of varied impacts across seasons align with the previous literature [18,19].

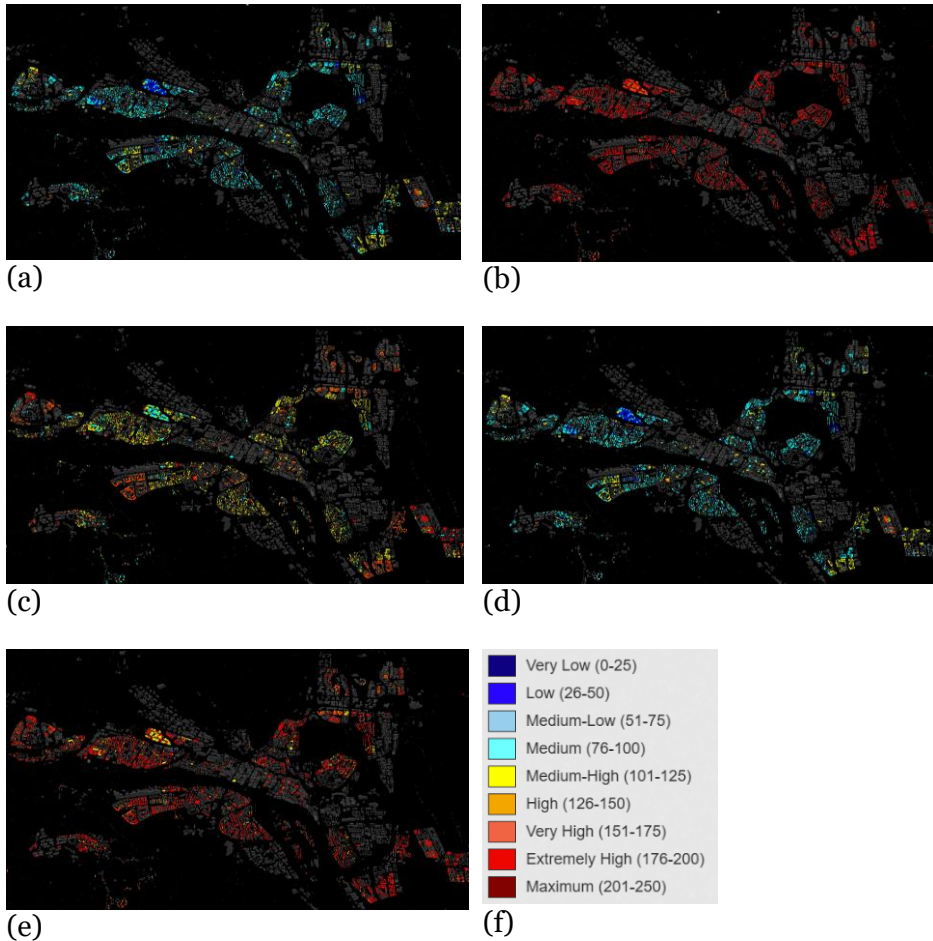
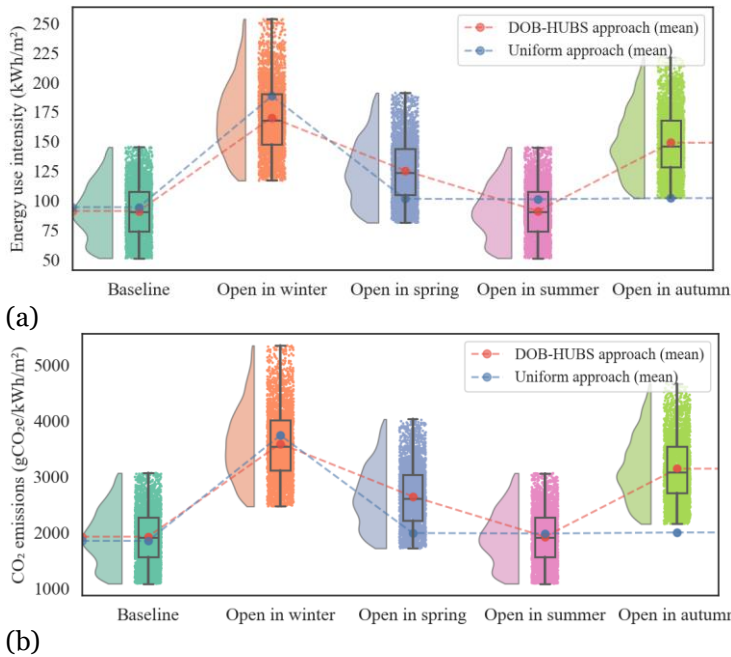


Figure 20. Urban level energy use intensity (kWh/m²/year) under excessive window-opening scenarios: a) baseline, b) open in winter, c) open in spring, d) open in summer, e) open in autumn, f) colour legend, (from Paper 3).

The distribution of these varied energy impacts with associated CO₂ emissions are shown in Figure 21. These variations emphasize the importance of moving beyond uniform OB and homogeneous modelling, to incorporate season specific variations into modelling. Conventional approaches often assume uniform OB schedules year-round and apply average U-values to all buildings. This simplification can misrepresent energy and environmental performance, as Figure 21 illustrates by comparing such uniform approach with the proposed approach.



(a)
 (b)
 Figure 21. Distribution of city-level energy and environmental impacts across excessive window-opening scenarios: a) energy use intensity ($kWh/m^2/year$), b) CO_2 emissions ($gCO_{2e}/kWh/m^2$), (from Paper 3).

Another key observation is the significant variation in impacts across buildings within the same seasonal scenario, highlighting the need to inherent heterogeneity in building stock modelling. Even under identical OB interactions, buildings can result in varied energy impacts [20]. For instance, poorly insulated buildings (high U-values) loose more heat during windows opening than better insulated buildings. Similarly, buildings occupied by retirees who remain home during the day generate different internal heat gains than those with working occupants. However conventional approaches often overlook this heterogeneity, leading to misrepresenting potentials.

3.3.2 Behavioural response under 2027 regulation

To evaluate the impacts of the 2027 capacity-based electricity tariff, three behavioural change scenarios were modelled: minor, moderate, and significant. These scenarios represent increased levels of behavioural response to the tariff policy, focusing on summer season to reduce modelling complexity and data constraints. Figure 22 presents stock level average hourly peak electricity load reduction potentials for a typical summer day. In minor change scenario, minimal reduction is observed throughout the day with a slight decrease in the afternoon

hours (15-16) and slight increase in late evening hours (22-24). The moderate change scenario shows nearly 10% reduction between 13-19 hours and showed rebound in late evening. The significant change scenario shows highest reduction particularly from 14-20 hours, peaking at 20% reduction at 16:00, with highest rebound during late evening hours. Overall, the results indicate significant peak load reduction potential under moderate and significant behavioural responses, reducing grid stress during critical late-afternoon hours.

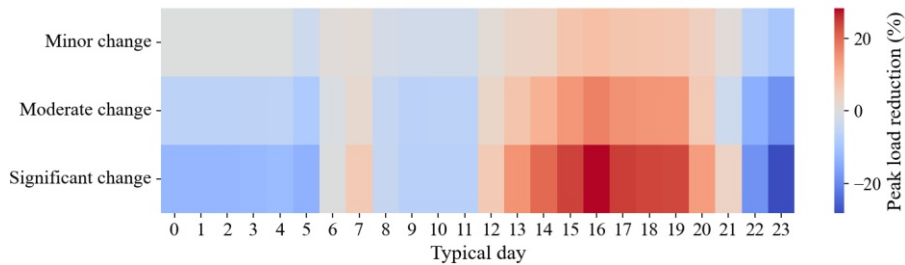


Figure 22. Hourly peak load reduction on a typical summer day under three behavioural change scenarios (from Paper 3).

