Data science for buildings, a multi-scale approach bridging occupants to smart-city energy planning

DATA SCIENCE FOR BUILDINGS, A MULTI-SCALE APPROACH BRIDGING OCCUPANTS TO SMART-CITY ENERGY PLANNING

JULIEN LEPRINCE

This work originates as a part of the research project "Using small data and big data: Neighbourhood Energy & Data Management Integration System" which is a part of the Commit2Data program funded by the Netherlands Organization for Scientific Research (NWO).

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN
Data science for buildings, a multi-scale approach bridging occupants to smart-city energy planning, by Julien Leprince

A catalogue record is available from the Eindhoven University of Technology Library.
ISBN: 978-90-386-5753-0
Bouwstenen 355

Cover design by Ben Meuleman generated with Midjourney

Ph.D. thesis, Eindhoven University of Technology, the Netherlands

DATA SCIENCE FOR BUILDINGS, A MULTI-SCALE APPROACH BRIDGING OCCUPANTS TO SMART-CITY ENERGY PLANNING

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr. SK Lenaerts, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op maandag 22 mei 2023 om 16:00 uur door

JULIEN LEPRINCE

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

Voorzitter:	prof.dr.ir. T.A.M. Salet
Promotor:	prof.ir. W. Zeiler
Copromotoren:	prof.dr. H. Madsen (Technical University of Denmark)
	dr.ir. C. Miller (National University of Singapore)
Leden:	prof.dr. L.C.M. Itard (Technische Universiteit Delft)
	dr. R. Li (Technical University of Denmark)
	prof.dr.ir. J.L.M. Hensen
Adviseur:	dr. A. Chong (National University of Singapore)

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

"Les baobabs, avant de grandir, ça commence par être petit."

Antoine de Saint-Exupéry

Acknowledgments

This dissertation represents the culmination of four years of hard work, dedication, struggles, and passion, in the pursuit of developing impactful methods and tools to make buildings more energy efficient. While the outcome products and achievements are important, it is the process of growth that lead to them that carries the most meaning. Much like the baobab tree started off as nothing but a seed, this thesis started out as nothing more than a curious student seeking answers to a puzzle that was all but apparent. And like any plant needs tending and water to grow, this work would not have been possible without the valuable support, help, guidance, and love of numerous people which I would here like to acknowledge.

First, I would like to thank my supervisor prof. Wim Zeiler for his trust, guidance, and early vision on the necessity of data-driven analytics for buildings. Under his supervision, the building service group cultivated a thriving working environment promoting stimulating teamwork and camaraderie. His counseling went beyond the professional sphere, and I truly appreciate his genuine care for the well-being of his students as well as his profound belief in the added value of collaboration.

These elements inspired me to reach out to dr. Clayton Miller, head of the Building Urban Data Science (BUDS) Lab group at the National University of Singapore (NUS), as well as prof. Henrik Madsen, Section Head for Dynamical Systems, of Applied Mathematics and Computer Sciences at the Technical University of Denmark (DTU) for what turned out to become durable and remarkably fruitful collaborations. I am extremely grateful for the opportunities each of them provided me with, for their perpetually constructive feedbacks, and invaluable guidance. Their encouragement, patience, and kindness have helped me better refine my research and fostered my confidence as a researcher.

Next, I wish to express my profound gratitude to dr. Kaustav Basu and Rik van der Vlist, for their trust, support, and partnership. Much of my work would not have been possible without the precious contribution of field measurements collected by Eneco and I sincerely appreciate the opportunity of working alongside them both.

On the infamous 6th floor of the Vertigo building, I would like to thank the people I spent my working days with. My first thought goes to my PhD comrade, colleague, and best friend, Waqas. Just as every adventure features its legendary pair, this one would certainly not have been viable without his companionship, a fair share of dark humor jokes, and an unwavering reciprocal care that would get us through any number of Dutchy rainy day. As of today, I am proud to count

him as one of my closest friends, and profoundly thankful for his support in a great many aspects of my life. To complete the team Oindrila, a.k.a. Cindy, and Kinga, a.k.a. Kinga, have brought much-needed laughter and outdoor excursions to this chapter of my life in Eindhoven. I would also like to extend my thanks to the rest of the building services group, too numerous to name, but with whom I have enjoyed working with, lively lunch discussions, and group gatherings.

Despite a two-year pandemic break in the midst of my PhD, I was lucky enough to undertake two research exchanges in NUS and DTU. These incredible experiences have led me to discover two magnificent cities as well as meet fantastic persons. In chronological order, being finally introduced in person to the BUDS team with whom I have been in contact through slack and Teams for close to two years was especially heartening. Matias and Yi Ting were the perfect duo to make my arrival feel as homely as possible. Thanks for being your two amazing selves. I have shared a collection of beautiful moments with you both that will always stay close to my heart. I had lots of fun both on- and off-campus with so many great persons of the BUDS team; to name a few, Mario, Martín, Chun, Filippo, Miguel, and Yun Xuan, thank you for being such great researchers and company. You have made my time in Singapore unforgettable and leaving all the more difficult.
Despite the hasty transition from Singapore to Copenhagen, it was a pleasure joining the great DynSys group. I very much enjoyed my time there and thank the entire team and the enthusiasm of Peder for all our warm social and professional exchanges. For the great lunches, drinks, beating me (far too easily) at chess, and launching midnight cluster computations from across the Atlantic Ocean in times of high stress, I wish to thank Amos. I had a blast collaborating and enjoying sunny summer days in Copenhagen with you.

'Home is where your heart is' as the popular saying goes.
To the ones that helped make Eindhoven feel like home, I extend my sincerest gratitude and affection; to my roommates, Niki and Liza, who made lockdowns feel *delicious*; to Rachel and Domino, for all your affection; to the Elephants, my rugby team, both battle and drinking comrades; to the Francofilous, with all its teachers, parents and children I got to work, meet with and teach french to; to the other expat-frenchies, Charlotte, Mathias, Pierro, Sylvain, Roxanne, Lisa, Charlie, Amandine, Aurélie, and Laure, for every poker game, raclette party, rugby or football supporting evening; to Sandra and the girls, for all your love and care.
To the ones that, despite the distance, have blessed me with the sincerest friendships, to my forever friends; to Paul-Antoine and Simon, a.k.a. les Fratés, forever present no matter when, where, or what for; to the rest of my XVI

family for all our memorable adventures and love giving times; to Peg, Julie, Etienne, and Amaury, the ones I grew up with and with whom there are too many memories to single any out.

As unorthodox as this may sound, I would like to acknowledge and appreciate my own efforts in this endeavor. Undertaking this PhD has, in fact, for the most part, left me face-to-face with myself, my doubts, numerous questionings, and a quest for purpose that I alone could overcome. This ordeal was magnified by a two-year pandemic that tested my self-resilience, perseverance, and ambitions, in a trialing but ultimately deeply rewarding way. Investing and believing in myself were pivotal key ingredients that guided the steady success of this project for which I am deeply thankful.

Lastly, I wish to thank my family, my parents, and my sister for their unconditional love and encouragement. Whether it rains or shines, whether today or a year from now, their unwavering support has been the foundation upon which I have been able to grow and flourish as a man. You are forever the anchors of my life, the best role models a son and big brother could ask for. I love you with all my heart.

Julien Leprince
May 1st, 2023

Abstract

In a context of global carbon emission reduction goals, buildings have been identified to detain valuable energy-saving abilities. With the exponential increase of smart, connected building automation systems, massive amounts of data are now accessible for analysis. These coupled with powerful data science methods and machine learning algorithms present a unique opportunity to identify untapped energy-saving potentials from field information, and effectively turn buildings into active assets of the built energy infrastructure. However, the diversity of building occupants, infrastructures, and the disparities in collected information has produced disjointed scales of analytics that make it tedious for approaches to scale and generalize over the building stock. This coupled with the lack of standards in the sector has hindered the broader adoption of data science practices in the field, and engendered the following questioning:

How can data science facilitate the scaling of approaches and bridge disconnected spatiotemporal scales of the built environment to deliver enhanced energy-saving strategies?

This thesis focuses on addressing this interrogation by investigating data-driven, scalable, interpretable, and multi-scale approaches across varying types of analytical classes. The work particularly explores descriptive, predictive, and prescriptive analytics to connect occupants, buildings, and urban energy planning together for improved energy performances.

First, a novel multi-dimensional data-mining framework is developed, producing distinct dimensional outlines supporting systematic methodological approaches and refined knowledge discovery. Second, an automated building heat dynamics identification method is put forward, supporting large-scale thermal performance examination of buildings in a non-intrusive manner. The method produced 64% of good quality model fits, against 14% close, and 22% poor ones out of 225 Dutch residential buildings. Third, a pioneering hierarchical forecasting method was designed, bridging individual and aggregated building load predictions in a coherent, data-efficient fashion. The approach was evaluated over hierarchies of 37, 140, and 383 nodal elements and showcased improved accuracy and coherency performances against disjointed prediction systems. Finally, building occupants and urban energy planning strategies are investigated under the prism of uncertainty. In a neighborhood of 41 Dutch residential buildings, occupants

were determined to significantly impact optimal energy community designs in the context of weather and economic uncertainties.

Overall, the thesis demonstrated the added value of multi-scale approaches in all analytical classes while fostering best data-science practices in the sector from benchmarks and open-source implementations.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Overview

- Why is data science key to unlock augmented building energy performances?
- State of the art: obstacles for the broader adoption of data science practices in the building sector
- Contributions and novelty of this thesis
- Thesis structure overview

"Knowledge is power."
Thomas Jefferson

"We are living in the era of information" is a popular dictum; truly, however, we are living in the data era. In 2020, the total amount of data created, captured, copied, and consumed globally reached a new high of 64.2 zettabytes, and is projected to grow to more than 180 zettabytes over the next five years up to 2025 [1]. This exponential increase in accessible data volume is due to the rapid advancement of powerful data collection and storage appliances coupled with the computerization of our society. Businesses worldwide generate colossal data sets, which, combined with the universal accessibility of data makes our time undeniably the data age [2]. In turn, functional and effective tools to automatically uncover the richness of information from gigantic amounts of data and transform it into organized knowledge soon became a necessity. This has led to the prominent ascension of data science.

Concurrently, the world is facing its greatest challenge to date: global warming. To tackle the global climate crisis and meet net-zero targets set by the European Green Deal [3], in line with the Paris agreement [4], countries around the world urgently need to decarbonize their economies by 2050. This requires them to simultaneously reduce their current energy demand while significantly increasing the penetration of renewable energy sources in decentralized energy systems [5]. However, the volatility of weather-driven renewable energy sources has introduced a shift in the electrical grid paradigm; originally unidirectional grid systems, providing energy to end-consumers from large dispatchable power plants, have effectively turned to bidirectional (smart-grid) systems, where energy consumers become *prosumers*

by producing a portion of the energy they consume [6]. This restructuring of the power grid demands adequate planning, monitoring, and automatic control supported by sensors, actuators hardware, telecommunication links, and computer-based algorithms. Such tools back optimal energy management techniques to match produced renewable energies to the varying energy demands supplied by the network for reduced carbon emissions.

Under these circumstances, buildings were identified as the largest energy consumer in the world, accounting for over one-third of overall final energy consumption [7], and from 40% to nearly 70% of global energy-related CO_2 emissions [8] by the Organisation for Economic Cooperation and Development (OECD). This effectively places buildings as a key centerpiece for reaching environmental and sustainability goals. Activating improved energy efficiencies in buildings, while ensuring smart energy management in a grid-connected context, has attracted a considerable amount of attention from industry and research both [9], [10]. Supported by building automation systems (BAS) [11], building planning and operational strategies have shifted from fixed schedules to smart, adaptive, responsive more energy-efficient ones. The abundance of collected data from buildings has opened the doors to a myriad of data-driven applications serving enhanced energy performances, such as automated-fault detection and diagnosis [12], building retrofitting [13], anomaly detection [14], load prediction [15], and demand-side management [16], supporting predictive control strategies [17].