Figure 23 presents impacts of behavioural changes across 24 heterogeneous building clusters, identified through representative sampling based on similar characteristics. Clusters 1, 5, 8, 13, 14, 15, 21 and 22 exhibit strong responses, with significant energy reductions under moderate and significant change scenarios. The remaining clusters exhibit minimal impacts due to summer vacation, consequent lack of electric equipment usage modelled. These results highlight the importance of incorporating OB diversity and urban building heterogeneity. Unlike traditional approaches that assume uniform OB across the stock, the proposed approach enables this analysis on how behavioural responses can vary across small building clusters within the stock, driven by OB diversity and inherent heterogeneity.

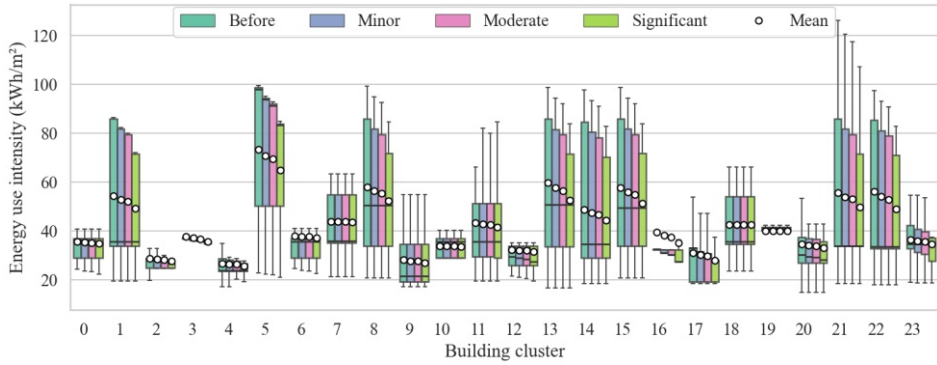


Figure 23. Energy use intensity impacts of three distinct behavioural changes across 24 heterogeneous building clusters within the urban stock (from Paper 3).

Furthermore, Figure 24 (a) presents the interaction between behavioural changes and building clusters defined by heterogeneous U-values. Figure 24 (b) shows the clustering based on average U-value distributions. Across all clusters, energy use intensity decreases consistently with increasing behavioural change. However, the magnitude of reduction varies significantly. High U-value clusters exhibit sudden and significant reductions, whereas low U-value clusters show gradual reductions.

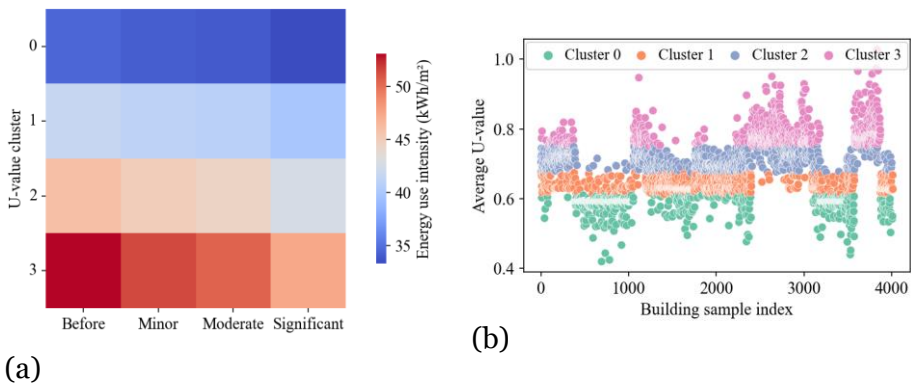


Figure 24. (a) Interaction impacts of behavioural changes across heterogeneous building clusters defined by U-value performance, (b) Clustering of building based on average U-value distributions, (from Paper 3).

These findings demonstrate that the impact of behavioural changes is highly influenced by building envelope U-value heterogeneity. Revealing such insights can enable designing tailored strategies to maximize policy effectiveness. The traditional UBEM approaches are limited to reveal these insights due its reliance on homogeneous U-values across all

buildings. By incorporating this heterogeneity, DOB-HUBS allows identifying small clusters that are more responsive to specific behavioural changes.

This research was validated by comparing simulated annual heating and cooling energy use with measured data from EPC dataset. The results show strong alignment between simulated and measured values at the stock level, with mean annual energy use of 14,279 kWh (simulated) and 13,811 kWh (measured). Individual and ensemble machine learning models demonstrate high predictive accuracy across all scenarios, with R^2 values generally exceeding 0.97 for both heating and cooling. Detailed validation methodology, performance metrics, and scenario-specific results are presented in Paper 3.

In addition, formal uncertainty quantification was conducted to assess the robustness of this framework predictions by propagating uncertainties from key input parameters, including occupant behaviour profiles and building envelope U-values obtained from data fusion, as well as their combined effect. Results indicate that occupant behaviour is the dominant source of uncertainty, with 90% confidence interval (CI) ranging from -4.73% to +6.90%. While envelope properties showed relatively less effect, with 90% CI between -2.38% and +0.09%. When both input uncertainties combined including their correlations, 90% CI resulted between -4.93% and +4.87%. The full uncertainty modelling framework, assumptions, and detailed results are described in detail in Paper 3.

3.4 Public engagement

To address RO4, this section presents main results related to public engagement in the urban energy transition. This is demonstrated through two aspects, 1) benchmarking of building performance, in section 3.4.1 and correspondence to Paper 4, 2) IVA based 3D urban visualization platform presents decision-making knowledge, allows users to explore current building performance, as well as the impacts of ‘what-if’ retrofit and occupant behavioural scenarios at their tailored cluster level. This is presented in section 3.4.2.

3.4.1 Benchmarking

With the user given inputs, the EUI is calculated and a user-specific group is identified from 270 reference peer groups. Then, the user building performance is evaluated against the buildings within the peer

group. The result is shown in multiple performance rating scale techniques, e.g., the percentile ranking of user building compared.

To demonstrate the model, a test case using sample user inputs is analysed and the resulting advice is presented. The test building is a multi-family house located in Umeå municipality, constructed in 1981, with a floor area of 52 m². The user input energy use is 6400 units (kWh/year) and approximately 30kg of firewood for external heating. Based on these inputs, the calculated user actual EUI is 128.5 kWh/m² and the relevant reference peer group is identified. The advice is presented as a percentile ranking that indicates the building's energy performance relative to its peer group.

Test user: *How is the user building performance compared to similar neighbourhood buildings?*

Advice: Illustrated in Figure 25.

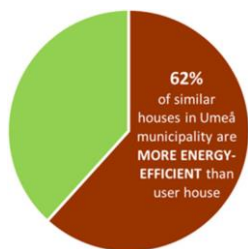


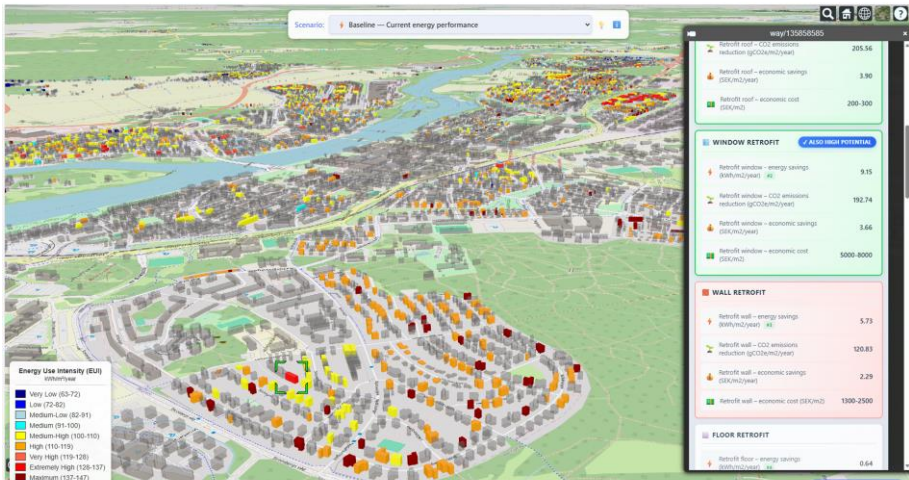
Figure 25. Benchmarking advice to the test user based on percentile ranking (from Paper 4).

From a public engagement perspective, this benchmarking result provides as an effective entry point for user awareness. This enables users to quickly understand their relative energy use and supports consideration of subsequent actions, such as retrofit measures or behavioural impacts.

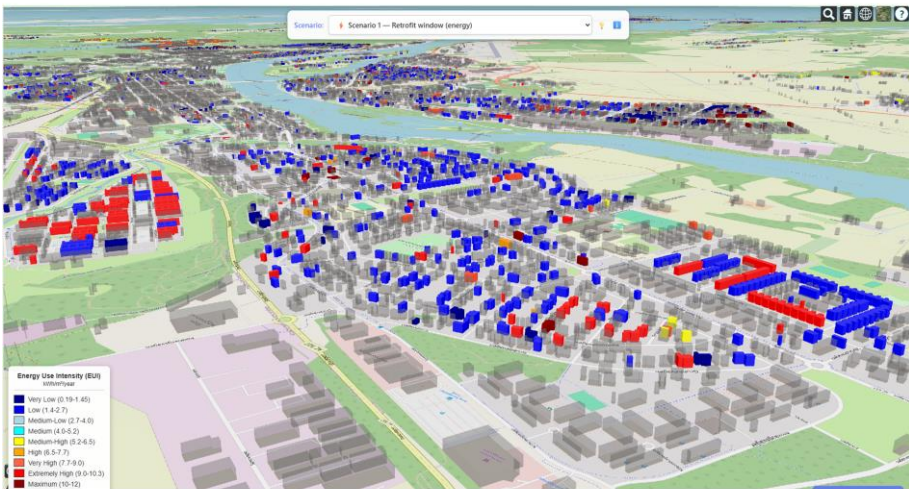
3.4.2 Interactive visual analytics (IVA) platform

Figure 26 shows the developed IVA 3D urban interactive visualization platform interface, which extends public engagement by enabling users to explore scenario-based outcomes and colouring buildings within the broader urban energy context, as detailed in section 2.6. As shown in Figure 26(a), users can select a building and view its key performance metrics, including potential impacts of retrofit measures, and behavioural impacts of seasonal window-opening, alongside comparative

benchmarks with respect to overall city performance. Figure 26(b) illustrates interpreting specific building performance relative to their neighbourhood buildings to reinforce peer-based understanding of the user, as detailed in Paper 5.



(a)



(b)

Figure 26. (a) Interactive platform interface showing energy-efficient measures and recommendations for the selected building, (b) Interactive platform interface comparing visualization against neighbourhood buildings. (from Paper 5).

These scenario-based insights and urban-scale visualization is enabled by integrating standardized open-source datasets (e.g., OSM, EPC, BETSI) and data-driven techniques (fuzzy string matching, spatial matching, geocoding and reverse geocoding) with bottom-up energy modelling, as detailed in section 2.6.

This developed platform was tested with real users in Umeå city, to quantify the impact of public engagement. This convergent evidence provides that IVA improved the user decision experience primarily by reducing processing costs, reflected in lowered self-reported effort, frustration as well as reduced physiological arousal.

This work translates complex energy performance modelling into accessible and interactive representations, thereby supporting in enhancing public awareness, participation, and informed decision-making within the BBSM context. The benchmarking and interactive visualization together demonstrate the potential of such approaches in addressing the limitation of existing BBSM research. While the majority of building stock level studies remain focused on technical modelling outcomes, this study demonstrates how such models can be extended to enhance public engagement in urban level energy transitions.

4 Implications

Overall, this thesis provides methodological and empirical insights that support evidence-based urban energy transition planning at both municipality and building occupant levels.

For municipal level stakeholders, including urban planners and policymakers, the proposed work supports:

- the identification of heterogeneous building clusters with high energy-saving potential, supporting targeted prioritization of retrofit investments and policy interventions.
- the evaluation of diverse occupant behavioural impacts, allowing the identification of building clusters that are particularly sensitive to behavioural change and the design of tailored engagement strategies in support of emerging regulations, such as the 2027 capacity-based electricity tariff.

And for building occupants, owners, and energy consumers, this work supports:

- benchmarking their building energy performance against comparable neighbourhood peer groups, reinforcing awareness through peer influence mechanism [8].
- evaluation of cluster-specific impacts of behavioural changes, enabling users to understand their potential contribution to ongoing energy transitions.
- identification of locally tailored and impactful retrofit measures to support informed decision-making when considering renovation investments.

Also, by integrating modelling output with user-oriented analytical tools, this thesis contributes to bridging the gap between technical assessment and practical decision-making in urban energy contexts.

5 Limitations and future work

While this thesis investigated how data fusion approach can enhance BBSM, its effectiveness could be compared with other approaches and improved further. The obtained data through data fusion is compared against the existing statistical data, and its associated uncertainty is quantified. However, future work could improve this process by incorporating building-specific real-world data, which remains challenging to collect but would strengthen the approach. Another limitation in exploring tailored and impactful retrofit insights is its narrowed focus on building envelope centric retrofit measures to improve heating energy efficiency. While these measures are generally considered important, further consideration of measures like smart controls, appliances upgrades, on-site renewables, efficient lighting etc., could discover comprehensive enhancement of overall energy efficiency. In addition, the modelling excludes other influential measures to building envelope such as thermal bridges, ventilation issues, heating system type due to lack of data. Future incorporation of these parameters could enhance comprehensive assessment of retrofit measures. Also, identifying the most impactful factors allows stakeholders to assess the impact of climate change scenarios, which supporting in planning long-term resilient strategies.

This work focused on residential building stocks across the empirical studies, future work should expand the analysis to include other building types, such as commercial and public buildings. Regarding occupant behaviour modelling, this study analysed only window opening interactions. Future research should also examine other occupant interactions, such as thermostat adjustments and air-conditioning usage, to enable comparisons between different behavioural patterns and their impact on building performance. Due to limited occupant-related information at the individual building level, assumptions were made to address data gaps in schedules, primarily based on typical occupant profiles reported in previous literature. This limitation could be addressed in future work through the use of high-resolution metered data to capture real-world occupant behaviour patterns [109]. Finally, the building geometries in this study were modelled using the TABULA reference database, which reflects typical construction standards over time in Sweden. This approach could be improved in future work by incorporating building-specific geometric data to reduce uncertainty.