However, the full exploitation of data science techniques is challenging for the sector. Indeed, due to the aging and exceedingly heterogeneous building stock [18], collected information is often disparate [19] which, coupled with the lack of established benchmarks [20], has hindered the development of data science practices in the sector. This development fostered the subsequent interrogation:

"How can data science support the decarbonization of the building sector?"

Let us investigate the potential of data science in uncovering knowledge from data and the specific challenges facing the building sector in embracing data science principles.

Data science to uncover knowledge from data

Data science is formally defined as a set of fundamental principles, processes, and techniques for understanding phenomena via the (automated) analysis of data [21] supporting decision-making processes. It provides practitioners with structure and methods to systematically approach the extraction of useful knowledge from data in order to avoid a "data rich but information poor" outcome [2]. This multi-disciplinary subject integrates combined techniques from mathematics and statistics, domain knowledge, and machine learning thanks to high-performance computing, as showcased by Figure 1.1. Machine learning is a category of artificial intelligence that enables computers to learn and resolve problems on their own.

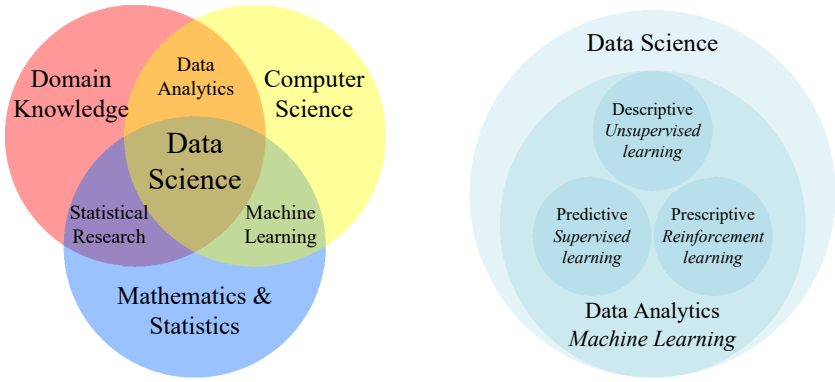


Figure 1.1: Data science in the context of computer science, domain knowledge, and mathematical and statistics (left), and data analytics and machine learning taxonomies in the context of data science (right).

The general learning problem was defined by Tom Mitchell as a computer program that learns from experience E with respect to some class of tasks T and performance measure P, if its performance on T, as measured by P, improves with experience E [22]. The data-driven foundations of machine learning techniques particularly provide high generalization capacities [23] which, in a context of high data availability, has allowed the method to strive.

Data science is commonly dissected into three classes of analytics: (i) descriptive analytics *"what is happening?"*, (ii) predictive analytics *"what is likely to happen?"*, and (iii) prescriptive analytics *"what should be happening?"* [15].

(i) **Descriptive analytics** aim to quantify events, report on them, and are a first step in turning data into actionable insight. They are diagnostic application-oriented analytics and intend on achieving a better understanding of the causes of a given process, e.g., identifying patterns or abnormal behaviors. Descriptive techniques have been judged most capable at discovering formerly unknown knowledge from large data sets [24] and are commonly related to unsupervised learning techniques which can organize instances without pre-specified attributes. These aim at finding underlying associations or data structures between variables from clustering and association rule mining (ARM) algorithms, but also visualization and data characterization techniques.

(ii) **Predictive analytics** intend on determining the likelihood of future events from historical data. From the analysis of sufficient numbers of training sets, i.e., data objects for which the desired output is known, it approximates a model or function to forecast future variable realizations. Predictive analytics are logically associated with predictive or supervised learning, its equivalent machine learning taxonomy. Supervised machine learning methods attempt to capture complex and nonlinear relationships between

inputs (independent) and outputs (dependent variable) of an observable phenomenon [25] by learning from historical data. It is then employed to predict the discrete or continuous value of observations yet unforeseen. Popular supervised methods comprise classification [26] and regression [27] methods.

(iii) **Prescriptive analytics** provide enterprises with automated, time-dependent, adaptive, and optimal decisions [28] and are exploited to generate more value to businesses thanks to strategic and operational decisions [29]. They derive optimal decision recommendations to assist decision-makers in reaching a desired outcome from either: mathematical optimization models and predictions, or actionable predictive models and their associated feedback data [30]. The former employs optimization, simulation, and evaluation methods [31], while the latter exploits a purely data-driven technique, i.e., reinforcement learning a computational approach to understand and automate goal-directed learning and decision-making [32].

The inherent capabilities of data science and machine learning to learn patterns from large datasets while generalizing acquired knowledge to new, unseen data have changed the status quo. Combined, they provide efficient and computationally tractable methods for big data knowledge extraction, resulting in improved operations, reduced costs, and increased system efficiency.

In the building sector, data science techniques are notably applied to support the optimization of energy consumption [33], reduce operational costs, and improve building performance [34]. Predictive models are exploited to forecast maintenance needs, thus reducing downtime and increasing building equipment lifespan [12], [35]. By analyzing measurements collected from BAS, building operators can effectively reduce energy usage and identify factors that affect building performance [36]–[38].

However, barriers in the sector still persist, preventing the full exploitation of data science potential for buildings.

State of the art: a multitude of approaches for a multitude of buildings

The main obstacles to a broader adoption of data science practices in the building sector can be grouped into three problematics:

(i) a lack of standards in building data analytics practices,

(ii) the tedious scalability and generalization capabilities of data-driven approaches over highly heterogeneous buildings and occupants,

(iii) the multiple yet disconnected scales linking smart buildings to smart cities.

These challenges are here briefly summarized while a more detailed review is later presented in their dedicated chapters.

On the need for standards in building data analytics

While data science extensively demonstrated its value potential for buildings due to its ability to generalize and scale from large data-driven insights, obstacles still persist preventing domain professionals from exploiting its full potential.

To start with, developed research methods commonly follow disparate steps, increasing analytical uncertainty with unsystematic analytical procedures. Data mining framework showcased clustering of energy time series as a data preprocessing step prior to predictive learning or association rule mining applications [39], [40]. C. Fan et al. [25] proposed a mining method tailored to large building automation system data encompassing data cleaning and transformation methods (preprocessing) while presenting the data partitioning phase separately. Data mining references, however, report data cleaning, integration, selection, and transformation as separate steps prior to knowledge discovery [2]. This has consequently left building professionals with the tedious task of determining which analytical steps to follow for targeted building analytics.

Secondly, the lack of consistency across building data sets hinders the widespread adoption of proposed data-driven solutions and increases the cost of scaling these applications and systems to different buildings [19]. Additionally, studies commonly relied on predefined problems employing only a small subset of available building data [40] with few established benchmarks to compare results from one investigation to the next. E. Keogh and S. Kasetty [41] notably exposed a need for a set of time series benchmarks coupled with more careful empirical evaluations of data mining research. Else contributions made would offer a negligible amount of improvements if not tested on sufficient representative real-world datasets. Establishing benchmarks in the building sector would effectively support the identification of the most suitable technique to a given application [20]. Building performance open data sets were introduced yet still necessitate a wider adoption, e.g., Building Data Genome project [20], Pecan Street [42], Low Carbon London [43].

Highly heterogeneous buildings and occupants

The existing and aging building stock is known to be highly diverse, consequently producing a challenge for the building industry. The International Energy Agency (IEA) reports that at least 40% of buildings in developed economies were built before 1980 [44]. In the United kingdom, it is estimated that over 75% of buildings in use today will still be standing in 2050 [45]. Building performances can consequently vary significantly based on factors such as building age, size, occupancy patterns, and geographic location [46]. For example, energy consumption in office buildings can vary by up to 200% due to differences in occupancy patterns and equipment use [47]. Similarly, A. Thornton et. al showed that

heating, ventilation, and air conditioning (HVAC) energy use in commercial buildings can vary by up to 60% due to differences in building systems, design, and occupancy [48]. The global building stock is highly diverse and fragmented, with over 70% of buildings in the world being either small or medium-sized [18]. This exacerbates further the heterogeneity challenge as there are few standardized approaches for building performance management across varying building types and sizes.

What's more, building occupants are a notoriously unpredictable constituent of building energy systems resulting in a myriad of energy needs, referred to as *occupant behavior*. Driven by multiple contextual, sociological, or psychological factors, occupant behavior is exceedingly tedious to characterize [49]. It has consequently become the leading source of uncertainty in predicting building energy use [50], [51] inducing the so-called building performance gap [52]. These behaviors commonly include interactions with thermostats, plug-in appliances, operable lights, windows, or blinds. The energy-saving potential originating from occupant behavior is evaluated to stand between 10-25% for residential buildings, and 5-10% for commercial buildings [53].

All in all, the unique characteristics of buildings and their occupants make the building stock a highly heterogeneous body. The development of tools and methodologies for building performance analysis and management that can scale across buildings is crucial to improve building energy efficiency and reducing carbon emissions.

From buildings to cities: a landscape of multi-disconnected scales

As of today, buildings and city energy system analytics encompass multiple spatiotemporal scales often disconnected.

In a predictive analytical context, this implies that produced forecasts from separate individual buildings (spatial scale) or hour-ahead estimations (temporal scale) do not typically match with their aggregated counterpart predictions, i.e., district-level (spatial) and day-ahead (temporal) ones. This results in inconsistencies in produced information across energy networks which prevent energy management strategies from being optimally conducted. Indeed accurate and *coherent* predictions across varying aggregation levels and horizons of the considered energy system are required for optimal decision-making, else decision-makers would be planning using separate and possibly conflicting views of the future.

From a prescriptive analytical perspective, this disconnect between spatiotemporal scales is commonly due to the differences in modeling details between building-focused energy management problems and strategic urban energy planning problems [54]. Energy planning problems at the neighborhood, city, or country scale typically need to reduce the encompassed dimensionality through spatial and temporal aggregations to render resulting optimization problems computationally tractable. For example, the planning of a residential neighborhood would consider both a typical, representative, year of operation, to reduce the

temporal dimension of the problem, as well as aggregated energy demands from clusters of buildings or apartments to simultaneously downscale its spatial granularity by using weighted representative elements only. Yet, these necessary simplifications deprive planners from exploiting the full extent of available synergies between prosumers of energy communities. Activating untapped energy flexibility potentials such as demand-side management in the planning phase could significantly improve system efficiency and reduce planning costs. The question of relevant scale identification in urban energy planning is in fact, not a new one. Cajot et al. [55] stated that it should be regarded rather as an open question, for future research to provide planners and decision-makers with rigorous and systematic tools necessary to quantify the gains and losses of different boundaries. And while extensive works have focused on the integration of smart buildings to grid-level energy management problems, the long-term provisioning of energy and sizing of power-system elements are typically not considered due to their resulting problem complexity [56]. This often results in over-simplified models of the grid edges and its users, as illustrated by the conclusion of Chiroma et al. [57], which state that building occupants and smart grid end users still require extensive research to be successfully incorporated to the smart grid operation scheme. Addressing the gap between these two distinct scales is essential for a holistic approach and linking user behavior assortment to smart city energy infrastructure planning.

Contributions and outline of the thesis

This thesis contributes to overcoming these limitations, thus bridging the gap between the different spatiotemporal layers of the built environment, grounded on data-driven, scalable, interpretable, and multi-dimensional methods. Particularly, connecting occupant behavior to urban energy system planning lead this thesis to explore descriptive, predictive, as well as prescriptive analytics. The main research question of this work can be defined as:

How can data science facilitate the scaling of approaches and bridge disconnected spatiotemporal scales of the built environment to deliver enhanced energy-saving strategies?

To further deconstruct the research question, we formulate four sub-questions to examine all analytical classes, namely descriptive, predictive, and prescriptive. These are formulated as follows:

- *Analytics* ■ How can data mining support the deconstruction of multi-dimensional analytics for the building sector?
- *Descriptive* ■ What data-driven methods can best provide scalable and interpretable models for building heat dynamics?
- *Predictive* ■ What is the added value of a unified bottom-up and top-down predictive learning approach for building load forecasting?
- *Prescriptive* ■ Can occupant behavior (bottom layer) affect city energy planning (top layer)?

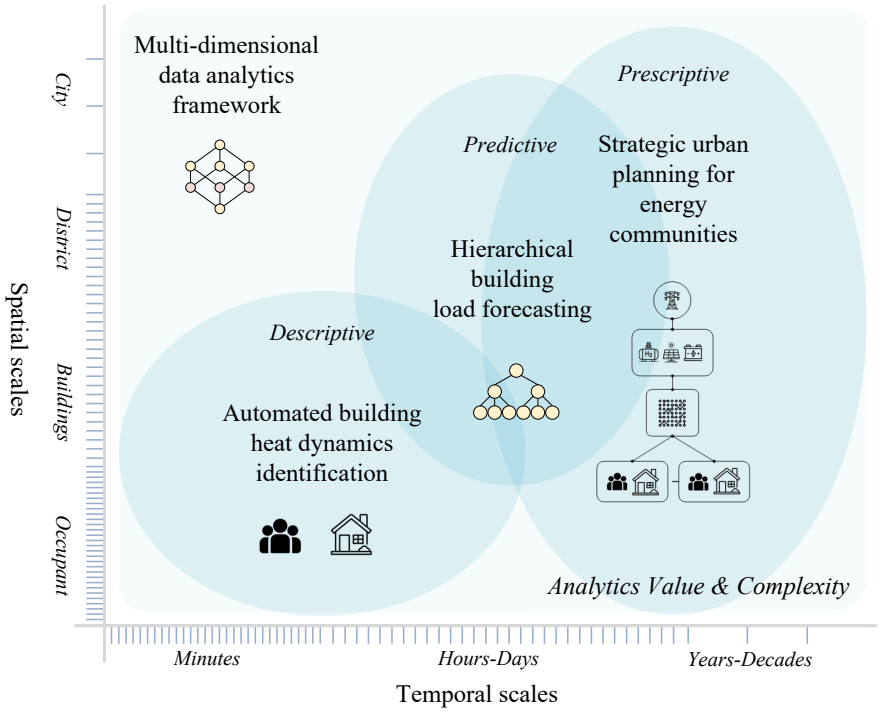


Figure 1.2: Thesis analytical and spatiotemporal scale frames.

To answer this question, the thesis proposes the following methodological contributions:

Chap. 2 ■ Data cube mining for buildings

A generic method is first framed to define best practices for data-driven analytics in the building sector in a high multi-dimensional data context. The framework particularly leverages data cubes as a foundation on which to structure dimensional frames of interest. The approach is exemplified by an automated pattern identification application of building performance data. The work notably puts forward the inherent connections between dimensional frames of interest and the insights uncovered from common data analytics approaches, namely bottom-up, top-down, and temporal drill-in.

Chap. 3 ■ Scalable building heat dynamics identification

To better identify and characterize the impact of occupant behavior on building energy demands, the thermal dynamics of residential buildings need to be identified in a scalable and interpretable manner. This is explored employing two principal data-driven modeling technics: grey- and black-box models. While black-box models are inherently powerful at scaling across multiple (buildings) data sets, their black-box analogy betrays a lack of interpretability. On the other hand, grey-box approaches

are endowed with physical knowledge of the modeled system thus producing highly meaningful models, yet their scaling across the building stock is tedious. This work proposes to tackle both of these shortcomings and produces, from a similar case study of 250 occupied residential buildings, interpretable, calibrated, building thermal models; an essential foundation of all building to grid energy management applications.

Chap. 4 ■ Hierarchical building load forecasting
To bring forecasted scales of the built environment together, i.e., building to cities, hierarchical forecasting is investigated as a holistic solution for multi-dimensional forecasts. Spatiotemporal hierarchical structures are first defined, and a novel approach is then introduced, producing coherent individual and aggregated forecasts of building energy time series data. The developed hierarchical regressor is further tested on varying deep neural network architectures echoing typical hierarchical structures to demonstrate the value brought by the coherency information of the multi-scale forecast. The outcome is a single, coherent forecast of an entire energy community, allowing aligned decision-making across the energy network.

Chap. 5 ■ From building occupants to urban energy planning
Finally, bringing together the different layers of the built environment, i.e., occupant behavior (bottom layer) and the urban energy system (top layer), a scalable, distributed, stochastic, and multi-objective energy planning solution for communities is designed. Calibrated building thermal models are exploited, providing detailed granular information to the urban energy planning problem. The stochastic optimization problem produces policies accounting for future uncertainties of the system, i.e., climate, economic, and occupant behavior, in a decentralized framework suitable for real-world deployment. Lastly, the impact of user-behavior uncertainty on the overall system design is estimated within the context of other system uncertainties, thus demonstrating the relative impact of occupants on urban energy systems, and effectively bridging both these scales.

To secure research reproducibility this thesis open-accesses produced codes for all chapters under distinct public GitHub repositories. These are regrouped under one account[1] with each contribution specified in the overview of their associated Chapter. Produced results are additionally anchored onto open-data-based benchmarks for Chapters 2, 4, and 5. The methodological contributions of the work particularly focus on tackling scalability purposes while uniting multi-disconnected scales together across analytical categories. The spatiotemporal scope of the presented work is illustrated under Figure 1.2.

The outline of the thesis is illustrated in Figure 1.3, where Chapter 2 presents a generic multi-dimensional data analytical framework applied to automated pattern identification.

[1] https://github.com/JulienLeprince

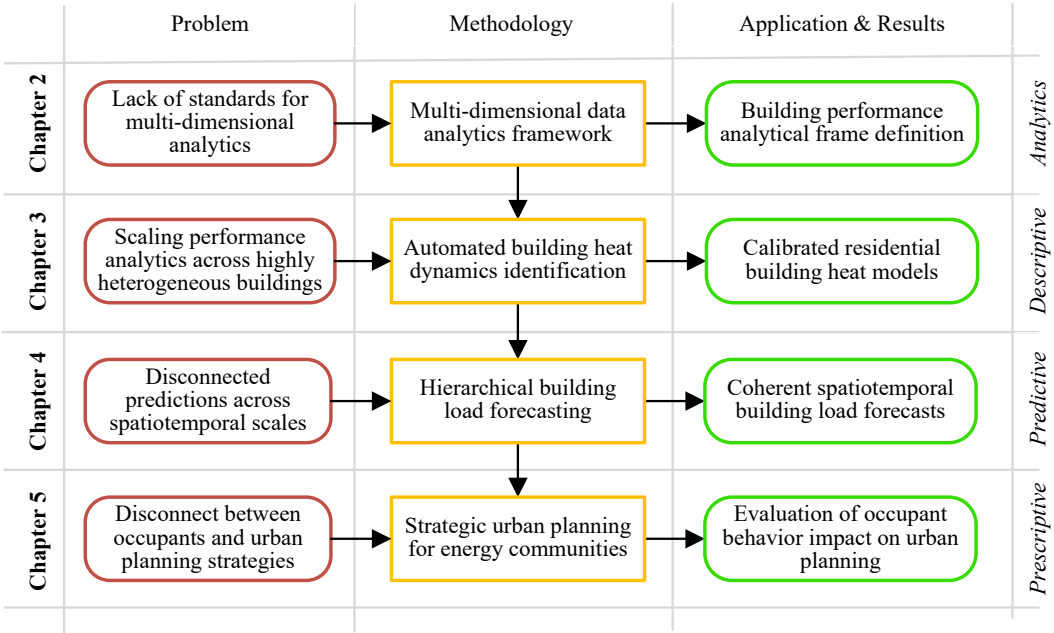


Figure 1.3: Thesis structure overview.

Chapter 3 puts forward an automated building heat dynamics identification applied to 250 Dutch residential buildings while Chapter 4 introduces hierarchical learning regressors as a novel coherent forecasting method applied to energy networks. In Chapter 5 occupant behavior is connected to the urban energy planning system scale while considering future system uncertainties. The latter generates optimal strategies evaluated and compared against varying uncertainty scenarios. Finally, concluding remarks summarize the main findings, provide recommendations and guidance for the broader adoption of multi-scale data science practices in the building sector, and envision future perspectives.

Annex A is complementary to Chapter 3 and details a black-box approach for the identification of building heat dynamics. The work uses symbolic regression as a means to identify interpretable analytical expressions from a similar data set. Its results are compared to typical grey-box lumped resistance capacity models.

Chapter 2

Data cube mining for buildings

Chapter overview

- Presentation of a multi-dimensional framework for building data analytics
- Application to automated pattern identification
- Case study: Building Data Genome Project 2 (BDG2) open dataset
- GitHub repository: /multidimensional-building-data-cube-pattern-identification

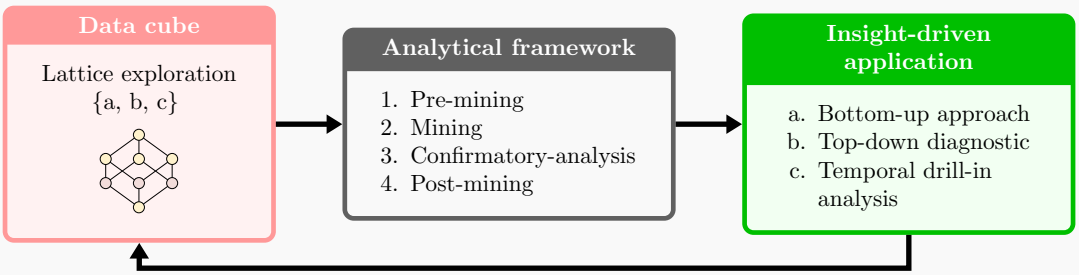


Figure 2.1: Overview of the multi-dimensional data (cube) analytical method

This chapter has been published as Leprince et al. [58].

"The real voyage of discovery consists not in seeking new lands but in seeing with new eyes."
Marcel Proust

2.1 Preface

As briefly introduced in the Introduction, advances in information systems, computing power and control technologies for optimal resource management, have endowed building automation systems (BAS) with enhanced energy savings ranging from 20 to 35% [59], while generating a huge amount of data from a wide range of appliances every day. These include essential building indoor environment quality condition processes such as ventilation, lighting, air conditioning, and heating but also home appliances such as dishwashers, laundry, kitchen

devices, and home entertainments. Yet, BAS data are rarely fully exploited and interpreted. Improving building energy efficiencies with such pools of data remains a challenge; how does one approach the analysis of various associations and correlations amongst multi-temporal, i.e., seconds to hourly resolutions, with daily to decades horizons, and high dimensional data? What methods to follow to acquire useful, interpretable insights on building energy performance and reduce its consumption? Such questions root upon causes usually involving poor data quality, resulting from a large share of missing values and outliers, coupled to lack of efficient and convenient analytical tools and methods for large data sets. Additionally, most BASs only perform basic data analytics and visualizations, such as historical tracking, moving averages, and threshold-based anomaly detections which have pushed the building automation industry to new data-driven methods and tools to harvest these data pools, namely, data mining.