6 Conclusion

Urban energy transitions require bottom-up building stock modelling (BBSM) that can represent heterogeneity, support context-specific decision-making, evaluate behavioural diversity for effective planning and public engagement. However, conventional BBSM is constrained by several limitations: incomplete data and homogeneous representation, generic retrofit prioritization, simplified behavioural assumptions, and limited integration of public engagement. These can lead to misrepresenting energy outcomes and policy effectiveness of energy transition initiatives.

This thesis presents an integrated, heterogeneity-aware methodological framework that addresses these limitations through a cumulative and interlinked techniques. By combining data fusion, machine learning, explainable AI, physics-based simulation, and interactive visual analytics, the proposed approach enhances BBSM capabilities. First, this thesis integrates multiple building performance datasets through data fusion framework based on probabilistic record linkage and inverse modelling. By filling the missing data in one dataset with information from another, instead of traditional reliance on average or statistical data. Thus, it overcomes the first limitation by capturing the inherent heterogeneity between the buildings within the same building stock, thereby providing robust foundation for subsequent analyses. Second limitation is addressed by introducing an integrated data-driven framework that combines ensemble machine learning, explainable AI and data fusion. Utilizing heterogeneous building performance data, it identifies the most impactful building envelope components for each local context. The results demonstrate that the retrofit measures to prioritize varies significantly across municipalities, climate zones, and building clusters. This shows traditional generic retrofitting can misrepresent energy potentials, highlighting the importance of local-specific, data-driven retrofit prioritization in effective retrofit planning. Third limitation is addressed by proposing DOB-HUBS framework that incorporates diverse occupant behaviour together with heterogeneous building properties into BBSM. The empirical results demonstrate that behavioural impacts vary significantly across seasons, with significant deviations compared to traditional approach. Results also show that behavioural responses to 2027 capacity-based electricity tariff can reduce peak electricity demand, with impacts varying across heterogeneous building clusters. Building on these modelling advances,

this thesis extended BBSM towards public engagement by developing data-driven benchmarking methods and an IVA based 3D urban interactive visualization platform. These translate complex modelling outputs into accessible, user-oriented insights that support peer comparison and ‘what-if’ scenario exploration. The empirical evidence from Umeå case study showed such IVA platform can improve user understanding by making complex information more approachable and less frustrating.

Overall, this thesis advances BBSM by capturing inherent heterogeneity, localized retrofit prioritization, evaluating behavioural diversity, and enhanced public engagement. The proposed approach supports urban planners and policymakers in designing more targeted and effective urban energy transition planning. These methods and evidence demonstrated in this thesis provide a foundation for future research focused on building stock decarbonisation.

Acknowledgements

This thesis is the result of immense support and contributions from many people, directly and indirectly, including family, supervisors, colleagues, and friends who made this long journey a meaningful and memorable one. I would like to take this opportunity to sincerely thank all of them.

Firstly, I'd like to thank my parents (**Subhash & Anusha**) for their unconditional sacrifices, love, limitless support, and patience. I am not equipped to repay even a fraction of what I have received from you. In fact, the process of trying to make you proud has become the strongest motivation pushing me forward in my career. I also express gratitude to universe for giving *all-in-one support system* in the form of my elder brother **Sarath**, who has always been there for me, and **Manasvi vadina**, warmly welcome to the family.

I am deeply thankful to my main supervisor **Weizhuo Lu** and co-supervisors **Kailun Feng, Thomas Olofsson, Anders Rebbling** for your consistent support and encouragement throughout the journey. **Weizhuo**, I'm grateful for the opportunity and for trusting me. Your guidance helped me expand my expertise and grow as a researcher. Thank you for always backing me and for your positive words during stressful times, which helped me overcome challenges with confidence. The freedom and encouragement you provided made this process deeply satisfying and fun. **Kailun**, the close collaboration with you, as a guide, as a friend helped me get hands-on to PhD significantly faster. Your ability to come up with ideas instantly and genuine commitment to support me is invaluable. I'm grateful for your fast-track support throughout the journey. **Thomas**, thanks for kindness and your lateral perspectives you always brought into discussions. Even our short conversations often had impact on my thinking and research process. **Anders**, your contribution and support especially within Auroral project is highly appreciated. Thank you for many encouraging and insightful conversations and for your practical feedback. I'd also like to thank my former reference person and mentor **Xingxing**, grateful for your support. I'm also grateful to my current reference person **Keni Ren** for her kind support. I thank all my colleagues at Applied Physics and Electronics Department for providing positive, friendly and fun working environment. I'm sincerely grateful to administrative colleagues

Christer (special thanks for warmly receiving me to Umeå), **Julia, Leif, Robert, Jörgen,** and **Helena.**

I would like to thank AURORAL project partners, internal team, and regional stakeholders for their contributions in ideating and navigating the research path. **Chanachok**, thanks for valuable contribution to the project, and for many good memories and as a friend, colleague and officemate. **Shoaib**, thanks for your contributions and your guidance during the initial phase of my PhD. I learned a lot about project management. Cheers to our travels, game evenings, and many fun memories. **Tobias**, it's always fun working with you in Auroral. Thanks for proofreading the Swedish translation text in thesis. Thanks for being a very good friend, I'll cherish our random lunches, coffee conversations. Also, thanks for initiating many group activities from food exploration to fun evenings. **Bokai**, thanks your contributions in critical moments of the project and for being such a kind colleague and friend. **Anjan**, thanks for being so kind and patiently listening to random talks, sharing insights, cooking good food, and game nights. **Moa**, thanks for bearing my brain dump conversations and for memorable adventures during summer school. I still can't believe you chose to walk under the shade in Zurich. **Gireesh**, thanks for checking up on me often, and for your thoughtful conversations and suggestions. **Chujie**, thanks for being fun officemate, the board games we played, the music we shared while working. **Pengju**, thanks for being very kind friend and colleague, for many laughs, and for opening the building door many times when I lock myself out. **Naresh**, thanks for desi conversations, sweets, and food. **Erik**, thanks for many fika conversations, and introducing best seafood into my life in Umeå. **Haitao** and **Jack**, thanks for your help with conducting experiments, and fun evenings. **Frej**, thanks for being kind officemate and for knowledge-sharing conversations. I would also like to thank colleagues who became wonderful friends. The Mafia game nights, Friday gatherings, and other activities helped a lot, especially during the stressful final phase of my PhD. **Fatemeh**, thank you for being such a kind friend. **Dinara**, thank you for the fun conversations. **Jonas, Elaheh, Vartan, Cui,** and **Rinad**, it has been wonderful spending time together. I'm grateful for nicest colleagues for meaningful interactions whether it was a lunch, a fika, or a quick talk or suggestion especially when my stress levels were high, so I thank **Antonio, Younes, Trong, Mark, Itai, Shuo, Ou.**

Thanks to *Telugu* acquaintances in Umeå **Yeswanth, Raviteja, Hema, Yamini, Aira** for being a relatable home away from home. Many fun activities and gatherings we did, especially food, has filled the part of the

void of being far from home. Yeswantha, I hope our journey *started in locus* continue forever, you're so fun together with Yamini. Ravi bro, thanks for your kind gestures, thoughtful efforts, and consideration you showed. Hema *garu*, you've been very kind, your fun-loving nature has made every gathering joyful. **Rangu**, thank you for bringing energy and joy in your bag every time you came from Skellefteå. Thanks to **Kwatra**, **Baura**, **Sourasekhar** for many good memories during early years of my PhD, from food, sports, celebrating Indian festivals, to Saloni's wedding. Those moments helped keep me going during the initial phase of my PhD journey. **Ayush**, thanks for being a kind friend, and for many kitchen conversations, and finally, *EE sala cup namdu*. I also thank **Shekhar** and **Yashwant** for fun memories. I'd like to acknowledge Borlänge friends for always giving me refreshing energy and for forgiving my absence during your special moments. **Juveria** thanks for bearing my cathartic and boring conversations. **Puneet**, thank you for consistently checking on me, especially during intense times. **Puneet & Sara**, I had great time at your special occasion. **Praneeth & Richa**, visiting you in Västerås has always been fun and memorable. **Samhita & Alex**, thanks for always being kind and very happy for you guys. Special thanks to long-term friends **Jayanth**, **Mahesh**, **Sankeerth** for occasional stress-busting group calls. I also thank many of family members and cousins for understanding my absence in their important occasions due to this marathon I embarked.