2.1.1 Data mining

Data mining (DM) has grown from a promising technology to an established powerful and effective analytical tool to interpret massive and complex data. In 2001, MIT reviewed DM as one of the top 10 emerging technologies that will change the world [60], while it has now accumulated over a hundred thousand publication records over the last 20 years in a wide variety of fields [61], including medicine, retails telecommunication, financial services, and target marketing [62]. In the building sector, the effervescence surrounding the technology was such that recent reviews employed text mining tools to fully uncover the extent of developments in the field [63]. DM is a multi-disciplinary subject, integrating combined techniques from statistics, machine learning, and artificial intelligence thanks to high-performance computing. It is the core process of identifying valid, useful, and understandable patterns from large and complex datasets, known as Knowledge Discovery in Databases (KDD). The DM taxonomy established by Oded and Lior [62] distinguishes two preeminent types of DM: verification oriented, where the system verifies a proposed hypothesis, and discovery-oriented, where the system identifies new rules and patterns autonomously. Verification methods include traditional statistical tests such as goodness of fit test, test of hypotheses (i.e. t-test of means, one sample Z-test), and analysis of variance (ANOVA). Discovery methods, on the other hand, are based on inductive learning, where a model is constructed from generalized sufficient numbers of training examples, assuming its applicability to future unseen data. Another terminology, widely used within the machine learning community, preferably considers discovery-oriented DM and separates the techniques into supervised and unsupervised learning. Supervised methods attempt to discover complex and non-linear relationships between input and output target attributes (referred to as independent and dependent variables respectively) by learning from historical data. This type of process largely composes the predictive learning component of DM discovery methods. It has been applied to the building operational stage [63] as it is directly linked to occupant comfort and responsible for 80-90% of the building's total green gas

emissions [24]. Applications of supervised learning notably include predictions of building energy consumption [64]–[67], thermal load [68], [69], indoor environment [70], [71], and system performance indices [72]–[74]. Popular supervised methods comprise two predominant groups: classification [75] and regression [26]. Unsupervised learning, also recognized as descriptive learning, groups techniques used to organize instances without pre-specified attributes. It aims at finding underlying associations or data structures between variables. The prominent advantage of unsupervised analytics is its ability to discover formerly unknown knowledge [2], [27]. Well-established techniques involve clustering, association rule mining (ARM), and anomaly detection. Visualization and summarization techniques, for instance, are DM descriptive methods that are not regarded as unsupervised learning. In opposition to predictive learning, descriptive learning can be viewed as a more flexible application that does not require model training or predefined targets during knowledge discovery. Its main applications encompass fault/anomaly detection and building performance diagnostics [76]–[78].

2.1.2 Cube multidimensional analytics

Dealing with the large volumes, velocities, and varieties characterizing high-dimensional big building data is a complex task. Common analytical tools developed to tackle multidimensional data rely on exploring different dimensional associations at different levels of aggregation leveraging the structures of a data cube [27]. A data cube is defined as a multidimensional data model allowing data exploration from its structured dimensions, i.e. dimension table and facts. A data cube is commonly organized around a central theme, represented by a fact table, which contains names of the different facts, or numeric values, and relational attribute keys. For example, a building fact table could include time, location or energy flow attribute keys linking them to their dimension table. Given a fixed set of dimensions, a cuboid can be generated for each subset of the given dimensions. Their combinations result in a lattice of cuboids, presenting the data at specific levels of summarization from which a multi-dimension analytical map can be defined. Cuboids forming the lowest level of summarization are denoted base cuboid, while the 0-D cuboid, holding the highest level of summarization, is designated the apex cuboid (typically referred to by all) [27], [79]. This lattice of cuboids defines the data cube. While data cubes are commonly represented as 3-D geometrical structures, they are naturally n-dimensional, where each dimension represents objects intended to keep record off. In BAS data, hierarchical relationships are often found within dimension tables, e.g., the time dimension includes a natural tree structure rooted in the year attribute, and progressively branching out to months, weeks, days, and hours. Other hierarchically structured dimensions typically include geographical location and site measurements. Multidimensional cube space analytics rely on the high-dimensional structure of the data to explore multi-lattice and abstraction levels of the cube. Common dimensional exploration methods rely on bottom-up, top-down approaches, namely rollup, where fine granularities are gradually aggregated in coarser ones, and drilldown, starting

from coarser dimensional granularities down to finer ones. This navigation across multiple cube spaces of interest is called OnLine Analytical Processing (OLAP) [80]. By summarizing and aggregating data subsets at different abstraction levels, this tool has greatly assisted multidimensional analytics. Leveraging this approach, R. Ramakrishnan and B. Chen [81] have put forward a cube-space mining method, taking advantage of the data-cube structure to define and select cuboids of interest to mine over. This way, data mining can be used as a building block within the OLAP analysis to exploit multi-scale knowledge discovery in a defined dimensional frame. Characteristics of the cube-space data mining scheme involve the following three steps: (1) relying on cube space to determine the space of candidates for mining, (2) employing OLAP queries to explore features and targets for mining and (3) adopting data-mining models as building blocks within a multi-step mining process. This exploratory multidimensional DM approach, also known as OnLine Analytical Mining (OLAM) [2], allows the user to effectively select and analyze a relevant subset of data at different granularities and present discovered knowledge at different abstraction levels.

2.1.3 Motivation

This being said, OLAM has, to the best of the authors knowledge, little to none been practiced in the built environment sector, and while DM extensively demonstrated strength and performance in this domain, barriers still persist avoiding professionals from exploiting the full potential of DM analytics. Previous studies usually relied on predefined problems using only a small subset of building data with few established benchmarks to compare results from one investigation to the next [81]. Additionally, developed research methods commonly follow disparate steps, increasing complexity with unsystematic mining analytical procedures. With the variety and complexity of the most recently developed DM techniques as well as the highly dimensional building data, it has become increasingly challenging for building professionals to (i) effectively target which data dimensions to explore and consider in their analytics, (ii) determine what analytical steps to follow for targeted building data mining and (iii) select the most suitable DM technique for a particular case study from established references. Realizing the prevalent demand for a common DM framework, noticeable studies have proposed methods applied to BAS data [24], [40], [82]. However, developed frameworks were usually tailored to DM application-specific cases and failed to address multi-dimensional analytical approaches from orderly steps required for systematic and benchmarked building data analytics. Detailed stepwise generic approaches with established good practices for preprocessing, application-specific and benchmarking procedures are frequently overlooked yet desperately needed. In order to adopt systematic analytical steps from a common framework within the building analysts and research community, several steps still need to be undertaken; (i) establishing and following a common DM framework and (ii) developing and employing open building data toy sets to serve both as benchmarks to case-specific studies while allowing (iii) the development of replicable implementations of typical building energy management applications for valuable knowledge transfer. Ensuing these steps would

cultivate more generalizable findings and insights while vastly contributing to the practical adoption of a common analytical frame. This study proposes a response to this appeal and puts forward a multi-dimensional analytical method grounded on a generic data mining framework for building data analysis. It puts together reviewed analytics best practices in a step-wise method tailored to DM application for systematic knowledge discovery in big building data. Contributions of this work can be summarized as three-fold;

- Putting forward a generic building-tailored DM framework for unified and systematic analytics,
- Framing a multi-dimensional analytical approach to big building data, cutting down the complexity endowed by high-dimensionality, and
- Providing an open access implementation of the presented method relying on a large and open building data set, serving as benchmarks to similar studies and appealing to more reproducible, comparable, and empirically validated analytics.

2.2 Method

Resulting from an in-depth analysis of DM methods and a comprehensive review of domain application-driven techniques we propose a generic DM framework tailored to multidimensional building data. The developed method is founded on established methodology from the literature. Notable existing frameworks typically involve four major phases, i.e. data preprocessing, data partitioning, knowledge discovery (data mining), and post-mining. In particular, the generic framework developed by C. Fan, F. Xiao, and C. Yan [24] designed for BAS data knowledge discovery englobed building performance assessment, diagnosis, and optimization as possible applications. Our method follows similar steps yet extends it from a multidimensional viewpoint leveraging both descriptive and predictive mining techniques while importantly stressing prerequisites for reproducible and generalizable results from benchmarks. It puts forward a generic feed-forward and back process to follow while attempting any building mining process and differentiates mining application-dependent steps from generic DM ones from a unified and interpretable method. Our method is illustrated in Fig. 2.2 where two tasks are performed in the data preprocessing phase, including data integration and data cleaning. Multidimensional data exploration then follows, incorporating benchmark reports and cube lattice selection with OLAP exploration. After, a pre-mining phase incorporates data transformation and mining-specific steps. Next, the mining stage takes place and an important confirmatory analysis phase is thereafter carried out with validation methods leading to algorithm selection. A feedback loop linking confirmatory analysis to the mining and pre-mining blocks is included in the framework to indicate potential iterative sequence allowing pre-mining steps and mining to be repeated to converge to the desired results for algorithm optimal selection. And the OLAM feedback loop illustrates the repeated mining process over different cube lattices for multidimensional mining. Knowledge

interpretation and extraction is then proceeded within the post-mining phase, supported by visualization tools. Finally, discovered knowledge can be used for a defined application, or serve as a preliminary step to another mining phase, as illustrated with the last feedback arrow. This is often the case when mining for association rules or undertaking predictive learning with prior profile clustering for example [82], [83]. Details of the evoked phases are developed in the following subsections.

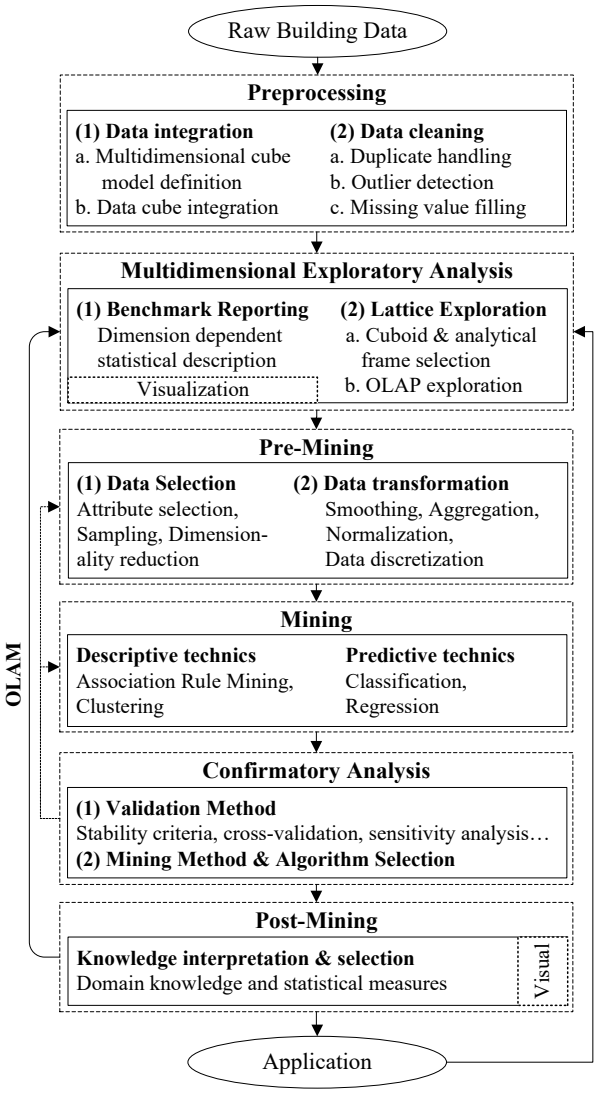


Figure 2.2: A generic multidimensional data mining framework for building data.

2.2.1 Preprocessing

Data preprocessing completes two main tasks, i.e. data integration and data cleaning (outlier identification, missing value handling). Data integration refers to the selection of a suitable structure format and data model for the later analysis. Cleaning aims at enhancing data

quality to obtain suitable results out of the intended analytics. It has been reported in DM literature that data cleaning should be performed prior to data integration, allowing information industry to benefit from clean 'usable' data stored in data warehouses [2]. This work considers the analytical process from a scientific point of view, where data may be cleaned in different ways consequently impacting the later analysis, which is why we recommend the data be stored raw rather than preprocessed[1].

Data integration

Data integration is composed of a first data model definition phase, from which the later integration process can be undertaken. Data model definition constitutes a fundamental first step to structure the multidimensional BAS data under a given schema. Data integration techniques can later be applied consequently providing consistency in naming conventions, encoding structures, and attribute measures [2].

Data cube map

Establishing the building data mapping serves as an imperative step to the framing and structuring of its various dimensions. Additionally, obtaining clear delineated dimensions allows leveraging the design of a data cube into decomposed lattices that will serve in shaping the later OLAM analytics. A common approach to the data cube model definition originates from the formulation of the analysts' interrogations. Specifying questions such as "what is the energy consumption relationship to time?", framing the analysis under the energy and time dimension, or "what was the total energy consumption of a building in a certain location during a specific time interval?", here querying along three dimensions, serves in the conceptualization of the state space to explore and, thus, in the definition of the data cube dimensions. Building data gathers six types of recorded data, from building operations to metadata combined [38], echoing quite conveniently the 6 facets of a cube; i.e, time, location, building data encapsulating operational and meta-data, climate conditions, occupant-related information and equipment data. Time data serve as a reference index to the other measured attributes, and indicates Year, Month, Day, Hour, Minute, Second, Day Type, generally formatted under the ISO 8601 [84] recognized complete format "YYYY-MM-DDTHH:MM:SSZ", e.g., 2019-07-16T19:20:30+01:00. Location straight forwardly regroups spatial delineations such as geographic coordinates or address, which can be divided into numerous granularities, i.e., device, room, zone, system, building, street, district, city, state and country. Building data regroups building characteristics and operational data. Operations cover energy demands originating from building comfort maintenance with heating and cooling loads, lighting and ventilation systems through electric power loads, heat flows, or natural gas consumption, but can also cover also water supply. Metadata

[1] Ideally, if data storage space permits it, both raw and preprocessed versions of the data should be stored, to allow preprocessing reproducibility evaluation as well as alternative variants that could be considered in other studies.

evokes the building's physical characteristics commonly encasing floor area, number of floors, global insulation coefficient, window-to-wall ratio, date of construction, and building type (school, dwelling, office building, hospital, education...). Climate conditions assemble indoor or external environmental conditions with attributes such as dry-bulb temperature, relative humidity, irradiance, wind speed, precipitations, pressure, and air quality but also non-temporal characteristics such as the Köppen climate classification [85]. Occupant information can deal with both occupant characteristics and comfort data. Occupant characteristics are seldom collected as a result of their privacy-sensitive nature as well as tediousness to gather through costly surveys. They cover attributes of age, gender, education, lifestyle, annual income, and other socio-economic parameters [53], [86]. Occupant comfort data relate to physiological, psychological, and environmental factors influencing human comfort perception, i.e. thermal, visual, and aural comfort [87]. Equipment data possess a non-temporal and operational entity, namely equipment characteristics, i.e. equipment type, efficiency, capacity, and operating system settings, i.e. set-point temperature, inlet and outlet equipment temperatures or pressures, control parameters, and events accompanied with eventual respective causes (human or agent initiators) [88]. Given these dimensions, one can group study-specific available data to form a dimensional mapping of the cube. Given an n-dimensional cube, each dimensional-element d_i, with $i \in [1, n]$ can thus be associated into groups of increasing size, i.e. cuboids. Cuboids of a given size compose a lattice, where each lattice $l \in [0, n]$ will thus be composed of $P_(n, l)$ possible partition cuboids from equation

$$p_{n,l} = \binom{n}{l} = \frac{n!}{l!(n-l)!} \ . \tag{2.1}$$

Figure 2.3 illustrates a 4D cube mapping example given the building cube dimensions: time (T), resource consumption (R), External conditions (E), and location (L). The cube can then be reduced by eliminating non-relevant dimensional associations from analyst inspection. Here the 2D cuboid association T, L can be eliminated as location is, by essence, non-temporal. Consequently, all emerging cuboids can also be eliminated from the cube space, resulting in a reduced state space mapping. Establishing the cube data mapping provides a conceptual and structured model, dividing data into clear separate dimensions, on which the later analytical processing can be founded on. It may also be noted that the hierarchical structure inherent to some dimensions, e.g., time or location, possess abstraction levels called footprints which represent granularities accessible for later OLAP exploration of the cube space [2].

Data cube integration

Integrating the data cube to a suitable format for mining processing then follows. Building data are typically recorded in two dimensional tables where a set of attributes (columns) representing a variable are stored across different instances (rows). Within the defined dimensions stated earlier, different levels of measurements are often required for in-depth

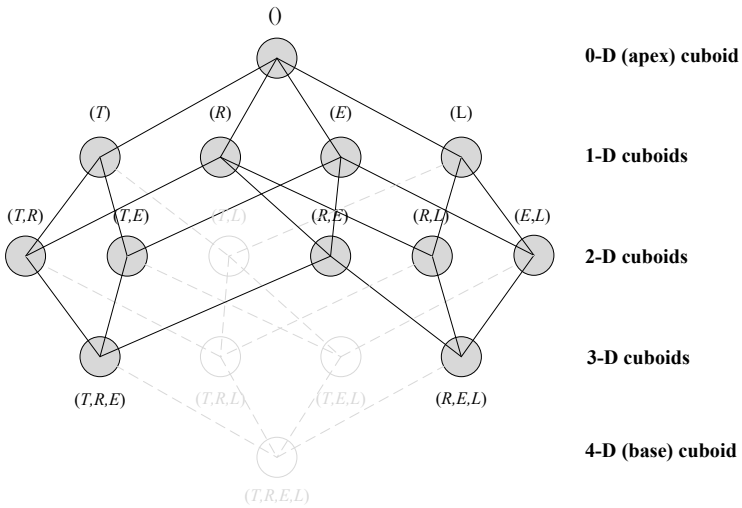


Figure 2.3: Building 4D data cube mapping and space reduction example, where (T) relates to time, (R) resource consumption, (E) External conditions and (L) location.

building energy performance analytics, i.e. from building site scale to room, equipment or component-point measurements, increasing data dimensionality and complexity. For instance, HVAC systems often require multiple outlet temperature, pressure, and air-flow point measurements, with one aggregated component energy consumption. Differentiating these relationships in an ergonomic and analytically efficient way becomes crucial for effective DM. A prevalent adopted solution proposes common markup language and data structure to organize the collected information: Project Haystack [89]. Data are organized hierarchically from three entities, i.e. site, a single building with a unique street address, Equip, physical or logical pieces of equipment within a site, and Point, referring to sensors, actuators or set point values of an equipment. Following this reference, a multi-column format is proposed where sets of attributes are grouped hierarchically under common sites, consequently structuring attributes from similar buildings under a common table.

Data cleaning

Data cleaning is a crucial step to efficient DM analytics aiming to improve data quality by dealing with duplicates, clearing outliers, and filling missing values from raw BAS data. It is unfortunately still common to find little to no information on the data cleaning phase of many studies [82], [90], [91]. This first major phase of DM legitimately effects the outcome of the later analysis and should always be clearly reported to assure proper result benchmarking. To the best of the author's knowledge, every existing BAS analytics from the literature perform missing values filling prior to outlier detection, as a consequence of the few existing

methods robust to missing values. This work introduces a shift in this established order to avoid using tampered sets for missing value filling which can result in a greater share of produced outliers, consequently making them harder to identify in the later step.

Duplicate data handling

Duplicates in data sets consist of data objects that are corresponding or identical to one another to some extent. In BAS data, these can consist of redundant attributes within a data set, or multiple attributes stored in a common instance (timestamp), sometimes with different values, also referred to as inconsistencies. Their sources cover use of denormalized tables, inaccurate data entry or updating some but not all data occurrences [2]. They can create major issues when merging data from heterogeneous sources and should be handled first within the cleaning phase. Duplicates handling is seldom depicted in BAS literature and usually consists of candid duplicate attribute, or instance, lookup functions coupled to targeted removals if the duplicates are identical. Handling inconsistencies however yields different alternatives, i.e., keep only one duplicate over the others, average the inconsistencies out or remove them from the data. Knowledge on the origin of a set of duplicates can help identify erroneous data and chose an appropriate strategy for duplicate handling.

Outlier detection

An outlier can be defined as data point that is significantly dissimilar to other data points or that does not imitate the expected behavior of others [92], [93]. In BAS data, outliers can come from measurement faults (sensor), transmission or transcription anomalies due system changes or human errors. Natural outliers reveal unusual but occasional behaviors of the monitored phenomenon. Outliers can be grouped in two main groups, i.e., point and subsequences outliers [93]. This phase of DM should only consider point outliers identification as recommended by the work of Fan et. al. [25], not to later overlap with mining typical/atypical patterns. Outlier detection methods include prediction models, profile similarity approaches, and deviants identification [93]. Prediction models spot-out outliers by comparing measured values from predicted ones with an outlier score threshold-based comparison. The primary variation across models concerns the particular prediction model considered (supervised, unsupervised). The profile similarity approach is based on a reference normal profile built upon historical data to which new time points are compared to. Outliers are then identified from time-dependent normal profiles and variance vector comparison with anomaly score. The deviant-based method estimates outliers from a minimum description length (MDL) standpoint originating from information theory. If the removal of a point in a time sequence results in a significantly simpler sequence to describe, then it is considered an outlier.

Missing value filling

Missing data in BAS are commonplace, with multiple processes being monitored from seconds to hourly frequencies on a yearly basis, gaps are typical within raw BAS data. Missing data originate from error or omissions when data is recorded or transferred [94],

imperfect procedures of manual data entry, incorrect measurements, and equipment error [25]. Discontinuities may lead to serious obstacles when analyzing findings [95], e.g., loss of efficiency, complications in data handling and analysis, bias estimates from dissimilar lengths of data, and reduction of statistical power (inefficient estimates) [96]. Selecting an appropriate method for missing data handling depends on the time series pattern and the missing data mechanism [97]. Challenges related to these techniques involve, maximizing available data use to preserve covariance structure in multivariate data [98], and incorporating variance estimates of the uncertainty rooted on imputed data [99]. If the gaps represent more than 60 percent of the set, however, then no method is judged suitable to cure the set [25]. Missing value-filling methods cover either deterministic approaches, known as single imputation, or stochastic ones, also referred to as multiple imputations, where several values are generated for each missing observation to reflect the uncertainty of the missing data [95]. This work proposes a single imputation approach dependent on the length of the missing sections. Explicit modeling with regression can be chosen for missing data sections smaller than 3 consecutive hours, i.e., moving average [25], while longer sections call for implicit modeling using the hot deck method, i.e., where missing values are averaged from identical time intervals and day of the week using sections of 2 weeks. Interested readers are invited to refer to the work of M. Norazian Ramli et. al. [100] for in-depth review of imputation methods.

Multidimensional exploratory analysis

This section intends on framing multidimensional data exploration leveraging the BAS data cube representation. It holds the essential role of identifying data structures, distributions, and trends, needed for benchmarking purposes and defining appropriate mining approaches for the investigated set. Additionally, it supports more generalizable, interpretable, and framed analytics by (i) cutting down the complexity of big data from cube lattice selection with OLAP exploration and (ii) putting forward important benchmark reporting characteristics. Exploratory Data Analysis (EDA) was originally defined by John W. Turkey as the act of "looking at data to see what is seems to say" [101], [102]. It aims at collecting insights into data characteristics to help with the following analysis by answering questions such as; what does the data look like? How can one visualize the data to get a better sense of it all? How are the values distributed and can similarities between attributes be measured [2]? Existing explored characteristics comprise attribute types, i.e., nominal, binary, ordinal, numeric, discrete or continuous, and statistical descriptions, i.e., central tendency, dispersion, variance, and correlations. Attribute-type exploration is carried out during the first data integration phase, however, statistical feature inspection can be performed a priori or posteriori, to data cleaning. EDA is here presented a posteriori to data preprocessing in the DM framework as a necessary step to encase multidimensional mining. First benchmark reporting presents the dimension-specific data structures, providing necessary insights to the later pre-mining phase to which follows, lattice/cuboid selection and OLAP exploration.