The unsung hero? IKSU sport, thanks for keeping me healthy and mentally awake. I'd also like to thank those I may have unintentionally missed. Your contribution directly or indirectly has been essential in this long journey.

References

- [1] Energy efficiency, (n.d.). <http://www.energimyndigheten.se/energieffektivisering/> (accessed June 11, 2021).
- [2] EBC || IEA EBC, (n.d.). <https://www.iea-ebc.org/ebc> (accessed June 8, 2021).
- [3] Energy-carbon-investment payback analysis of prefabricated envelope-cladding system for building energy renovation: Cases in Spain, the Netherlands, and Sweden, *Renewable and Sustainable Energy Reviews* 145 (2021) 111077. <https://doi.org/10.1016/j.rser.2021.111077>.
- [4] 1946–1975 Allmännyttan byggs ut och bostadsbristen byggs bort, *Sveriges Allmännytta* (n.d.). <https://www.allmannyttan.se/historia/historiska-epoker/1946-1975-allmannyttan-byggs-ut-och-bostadsbristen-byggs-bort/> (accessed June 8, 2021).
- [5] K. Feng, W. Lu, S.R. Penaka, E. Eklund, S. Andersson, T. Olofsson, Energy-efficient retrofitting with incomplete building information: a data-driven approach, *E3S Web Conf.* 356 (2022) 01003. <https://doi.org/10.1051/e3sconf/202235601003>.
- [6] Radikal implementering för klimatneutrala städer, *News Powered by Cision* (2023). <https://news.cision.com/se/viable-cities/r/radikal-implementering-for-klimatneutrala-stader,c3890493> (accessed October 31, 2024).
- [7] D. Maiullari, A. Palm, H. Wallbaum, L. Thuvander, Pathways towards carbon neutrality: A participatory analysis of the Gothenburg’s energy plan, *IOP Conf. Ser.: Earth Environ. Sci.* 1085 (2022) 012041. <https://doi.org/10.1088/1755-1315/1085/1/012041>.
- [8] K.S. Wolske, K.T. Gillingham, P.W. Schultz, Peer influence on household energy behaviours, *Nat Energy* 5 (2020) 202–212. <https://doi.org/10.1038/s41560-019-0541-9>.
- [9] M. Kavgic, A. Mavrogianni, D. Mumovic, A. Summerfield, Z. Stevanovic, M. Djurovic-Petrovic, A review of bottom-up building stock models for energy consumption in the residential sector, *Building and Environment* 45 (2010) 1683–1697. <https://doi.org/10.1016/j.buildenv.2010.01.021>.
- [10] P. Caputo, G. Pasetti, Boosting the energy renovation rate of the private building stock in Italy: Policies and innovative GIS-based tools, *Sustainable Cities and Society* 34 (2017) 394–404. <https://doi.org/10.1016/j.scs.2017.07.002>.

- [11] M. Österbring, C. Camarasa, C. Nägeli, L. Thuvander, H. Wallbaum, Prioritizing deep renovation for housing portfolios, *Energy and Buildings* 202 (2019) 109361. <https://doi.org/10.1016/j.enbuild.2019.109361>.
- [12] R. Andersen, L.B. Jensen, M. Ryberg, Using digitized public accessible building data to assess the renovation potential of existing building stock in a sustainable urban perspective, *Sustainable Cities and Society* 75 (2021) 103303. <https://doi.org/10.1016/j.scs.2021.103303>.
- [13] Boverket, Öppna data - Betsi är en undersökning om byggnaders tekniska status, Boverket (n.d.). <https://www.boverket.se/sv/om-boverket/publicerat-av-boverket/oppna-data/betsi-oppna-data/> (accessed May 26, 2023).
- [14] S.B. Sadineni, S. Madala, R.F. Boehm, Passive building energy savings: A review of building envelope components, *Renewable and Sustainable Energy Reviews* 15 (2011) 3617–3631. <https://doi.org/10.1016/j.rser.2011.07.014>.
- [15] P. Eriksson, T. Johansson, Towards Differentiated Energy Renovation Strategies for Heritage-Designated Multifamily Building Stocks, *Heritage* 4 (2021) 4318–4334. <https://doi.org/10.3390/heritage4040238>.
- [16] Boverket, Boverket (n.d.). <https://www.boverket.se/sv/om-boverket/publikationer/2011/energi-i-bebyggelsen---tekniska-egenskaper-och-berakningar/> (accessed May 2, 2025).
- [17] S. Dabirian, N. Rahimi, U. Eicker, Stochastic-based occupant-centric building archetype modelling using plug loads, in: *IBPSA*, 2023: pp. 298–305. <https://doi.org/10.26868/25222708.2023.1381>.
- [18] C. Du, W. Yu, Y. Ma, Q. Cai, B. Li, N. Li, W. Wang, R. Yao, A holistic investigation into the seasonal and temporal variations of window opening behavior in residential buildings in Chongqing, China, *Energy and Buildings* 231 (2021) 110522. <https://doi.org/10.1016/j.enbuild.2020.110522>.
- [19] F. Guo, Z. Chen, J. Wang, Effect of Dynamic Window Opening Behaviors on Indoor Thermal Environment and Energy Consumption in Residential Buildings of Different Chinese Thermal Climate Zones, *Buildings* 15 (2025) 28. <https://doi.org/10.3390/buildings15010028>.
- [20] P.R. Pandey, Y. Liu, N. Wilson, B. Dong, Dataset on occupant behavior, indoor environment, and energy use before and after dormitory retrofit, *Sci Data* 12 (2025) 798. <https://doi.org/10.1038/s41597-025-05166-7>.
- [21] H. Ghorbani Naeini, A. Norouzasas, F. Piraei, M. Kazemi, M. Kazemi, M. Hamdy, Impact of building envelope parameters on