Benchmark reporting

EDA serves as a necessary data structure reporting appliance to any scientific study. As the work of B. Yildiz et. al. demonstrates, BAS data characteristics should systematically be described to allow validation of a study's true success thanks to a defined analytical framework [103]. Yet, too many studies fail to report these features. Description of household characteristics such as dwelling types, age and physical condition, household loads statistical components and climatic conditions using established classifications, e.g. Köppen Climate Classification [85], should henceforth systematically be reported [103]. While undertaking EDA, it becomes necessary to define what the authors propose to call the analytical window frame which encompasses three elements, i.e., data granularity, horizon and frame. Granularity refers to the sampling rate of the data set over the selected dimensions. The horizon entails the largest dimensional attribute considered and the frame defines the dimensional region of interest within the analysis. For example, typical building energy pattern analytics tend to use temporal analytical windows with 15-min to hourly granularity, yearly horizon and daily frame [35]. Statistical features examination is often performed through data visualization. It communicates data structures and tendencies clearly and effectively from graphical representation endowing users with a straightforward understanding of the data [2]. A series of three data visualization techniques are hereby presented capturing the data dimensions' inter- and/or intra-attribute structures.

1. Combined half-violin and boxplots allow appreciation of central tendency, distribution and variance with an assessment of statistical inference at a glance via overlaid boxplots [104], while avoiding the redundant mirrored probability density functions of violin plots.

2. Scatterplot matrix coupled to correlation matrix display the marginal dependence structure of the data [105], granting examination of intra-attribute correlations and attribute distributions from bivariate relationships. These plots are favored as particularly effective for feature engineering and visualization [106].

3. Weekly framed heat maps are suggested as a substitute to run charts, enabling inspection of per attribute patterns leveraging a weekly to daily analytical frame of interest using hourly resolution and yearly horizon.

Lattice exploration

Lattice exploration embodies the starting step of the iterative cube-space mining, i.e., OLAM loop [81]. A cuboid is firstly chosen from the established data cube dimensions for OLAP exploration of the multidimensional data. For instance, given a 3-D building data cube covering time, site and attributes dimensions, see Fig. 2.4, three sets of 2-D cuboids can be iteratively selected and explored, i.e. time, site, time, attribute and site, attribute. Typically, analytical frames explored in building performance mining encompass only one of the three presented cuboids, i.e., top-down, bottom-up and the less common temporal

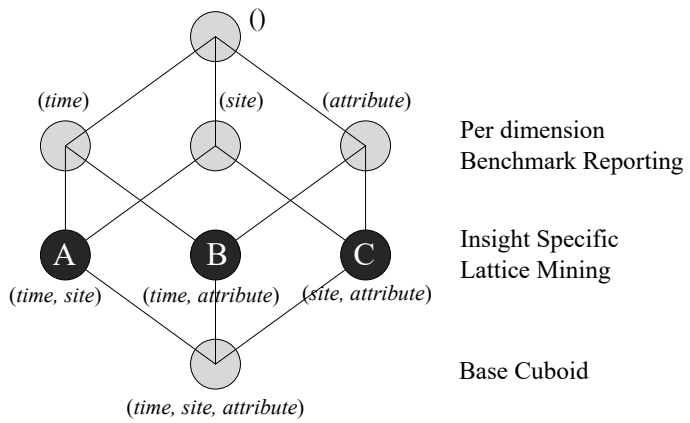


Figure 2.4: Building 3D data cube mapping with Benchmark reporting and lattice exploration.

drill-in approach, respectively corresponding to cuboids A time, site, B time, attribute and C site, attribute. By examining varying levels of abstractions through lattice exploration, information and insights sharing between them can be exploited; a concept also employed in transfer learning [107]. C. Fan et. al. [108] recently demonstrated its value across buildings for short-term building energy predictions, particularly when measured data are limited. Multidimensional mining can then exploit varying levels of data abstractions from drilling, pivoting, filtering, dicing and slicing of the data cube. Leveraging data visualization to these ends notably expands the power and flexibility of data mining [2].

2.2.2 Pre-mining

Pre-mining is by nomenclature the phase completed prior to mining. Customarily, this process is treated within pre-processing as it shares the objective of preparing the data for mining [2], [24], [25], [38], [40], [109]. This study suggests differentiating application-independent steps from the dependent ones and introduces pre-mining in the DM framework as a mining-specific preprocessing phase which can be iterated over in response to confirmatory analysis results. Pre-mining englobes two principal functions, i.e. data selection, for targeted and computationally efficient mining, and data transformation, to prepare the data to a suitable type and range for mining.

Data selection

Data selection, also referred to as data reduction, answers to a necessary step in big-data mining originating from the sheer volume and high dimensionality of the data. Indeed,

addition of data volumes from keeping irrelevant attributes or loss of decisive information from withdrawing relevant ones will likely be detrimental to the mining process; it may slow the mining algorithm employed while leading to discovered patterns of poor quality [2] and has been recognized to play an equally important role as ML model development throughout the pipeline of DM [106]. To that end, data selection encompasses measures that are attribute selection, sampling, and dimensionality reduction techniques. Attribute selection proposes to straightforwardly reduce the data set dimension by removing irrelevant or redundant attributes (or features). Note that this process can also involve the creation of new attributes, from combined information of removed ones. Its aims at finding a minimum set of attributes while keeping the original probability distribution of the classes as unaffected as possible [2]. Feature selection is commonly conducted by sequential backward selection (SBS), where attributes are sequentially removed till the reduced space contains the desired number of features [106]. Existing techniques commonly evaluate and rank individual or subsets of data attributes, e.g., information gain attribute ranking, relief, principal component, and correlation-based feature selection [110]. Data sampling involves using statistical techniques to select, manipulate and analyze a representative (sub)set of data, usually resulting from the need to reduce the size (dimension) of the enormous data set considered, i.e. under-sampling. Over-sampling on the other hand is less frequent within the big data era, as a result of the overabundance of already collected data. Yet, over-sampling can be used to test the robustness of mining results and highlight sensitivity of the approach to sampled realizations. Dimensionality reduction techniques, or data reconstruction methods [24], serve as a means to reach reduced representations of the data while minimizing information loss [2]. Main techniques include wavelet transform, which provides high and low frequency decompositions of signals based on wavelet approximation coefficients [111], and principal component analysis, where low-dimensional attributes are created from orthogonal linear transformations of the original high-dimensional ones.

Data transformation

Data transformation addresses data conversion to suitable types, ranges and noisiness to serve as DM algorithms input. Indeed, depending on the mining technique considered different data formats are required, e.g. categorical or numerical, while BAS data can exhibit varying units, scales and data type [25]. To this end, this phase covers data normalization, aggregation, smoothing and discretization. Normalizing a time series consists in scaling its attributes within, or around a smaller range or value, typically [-1, 1] or [0, 1]. This step is commonly performed to allow scaled comparisons between dissimilar ranges of attributes, e.g. normalizing features allows balanced contributions in the update of model weights during the training phase of predictive learning. Typical normalization methods cover min-max, z-score and decimal point normalizations [2]. Aggregation similar data groups together, also known as binning or bucketing, consists in applying summary operations to the data. For instance, sample-rate conversions resample the data by aggregating values together at regular

instances, i.e., down-sampling, where daily intervals are reduced to weekly or monthly ones. The reverse operation, interpolates data across larger resolutions, e.g. up-sampling to convert hourly instances to 15 minutes interval ones. Smoothing serves to remove noise from data which is frequently used to uncover trends in noisy time series and can alleviate overfitting pitfalls of regression models. Usual techniques include binning, with either equal width or frequency, regression or clustering [2]. It can be interestingly noted that the previously presented down-sampling rate conversion method, can also achieve smoothing effects as a data binning technique. Discretization involves data type transformations such as converting numeric features to interval or conceptual labels [2], e.g., 30-50, 50-70 or adult, senior, respectively. This step is consistently required for mining algorithms such as Association Rule Mining (ARM), i.e., the frequent-pattern growth and Apriori algorithms that can only handle categorical data [24]. It should be noted that feature construction relates more to data transformation, as the work of J. Han et al. reports [2]. Yet, because this framework treats the ordering in which these steps should be taken, it was chosen to include it within feature selection, to apply data transformation techniques a posteriori to feature engineering.

2.2.3 Mining

Mining, or knowledge discovery, encapsulates the algorithmic mining of the data which entails a large number of varying techniques. Selecting the appropriate one for a given application is part of the difficulties most data practitioners are faced with and is, naturally, function of the nature of the problem and the given data set, or case study. Going towards more interpretable DM analytics, we propose to group these techniques in two application-oriented groups, i.e., descriptive and predictive techniques. These groups echo the, well-established machine-learning families that constitute unsupervised and supervised learning respectively, while clearly distinguishing the typical end goals one can expect from such methods. It is beyond the scope of this study to give a complete review of all existing DM methods, however, principal mining groups will here be revised.

Descriptive techniques

By definition, descriptive techniques are diagnostic application-oriented analytics and intend on achieving a better understanding of the causes of a given process, i.e., identifying patterns or abnormal behaviors. Descriptive DM techniques, as opposed to predictive ones, have been judged more capable at discovering previously unknown knowledge from BAS data [24]. Descriptive techniques cover the important DM groups of clustering, and Association Rule Mining (ARM). Clustering is the process of grouping a set of data objects (or observations) into subsets or clusters. Each object within a cluster is similar to one another, yet dissimilar to objects in other clusters. Similarities and dissimilarities are assessed based on attribute values describing the objects and often involve distance measures, or metrics [2]. Clustering algorithms have been broadly applied to identify typical building operation patterns, e.g.,

building energy demand patterns, indoor environment distribution and building energy system operation patterns. Main clustering algorithms involve k-means clustering with many variants including adaptive k-means and k-shape clustering, Fuzzy C-Means (FMC), support vector clustering, hierarchical clustering or decision tree-based clustering and Self-Organizing Maps (SOM) [35]. ARM is a powerful tool designed to extract association rules amongst attributes from large amounts of operation data. Association rules are commonly an implication of the form "A→B", where A is defined as the antecedent and B the consequent. In general, ARM can be viewed as a two-step process where all frequent item sets are firstly identified, from which strong association rules from the frequent item sets satisfying minim support and confidence can then be generated [2]. Variations of ARM recently applied in BAS data include Temporal Association Rule Mining (TARM), or sequential rule mining, to encapsulate the temporal dimension within the discovered rule. Common ARM algorithms encompass TRuleGrowth, Weighted ARM, QuantMiner, Apriori, ParaMiner and CloseGraph. Some notable TARM algorithms include TRuleGrowth, SPADE and CMRules [25].

Predictive techniques

Predictive mining intends on determining the likelihood of future events from historical data. It constructs a model, or function from the analysis of sufficient numbers of training sets, i.e., data objects for which the desired output is known. Predictive DM is often employed to capture complex and nonlinear relationships between inputs (independent) and outputs (dependent variable) of an observable phenomenon [25]. It is then employed to predict the discrete or continuous value of observations yet unforeseen. Familiar DM predictive techniques comprise regression and classification-based methods. Regression analysis is a statistical methodology most often applied for numeric prediction of missing or unavailable data values. It also covers the identification of distribution trends from available data [2]. Methods include Artificial Neural Networks (ANN), deep neural networks, Support Vector Machines (SVM), Decision-Trees (DT), Genetic Algorithms (GA) and ensemble learning [112]. On the other hand, classification predicts categorical labels (unordered, discrete). Classification forms an analysis that identifies a model describing the data into distinguishable classes or concepts. The models are built on targeted attributes fitted to the value of predictor attributes. Data classification aims to classify data into distinct predefined classes, providing the description categorization and generalization of a given database [82]. It includes algorithms such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Decision-Trees (DT), Bayesian Network (BN) and ensemble models, i.e., random forest . [113]

2.2.4 Confirmatory analysis

Confirmatory analysis provides answers to the questions of model accuracy estimation; "what are appropriate measures of a model's goodness?" and, if there are multiple models to choose

from, "how to selection the best model?" from them. These inquiries relate to method validation, and model selection respectively. This phase embottles the two earlier pre-mining and mining phases and constitutes the keystone of this iterative process. It defines the method around which the mining will be performed and consequently arises as a founding phase of the analytical approach. Confirmatory analysis in DM echoes the statistical process of evidence evaluation from significance, inference, and confidence tests; it is the phase where findings and arguments are put to trial. This phase is usually implicitly included in the earlier mining step and explicitly framing such a key step of the mining process is becoming imperative to approach more interpretable mining analytics. The confirmatory analysis step ergo includes validation method and model selection.

Validation Method

Determining what decides the goodness of a mining's process and how to assess it is by all means what the validation method deals with. The what serves to quantify the evaluated characteristics, commonly employing performance metrics. These characteristics can cover speed, robustness, scalability, interpretability, and mining-dependent validity indicators, e.g., purity, similarity index, or accuracy [2]. The how works towards obtaining representative evaluated characteristics and assures reliable results are obtained. Common methods employed for model assessment contain cross-validation, bootstrap, sensitivity analysis, and hyper-parameter tuning. Defining how a mining model is validated and under what criteria is the fundamental foundation of any mining process.

Algorithm Selection

With evaluation characteristics, metrics, and validation method defined and undertook, model and associated parameters are selected from multi- or single-criteria assessments. Most DM work evaluate model quality from one criterion at a time such as accuracy or interestingness with a single-criterion assessment [114]. Some works have proposed multi-criteria evaluation methods to combine multiple measures in their model selection. The review of Aruldoss et al. cover a few of them, namely fuzzy and non-fuzzy analytics hierarchy process, TOPSIS, grey theory, data envelopment analysis, weigthed sum models [115]. Panapakidis and Christoforidis have notably developed a multi-criteria decision method for optimal selection of clustering algorithms applied to load profiling applications [116].

2.2.5 Post-mining

Post-Mining intends on bridging practical applications with mined discovered knowledge. This step requires domain expertise for knowledge selection and interpretation which can become particularly time-consuming [24], [76]. Knowledge selection can refer to varying application-dependent processes, e.g., characterizing identified load profiles from clusters [117], [118], or selecting relevant rules for interpretation amongst massive ARM outputs [25].

Typical end-use applications of building data mining englobe building energy load prediction, predictive maintenance, fault detection, and diagnosis, building performance analysis, and energy management optimization.

2.3 Implementation

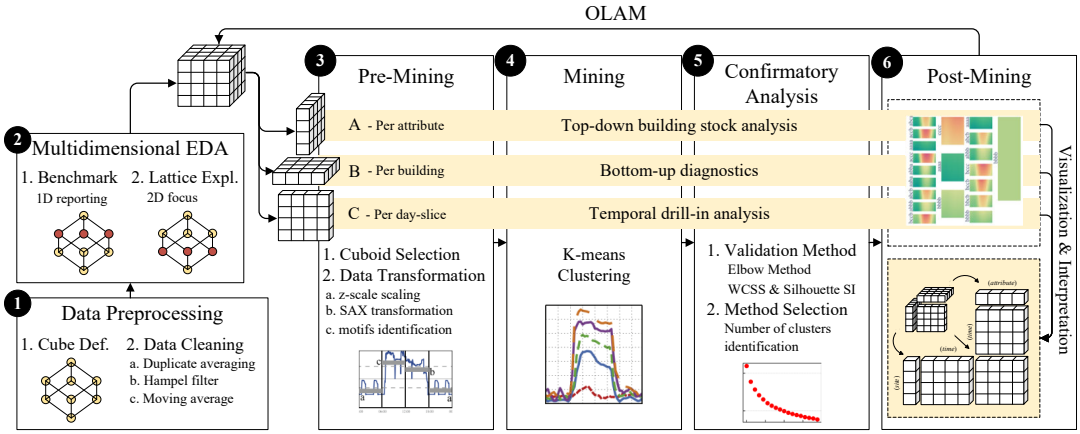


Figure 2.5: Diagram of three-dimensional cube space SAX pattern mining steps.

We implement the given method on an established automated pattern filtering application for building performance analysis proposed by Miller et. al, namely DayFilter [76]. Using time series Symbolic Aggregate approXimation (SAX) daily profiles are first segmented into W equal sized segments, piece-wise approximated across each of these segments and finally transformed to alphabetic characters according to a chosen alphabet size A, creating breakpoints of equiprobable regions from Gaussian distribution. For example, fixing W=4 and A=3 could produce the sequence 'abca' where each alphabet character would correspond respectively to a 'low-medium-high-low' segment values. SAX transformation results in reduced representations of daily profiles allowing computationally efficient differentiation of daily motifs from discords. The method steps are visually presented under Fig. 2.5, where the arrows in the diagram designate the sequence execution flow from steps 1 to 6. The iterative cube space OLAM process is repeated within steps 3 to 6, where each cuboid selection orients the analysis to either prevalent top-down, bottom-up or temporal drill-in approaches. The complete code implementation of the reported study can be found under an open github repository[2], for more interpretable as well as transparent knowledge transfer.

We use the open data set from the Building Data Genome project 2 (BDG2) [20]. This open set was chosen to allow reproducibility of the given analytics while illustrating how large open-source reference sets can be beneficial for DM analytics benchmarking. The BDG2

[2]https://github.com/JulienLeprince/multidimensional-building-data-cube-pattern-identification

includes 3,053 energy meters from 1,636 non-residential buildings located in Europe and, principally, North America. The set covers two full years (2016-2017) at an hourly resolution with multi-meter building measurements as well as weather and building meta-data.

We consider a simple 3D data cube regrouping dimensions of time, site, attribute, recall Fig. 2.4, to illustrate the given pattern identification. The attribute dimension encapsulates weather and building meter-data combined to allow a simplified space cube mapping where its 2D lattice echoes typical study frames of inter- and intra-building analytics. The inter-building analytical frame, i.e. A=time, site cuboid, is typically relevant for building stock diagnosis or benchmarking given a fix attribute, while the intra-building frame, i.e. B=time, attribute cuboid, serves for within-site diagnosis for the selected building across time. The rather unfamiliar C cuboid regrouping site, attribute dimensions, allows diurnal drill-in exploration of cross-building/attributes combined information from a certain time slice of interest. To grasp the complexity endowed with high dimensionality while keeping the use case relatively simple, we select temporal attributes of electricity, gas, hot water and chilled water of site meter energy consumptions and pair them with external condition attributes of air temperature and sea level pressure. Meter data contained approximately 4.74%, 6.20% and 6.94% of missing values for electricity, hot and chilled water respectively with maximum lasting periods ranging between 0.25 and 1.5 days, lower and upper quantiles for electricity respectively. First, the Hampel filter, an outlier robust rolling window method, was applied to detect point-wise outliers with a window size of 6 time-steps (hours) and a standard-deviation threshold of 3 [119]. A moving average was then used to fill in missing data points for consecutive gaps smaller than 4 hours. Greater gaps were averaged from identical time intervals and days of the week using sections of 2 weeks. The dimension-dependent EDA of the building stock meta-, weather, and meter-data can be found under the publication's GitHub repository [20] and will thus not be repeated here.

The 2D lattice is then selected for cube-space exploration as it encompasses typical study frames while paving the way to the dimensionally more complex 3D base cuboid. The following mining steps treat the OLAM iterative process of data selection, transformation, clustering, validation and knowledge interpretation, over the selected 2D lattice of the cube. The time series are first normalized through a z-scale transform to obtain an approximate mean of 0 with standard deviations approaching 1 [120]. Echoing the work of Miller et al. [76] we do not normalize the series based on individual sub-sequences and rather take the full temporal scope of the time series into consideration. This allows us to discover patterns leveraging both the magnitude and shape of the original profiles revealing the seasonality within the series. SAX transformation then serves as a blended data dimensionality reduction, aggregation and smoothing technic. It is performed over the time series considering segments W=4 and alphabet size A=3. This selection of SAX parameters is driven by the desired signal granularity and coarseness of the reduced time series approximation. More detailed patterns could be generated with increasing segment and alphabet size, however, ensuing

the findings of Miller et al. these parameters have been found to provide the best balance between the number of patterns generated and detailed resolution required to filter discords in a diurnal frame context. Heuristically, we set an absolute count threshold at 10 to filter motifs from discords, to which succeeds Euclidean distance-based K-means clustering, further reducing the pattern groups. We validate the optimal number of clusters through an elbow method assessment using two similarity indexes, i.e. within-cluster sum of squares (WCSS) and silhouette score. Finally, we present results using expressive visualization tools, allowing human result inspection for efficient and interpretable knowledge extraction. Diurnal heatmaps were retained as a particularly impacting visualization plot allowing 3-dimensional inspections within a 2D domain. The below sub-sections present the cuboid-specific analytical particularities and diagnostic focus of the evoked pre-mining, mining, confirmatory analysis and post-mining steps, i.e., building benchmarking, in-site view, and temporal drill-in analysis, with a closing multi-cube-space visualization interpretation step.

2.3.1 Building benchmarking

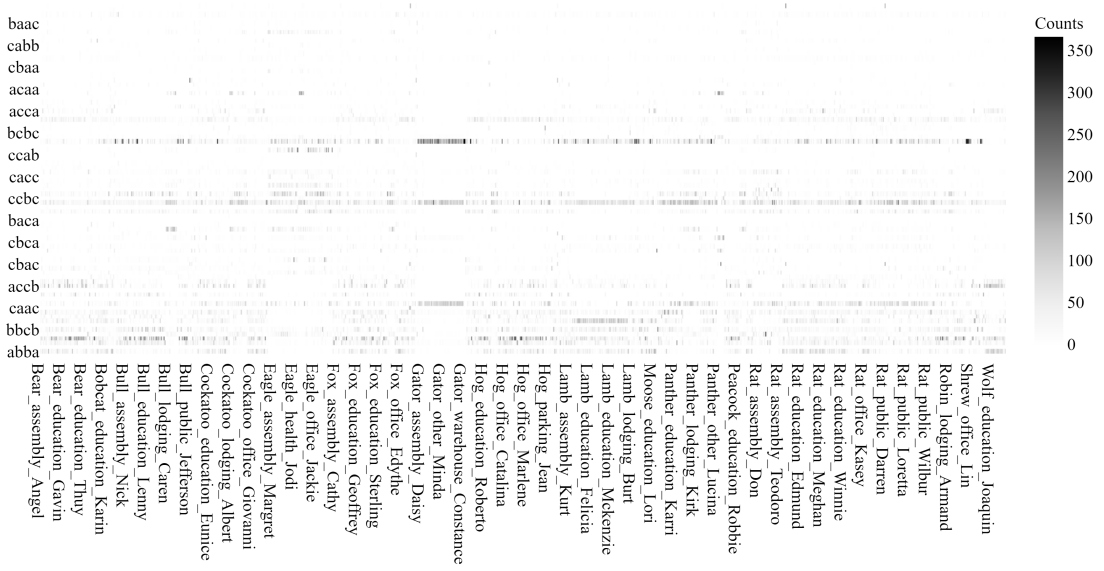


Figure 2.6: Heatmap of electric meter consumption SAX counts accross building stock.