- occupants' thermal comfort and energy use in courtyard houses, *Architectural Engineering and Design Management* 20 (2024) 1725–1751. <https://doi.org/10.1080/17452007.2023.2254294>.
- [22] I.K. Kasmeridis, N. Skandalos, T. Dimitriou, V.V. Dimakopoulos, D. Karamanis, Digitized Energy Systems and Open-Access Platforms: Accelerating Cities' Transition to Carbon Neutrality, *Urban Science* 9 (2025) 364. <https://doi.org/10.3390/urbansci9090364>.
- [23] E. Wilson, N. Merket, An Interactive Visualization Tool for Large-Scale Building Stock Modeling: Preprint, *Renewable Energy* (2022).
- [24] V. Coors, R. Padsala, Urban Digital Twins Empowering Energy Transition: Citizen-Driven Sustainable Urban Transformation towards Positive Energy Districts, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLVIII-4-W10-2024* (2024) 51–56. <https://doi.org/10.5194/isprs-archives-XLVIII-4-W10-2024-51-2024>.
- [25] J. Thiele, J. Wiehe, C. von Haaren, Participation 3.0 in the implementation of the energy transition—Components and effectiveness of an interactive dialogue tool (Vision:En 2040), *PLOS ONE* 19 (2024) e0299270. <https://doi.org/10.1371/journal.pone.0299270>.
- [26] Comprehensive investigation on energy retrofits in eleven multi-family buildings in Sweden, *Energy and Buildings* 84 (2014) 704–715. <https://doi.org/10.1016/j.enbuild.2014.08.044>.
- [27] Development and validation of energy signature method – Case study on a multi-family building in Sweden before and after deep renovation, *Energy and Buildings* 210 (2020) 109756. <https://doi.org/10.1016/j.enbuild.2020.109756>.
- [28] Z. Wang, Y. Wang, R.S. Srinivasan, A novel ensemble learning approach to support building energy use prediction, *Energy and Buildings* 159 (2018) 109–122. <https://doi.org/10.1016/j.enbuild.2017.10.085>.
- [29] K. Feng, W. Lu, Y. Wang, Q. Man, Energy-Efficient Retrofitting under Incomplete Information: A Data-Driven Approach and Empirical Study of Sweden, *Buildings* 12 (2022) 1244. <https://doi.org/10.3390/buildings12081244>.
- [30] A. Alrobaie, M. Krarti, A Review of Data-Driven Approaches for Measurement and Verification Analysis of Building Energy Retrofits, *Energies* 15 (2022) 7824. <https://doi.org/10.3390/en15217824>.
- [31] S.R. Penaka, K. Feng, T. Olofsson, A. Rebbling, W. Lu, Improved energy retrofit decision making through enhanced bottom-up

- building stock modelling, *Energy and Buildings* 318 (2024) 114492. <https://doi.org/10.1016/j.enbuild.2024.114492>.
- [32] J. von Platten, C. Sandels, K. Jörgenson, V. Karlsson, M. Mangold, K. Mjörnell, Using Machine Learning to Enrich Building Databases—Methods for Tailored Energy Retrofits, *Energies* 13 (2020) 2574. <https://doi.org/10.3390/en13102574>.
- [33] H. Yang, L. Liu, X. Li, C. Liu, P. Jones, Tailored domestic retrofit decision making towards integrated performance targets in Tianjin, China, *Energy and Buildings* 140 (2017) 480–500. <https://doi.org/10.1016/j.enbuild.2016.12.040>.
- [34] B. Dong, Y. Liu, W. Mu, Z. Jiang, P. Pandey, T. Hong, B. Olesen, T. Lawrence, Z. O’Neil, C. Andrews, E. Azar, K. Bandurski, R. Bardhan, M. Bavaresco, C. Berger, J. Burry, S. Carlucci, K. Chvatal, M. De Simone, S. Erba, N. Gao, L.T. Graham, C. Grassi, R. Jain, S. Kumar, M. Kjærgaard, S. Korsavi, J. Langevin, Z. Li, A. Lipczynska, A. Mahdavi, J. Malik, M. Marschall, Z. Nagy, L. Neves, W. O’Brien, S. Pan, J.Y. Park, I. Pigliautile, C. Piselli, A.L. Pisello, H.N. Rafsanjani, R.F. Rupp, F. Salim, S. Schiavon, J. Schwee, A. Sonta, M. Touchie, A. Wagner, S. Walsh, Z. Wang, D.M. Webber, D. Yan, P. Zangheri, J. Zhang, X. Zhou, X. Zhou, A Global Building Occupant Behavior Database, *Sci Data* 9 (2022) 369. <https://doi.org/10.1038/s41597-022-01475-3>.
- [35] A. Banfi, M. Ferrando, P. Li, X. Shi, F. Causone, Integrating Occupant Behaviour into Urban-Building Energy Modelling: A Review of Current Practices and Challenges, *Energies* 17 (2024) 4400. <https://doi.org/10.3390/en17174400>.
- [36] M. Ferrando, F. Causone, T. Hong, Y. Chen, Urban building energy modeling (UBEM) tools: A state-of-the-art review of bottom-up physics-based approaches, *Sustainable Cities and Society* 62 (2020) 102408. <https://doi.org/10.1016/j.scs.2020.102408>.
- [37] C. Wang, X. Wang, F. Causone, Y. Yang, N. Gao, Y. Ye, P. Li, X. Shi, Addressing uncertainty to achieve stability in urban building energy modeling: A comparative study of four possible approaches, *Building and Environment* 267 (2025) 112197. <https://doi.org/10.1016/j.buildenv.2024.112197>.
- [38] S. Dabirian, K. Panchabikesan, U. Eicker, Occupant-centric urban building energy modeling: Approaches, inputs, and data sources - A review, *Energy and Buildings* 257 (2022) 111809. <https://doi.org/10.1016/j.enbuild.2021.111809>.
- [39] G. Happle, J.A. Fonseca, A. Schlueter, A review on occupant behavior in urban building energy models, *Energy and Buildings* 174 (2018) 276–292. <https://doi.org/10.1016/j.enbuild.2018.06.030>.

- [40] C. Bianchi, L. Zhang, D. Goldwasser, A. Parker, H. Horsey, Modeling occupancy-driven building loads for large and diversified building stocks through the use of parametric schedules, *Applied Energy* 276 (2020) 115470. <https://doi.org/10.1016/j.apenergy.2020.115470>.
- [41] A. Doma, M. Ouf, Modelling occupant behaviour for urban scale simulation: Review of available approaches and tools, *Build. Simul.* 16 (2023) 169–184. <https://doi.org/10.1007/s12273-022-0939-3>.
- [42] OpenStudio Coalition User Docs, (n.d.). <https://openstudiocoalition.org/> (accessed September 1, 2025).
- [43] W.J. Cole, K.M. Powell, E.T. Hale, T.F. Edgar, Reduced-order residential home modeling for model predictive control, *Energy and Buildings* 74 (2014) 69–77. <https://doi.org/10.1016/j.enbuild.2014.01.033>.
- [44] O.K. İşeri, İ.G. Dino, Building Archetype Characterization Using K-Means Clustering in Urban Building Energy Models, in: D. Gerber, E. Pantazis, B. Bogosian, A. Nahmad, C. Miltiadis (Eds.), *Computer-Aided Architectural Design. Design Imperatives: The Future Is Now*, Springer, Singapore, 2022: pp. 222–236. https://doi.org/10.1007/978-981-19-1280-1_14.
- [45] L. Czétány, V. Vámos, M. Horváth, Z. Szalay, A. Mota-Babiloni, Z. Deme-Bélafi, T. Csoknyai, Development of electricity consumption profiles of residential buildings based on smart meter data clustering, *Energy and Buildings* 252 (2021) 111376. <https://doi.org/10.1016/j.enbuild.2021.111376>.
- [46] S. Eiraudó, L. Barbierato, R. Giannantonio, A. Porta, A. Lanzini, R. Borchellini, E. Macii, E. Patti, L. Bottaccioli, A Machine Learning Based Methodology for Load Profiles Clustering and Non-Residential Buildings Benchmarking, *IEEE Transactions on Industry Applications* 59 (2023) 2963–2973. <https://doi.org/10.1109/TIA.2023.3240669>.
- [47] C. Lu, S. Li, J. Gu, W. Lu, T. Olofsson, J. Ma, A hybrid ensemble learning framework for zero-energy potential prediction of photovoltaic direct-driven air conditioners, *Journal of Building Engineering* 64 (2023) 105602. <https://doi.org/10.1016/j.jobe.2022.105602>.
- [48] Directive 2002/91/EC of the European Parliament and of the Council of 16 December 2002 on the energy performance of buildings, 2002. <http://data.europa.eu/eli/dir/2002/91/oj/eng> (accessed December 29, 2023).
- [49] Energy performance certificate, Boverket (2023). <https://www.boverket.se/en/start/building-in-sweden/contractor/inspection-delivery/energy-performance-certificate/> (accessed May 7, 2024).

- [50] 1. Indexing — Python Record Linkage Toolkit 0.15 documentation, (n.d.). <https://recordlinkage.readthedocs.io/en/latest/ref-index.html> (accessed January 1, 2024).
- [51] Python Record Linkage Toolkit Documentation — Python Record Linkage Toolkit 0.15 documentation, (n.d.). <https://recordlinkage.readthedocs.io/en/latest/index.html> (accessed January 1, 2024).
- [52] M. Bachmann, JaroWinkler, (2023). <https://github.com/maxbachmann/JaroWinkler> (accessed April 18, 2023).
- [53] William E. Yancey, Evaluating String Comparator Performance for Record Linkage, (n.d.). <https://www.census.gov/content/dam/Census/library/working-papers/2005/adrm/rrs2005-05.pdf> (accessed May 4, 2023).
- [54] Python library, Python Documentation (n.d.). <https://docs.python.org/3/library/functions.html> (accessed May 29, 2023).
- [55] R. Rezaee, J. Brown, J. Haymaker, G. Augenbroe, A novel inverse data driven modelling approach to performance-based building design during early stages, *Advanced Engineering Informatics* 41 (2019) 100925. <https://doi.org/10.1016/j.aei.2019.100925>.
- [56] H. Haario, E. Saksman, J. Tamminen, Adaptive proposal distribution for random walk Metropolis algorithm, *Computational Statistics* 14 (1999) 375–395. <https://doi.org/10.1007/s001800050022>.
- [57] S.-S. Jin, H. Ju, H.-J. Jung, Adaptive Markov chain Monte Carlo algorithms for Bayesian inference: recent advances and comparative study, *Structure and Infrastructure Engineering* 15 (2019) 1548–1565. <https://doi.org/10.1080/15732479.2019.1628077>.
- [58] Klimatkontrakt 2030, (n.d.). <https://www.energimyndigheten.se/forskning-och-innovation/forskning/strategiska-innovationsomraden/viable-cities/klimatkontrakt-2030/> (accessed December 16, 2024).
- [59] Tabula, (n.d.). <https://webtool.building-typology.eu/#bm> (accessed October 19, 2022).
- [60] Boverket, (n.d.). <https://forfattningssamling.boverket.se/detaljer/BFS2011-6> (accessed October 23, 2024).
- [61] N. Arsov, M. Pavlovski, L. Basnarkov, L. Kocarev, Generating highly accurate prediction hypotheses through collaborative ensemble learning, *Sci Rep* 7 (2017) 44649. <https://doi.org/10.1038/srep44649>.

- [62] S. Behera, A. Choubey, C.S. Kanani, Y.S. Patel, R. Misra, A. Sillitti, Ensemble trees learning based improved predictive maintenance using IIoT for turbofan engines, in: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, Association for Computing Machinery, New York, NY, USA, 2019: pp. 842–850. <https://doi.org/10.1145/3297280.3297363>.
- [63] P.W. Tien, S. Wei, J. Darkwa, C. Wood, J.K. Calautit, Machine Learning and Deep Learning Methods for Enhancing Building Energy Efficiency and Indoor Environmental Quality – A Review, *Energy and AI* 10 (2022) 100198. <https://doi.org/10.1016/j.egyai.2022.100198>.
- [64] S. A., S. R., A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends, *Decision Analytics Journal* 7 (2023) 100230. <https://doi.org/10.1016/j.dajour.2023.100230>.
- [65] S. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, (2017). <https://doi.org/10.48550/arXiv.1705.07874>.
- [66] Y. Ma, W. Deng, J. Xie, T. Heath, Y. Xiang, Y. Hong, Generating prototypical residential building geometry models using a new hybrid approach, *Build. Simul.* 15 (2022) 17–28. <https://doi.org/10.1007/s12273-021-0779-6>.
- [67] Y. Sun, F. Geng, L. Li, Research on Individual Aesthetic Taste Generation of Hezhe Nationality, *Asian Social Science* 10 (2014) p177. <https://doi.org/10.5539/ass.v10n18p177>.
- [68] KMeans, Scikit-Learn (n.d.). <https://scikit-learn/stable/modules/generated/sklearn.cluster.KMeans.html> (accessed July 29, 2025).
- [69] E. Tipton, Stratified Sampling Using Cluster Analysis: A Sample Selection Strategy for Improved Generalizations From Experiments, *Eval Rev* 37 (2013) 109–139. <https://doi.org/10.1177/0193841X13516324>.
- [70] Rapid Application Development with OpenStudio, (n.d.). <https://www.aceee.org/files/proceedings/2012/data/papers/0193-000371.pdf> (accessed July 26, 2025).
- [71] A Flexible Framework for Building Energy Analysis, (n.d.). <https://www.aceee.org/files/proceedings/2012/data/papers/0193-000370.pdf> (accessed July 26, 2025).
- [72] X. Zhang, F. Pellegrino, J. Shen, B. Copertaro, P. Huang, P. Kumar Saini, M. Lovati, A preliminary simulation study about the impact of COVID-19 crisis on energy demand of a building mix at a district in Sweden, *Applied Energy* 280 (2020) 115954. <https://doi.org/10.1016/j.apenergy.2020.115954>.

- [73] Startsidea — OpenDataUmea, (n.d.).
<https://opendata.umea.se/pages/startsidea/> (accessed July 27, 2025).
- [74] Sweden aging population 2023, Statista (n.d.).
<https://www.statista.com/statistics/525637/sweden-elderly-share-of-the-total-population-by-age-group/> (accessed July 27, 2025).
- [75] Sweden Employment Rate, (n.d.).
<https://tradingeconomics.com/sweden/employment-rate> (accessed July 27, 2025).
- [76] The Big Question: Where Do Swedes Go On Holiday, Farmstay Sweden (2020). <https://www.farmstaysweden.com/the-big-question-where-do-swedes-go-on-holiday/> (accessed July 27, 2025).
- [77] Var tredje planerar att spendera mer på årets semester, Mynewsdesk (2023).
<https://www.mynewsdesk.com/se/countryside-hotels-sweden/pressreleases/var-tredje-planerar-att-spendera-mer-paa-aarets-semester-3263712> (accessed July 27, 2025).
- [78] Inomhusklimatguiden_ENGELSKA.pdf, (n.d.).
https://intra.kth.se/polopoly_fs/1.1254438.1709111207!/Inomhusklimatguiden_ENGELSKA.pdf (accessed July 27, 2025).
- [79] M. Farjadnia, A. Fontan, A. Russo, K.H. Johansson, M. Molinari, What influences occupants' behavior in residential buildings: An experimental study on window operation in the KTH Live-In Lab, (2023). <https://doi.org/10.48550/arXiv.2307.08090>.
- [80] V. Fabi, R.V. Andersen, S. Corgnati, B.W. Olesen, Occupants' window opening behaviour: A literature review of factors influencing occupant behaviour and models, *Building and Environment* 58 (2012) 188–198.
<https://doi.org/10.1016/j.buildenv.2012.07.009>.
- [81] OpenStudio:model: openstudio::model::DefaultConstructionSet Class Reference - OpenStudio SDK Documentation, (n.d.).
https://openstudio-sdk-documentation.s3.amazonaws.com/cpp/OpenStudio-2.7.0-doc/model/html/classopenstudio_1_1model_1_1_default_construction_set.html (accessed July 28, 2025).
- [82] Data Aggregation with Ruby Hashes, CodeSignal Learn (n.d.).
<https://codesignal.com/learn/courses/in-depth-exploration-of-ruby-hashes-and-their-practical-use/lessons/data-aggregation-with-ruby-hashes> (accessed July 28, 2025).
- [83] Energibalansen Umeå kommun 2040, (n.d.).
<https://www.umea.se/download/18.7fe252db194a596f3289421/1739784925472/Nul%C3%A4ge,%20Energibalansen%20Ume%C3%A5%20kommun.pdf> (accessed July 29, 2025).

- [84] Ursprungsmärkning och utsläpp | Umeå Energi, (n.d.).
<https://www.umeaenergi.se/om-oss/ursprungsmarkning-och-utslapp> (accessed July 29, 2025).
- [85] Sammanfattning: Underlag till regeringens kommande klimathandlingsplan och klimatredovisning, (n.d.).
<https://www.naturvardsverket.se/496a0e/globalassets/amnen/klimatomställning/klimatredovisning/sammanfattning-underlag-2023-04-13.pdf> (accessed July 29, 2025).
- [86] Emission factors and methods used in climate impact assessments of energy use in Swedish heated buildings, (n.d.).
<https://lup.lub.lu.se/luur/download?func=downloadFile&recordId=9122136&fileId=9122142> (accessed July 29, 2025).
- [87] C.R. Upshaw, J.D. Rhodes, M.E. Webber, Modeling peak load reduction and energy consumption enabled by an integrated thermal energy and water storage system for residential air conditioning systems in Austin, Texas, *Energy and Buildings* 97 (2015) 21–32. <https://doi.org/10.1016/j.enbuild.2015.03.050>.
- [88] Y. Xie, A.I.M. Noor, Factors Affecting Residential End-Use Energy: Multiple Regression Analysis Based on Buildings, Households, Lifestyles, and Equipment, *Buildings* 12 (2022) 538.
<https://doi.org/10.3390/buildings12050538>.
- [89] (n.d.).
http://cesb.cz/cesb13/proceedings/5_tools/CESB13_1409.pdf (accessed November 8, 2022).
- [90] F. Re Cecconi, A. Khodabakhshian, L. Rampini, Data-driven decision support system for building stocks energy retrofit policy, *Journal of Building Engineering* 54 (2022) 104633.
<https://doi.org/10.1016/j.jobe.2022.104633>.
- [91] Overpass turbo - OpenStreetMap Wiki, (n.d.).
https://wiki.openstreetmap.org/wiki/Overpass_turbo (accessed December 30, 2025).
- [92] A. Alhamwi, W. Medjroubi, T. Vogt, C. Agert, OpenStreetMap data in modelling the urban energy infrastructure: a first assessment and analysis, *Energy Procedia* 142 (2017) 1968–1976.
<https://doi.org/10.1016/j.egypro.2017.12.397>.
- [93] A. Sarretta, M. Minghini, TOWARDS THE INTEGRATION OF AUTHORITATIVE AND OPENSTREETMAP GEOSPATIAL DATASETS IN SUPPORT OF THE EUROPEAN STRATEGY FOR DATA, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLVI-4-W2-2021* (2021) 159–166. <https://doi.org/10.5194/isprs-archives-XLVI-4-W2-2021-159-2021>.
- [94] RapidFuzz: rapid fuzzy string matching, (n.d.).
<https://pypi.org/project/RapidFuzz/> (accessed January 2, 2026).

- [95] N. Elmobark, A Comparative Analysis of Python Text Matching Libraries: A Multilingual Evaluation of Capabilities, Performance and Resource Utilization, *International Journal of Environment, Engineering and Education* 7 (2025) 48–60. <https://doi.org/10.55151/ijeedu.v7i1.188>.
- [96] S. Pons, R. Huertas-García, J. Lengler, D.L. de Mattos Nascimento, Natural Language Processing Algorithms to Improve Digital Marketing Data Quality and Its Ethical Implications, *Psychology & Marketing* 42 (2025) 1946–1957. <https://doi.org/10.1002/mar.22211>.
- [97] H. Liu, C.J. Hellín, A. Tayebi, F. Calles, J. Gómez, H. Liu, C.J. Hellín, A. Tayebi, F. Calles, J. Gómez, Vertex-Oriented Method for Polyhedral Reconstruction of 3D Buildings Using OpenStreetMap, *Sensors* 24 (2024). <https://doi.org/10.3390/s24247992>.
- [98] WGS 84 / UTM zone 33N - EPSG:32633, (n.d.). <https://epsg.io/32633> (accessed January 2, 2026).
- [99] geopandas: Geographic pandas extensions, (n.d.). <https://pypi.org/project/geopandas/>.
- [100] CesiumJS, Cesium (2025). <https://cesium.com/platform/cesiumjs/> (accessed January 2, 2026).
- [101] CesiumGS/cesium, (2026). <https://github.com/CesiumGS/cesium> (accessed January 2, 2026).
- [102] P. Alva, M. Mosteiro-Romero, C. Miller, R. Stouffs, Mitigating operational greenhouse gas emissions in ageing residential buildings using an Urban Digital Twin dashboard, *Energy and Buildings* 322 (2024) 114681. <https://doi.org/10.1016/j.enbuild.2024.114681>.
- [103] W.E. Winkler, String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, 1990. <https://eric.ed.gov/?id=ED325505> (accessed May 5, 2023).
- [104] A. Liaw, M. Wiener, Classification and Regression by randomForest, in: 2007. <https://www.semanticscholar.org/paper/Classification-and-Regression-by-randomForest-Liaw-Wiener/6e633b41d93051375ef9135102d54fa097dc8cf8> (accessed March 13, 2024).
- [105] Boverket, Svensk byggnorm : SBN 1980 : Utgåva 1, (n.d.). <https://www.boverket.se/contentassets/c4c3f9ae57294ae889bfaf710b08b125/sbn-1980-utg-1.pdf> (accessed May 17, 2024).
- [106] T.D. Pettersen, Variation of energy consumption in dwellings due to climate, building and inhabitants, *Energy and Buildings* 21 (1994) 209–218. [https://doi.org/10.1016/0378-7788\(94\)90036-1](https://doi.org/10.1016/0378-7788(94)90036-1).

- [107] Using the energy signature method to estimate the effective U-value of buildings, (n.d.). <https://www.diva-portal.org/smash/get/diva2:1014180/FULLTEXT01.pdf> (accessed April 28, 2025).
- [108] M.R. Seyedabadi, U. Eicker, A critical review of urban scale life cycle assessment of the built environment, *Sustainability Analytics and Modeling* 3 (2023) 100026. <https://doi.org/10.1016/j.samod.2023.100026>.
- [109] M. Jia, R.S. Srinivasan, A.A. Raheem, From occupancy to occupant behavior: An analytical survey of data acquisition technologies, modeling methodologies and simulation coupling mechanisms for building energy efficiency, *Renewable and Sustainable Energy Reviews* 68 (2017) 525–540. <https://doi.org/10.1016/j.rser.2016.10.011>.