A high-level top-down building stock diagnosis is first undertaken to gather insights from building energy consumption profile ranges and orient the subsequent lattice exploration. The time, site dimensions are sliced from the building data cube, and the site electric meter consumption attribute is chosen as both a conventional and representative energy resource consumption metric. Time series normalization is performed per site and across time, capturing each buildings' energy consumption profile shape and seasonal range. SAX sequences are obtained from the z-scaled time series and discords are filtered out from the

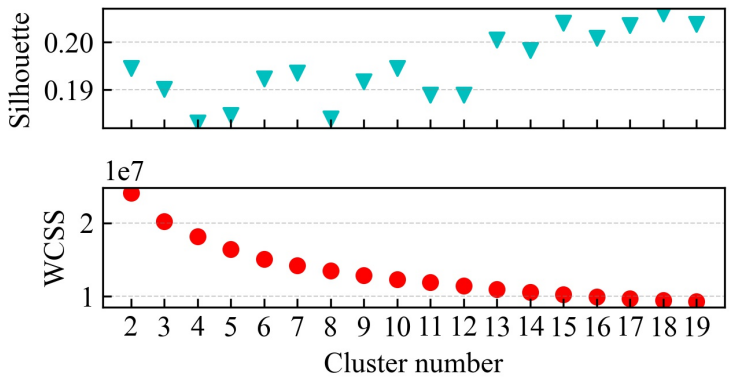


Figure 2.7: Cluster similarity index assessment of cross-building stock from electric-meter SAX motif counts.

settled count threshold. Fig. 2.6 presents the motif counts obtained across the stock, from which the most frequently observed motif count is aaaa, a constant low to null consumption steady-state sequence accounting for 12.58% of overall building stock SAX sequence counts.

To group buildings into similarly operating clusters we leverage the dimensionality reduction brought by the SAX transformation and alleviate the computational burden that would follow undertaking clustering on the original daily profiles. We consequently perform clustering on the motif sequence cumulated counts across the building pool. This allows to group buildings together based on their motifs distribution across the entire time horizon considered, while being very computationally light. A limitation of considering solely motif counts in the clustering process is however that the similarity between sequences is not accounted for, e.g., aaba will be considered as different from ccbc as aaaa, although it is naturally much closer to the later. While the authors are conscious of this limitation, it is beyond the scope of this work to develop a clustering method accounting for SAX sequences similarities. From the confirmatory analysis results presented in Fig. 2.7, we fix the optimal number of clusters as 6; a value showing a peak in silhouette score, indicating a slightly higher cluster cohesions, while displaying a sufficiently lowered WCSS and acceptably large number of clusters. The clustering results present the distribution of motifs across the obtained clusters under Fig. 2.8. Cluster 2 stands out as being composed solely of flat daily profiles from either aaaa, bbbb or cccc sequences. These electrical yearly consumption typically point to constant daily rule-based operationally controlled buildings, here mainly present in education, office, assembly and public type buildings, as Figure 2.9 shows. Clusters 3 and 5 behave quite similarly, with predominant flat profiles and low numbers of different sequences across the time horizon. Cluster 4 presents a variety of patterns yet with a clear predominant abcc sequence across the temporal horizon. Clusters 0 and 1 both show a diversity of profiles, although cluster 1 presents less variability in SAX counts, a likely consequence of it being less populated than cluster 0, the most populated of all 6 groups, collecting close to 400

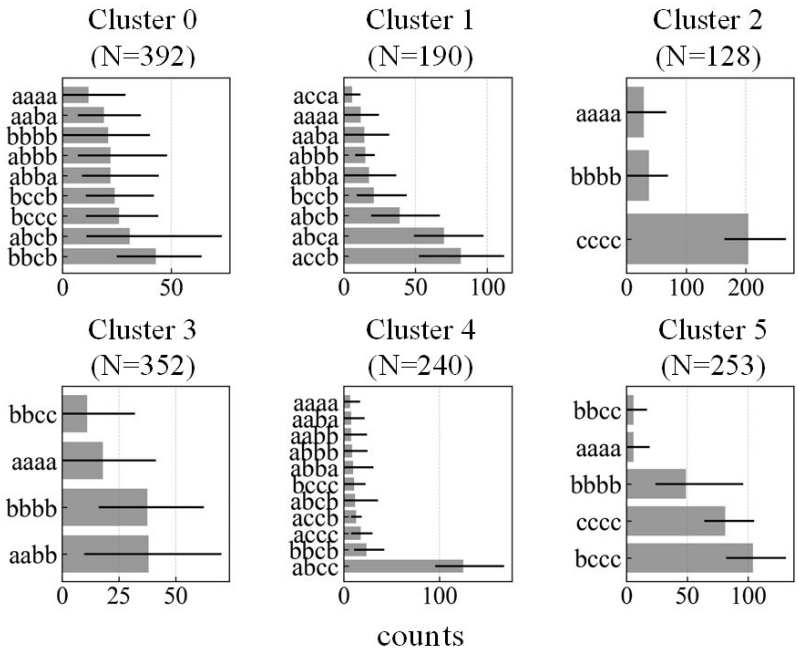


Figure 2.8: Building stock electric-meter SAX motifs distribution across clusters. Bar plots represent the sequence count median value while the error bars indicate the lower and upper quantiles.

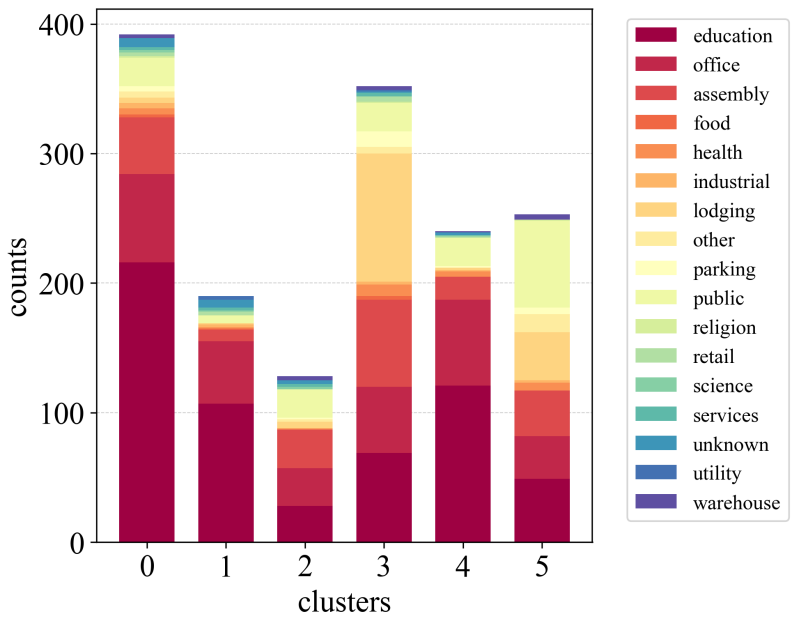


Figure 2.9: Building type distribution across identified clusters.

buildings. We turn our attention to cluster 0, the most populated cluster of the six, also presenting an interesting variety of motifs across both evaluated years. In the following sub-section, we explore a within-building analytical frame and switch to cuboid B for an in-site, bottom-up, analytical approach.

2.3.2 In-site view

This analytical frame follows more closely the presented DayFilter process of Miller et al., yet extends it with a multi-attribute temporal exploration through the time, attribute cuboid, further bridging the gap between top-down and bottom-up approaches. We explore buildings grouped within the afore-determined cluster 0, thus substantially reducing the initially considered cube-space, and iteratively slice the reduced cube by selecting individual buildings on which to perform automated pattern filtering. Time series are scaled within each attribute dimensional-frame, transformed to SAX sequences and filtered for motif identification using the same formerly presented process. Fig. 2.10 presents the sax-grouped daily profiles of

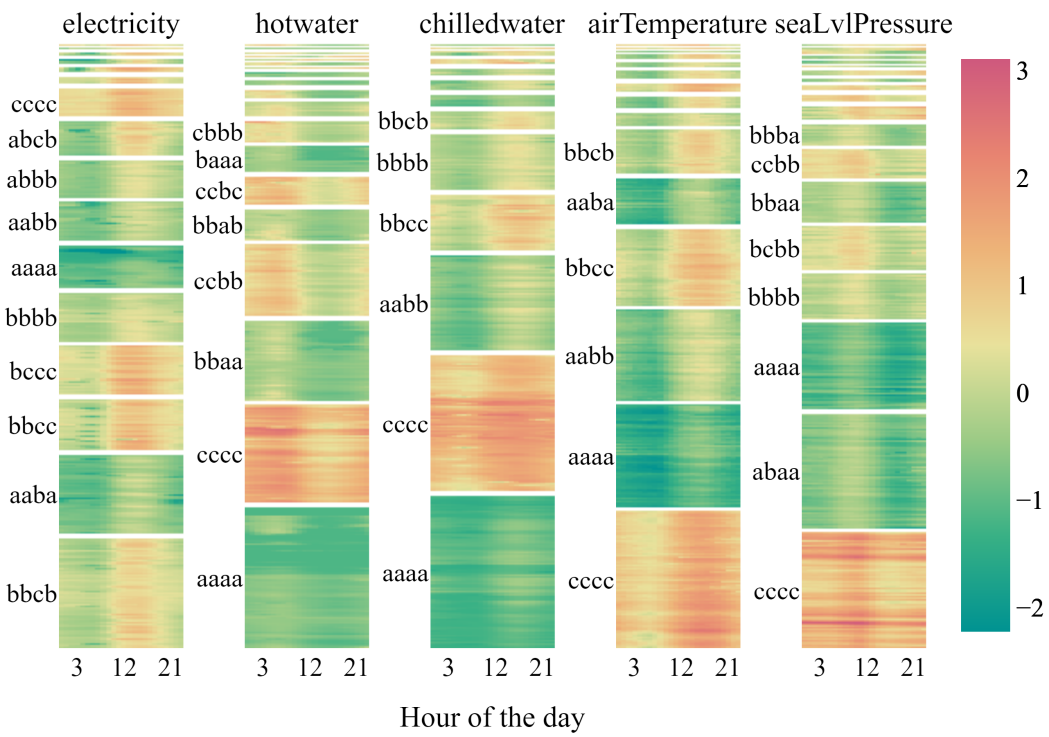


Figure 2.10: SAX sequences across Fow_education_Melinda attributes illustrated by daily heatmaps normalized per attribute. SAX motifs are explicitly referenced while discords are plotted but not tagged.

building Fow_education_Melinda in the form of a cascade heatmap, where motif sequences

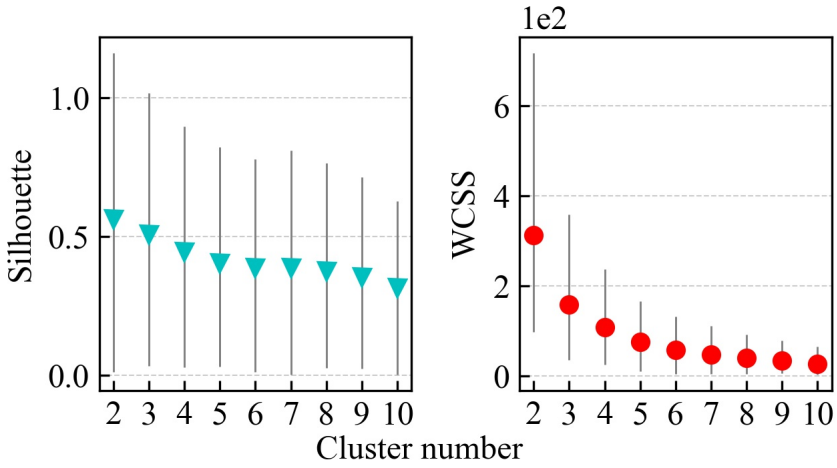


Figure 2.11: Cluster similarity index assessment of cross-attributes from Fow_education_Melinda building diurnal motifs. Scatter points illustrate median values of the evaluated similarity index across attributes, while the error bars cover the upper and lower quantiles, representing value variance.

are explicitly tagged. This allows impactful visual representation of the profiles, as well as their per-sequence group size, in a cross attribute context.

Links between distributions of patterns from one dimensional attribute to the other can thus be visually explored, e.g. two largest presented sequences within air temperature across the time horizon are cccc and aaaa (can also include aaba), referring quite straightforwardly to typical winter and summer periods, while chilled water possess two similar principal groups, i.e., cccc and aaaa, hinting to these identical seasonal periods. While this visualization display is powerful, the complexity endowed from exponentially increasing association possibilities between cross-attribute motifs can be limiting for human inspection. The dimensionality reduction provided by the ensuing clustering step takes up this problem, in an attribute dimensional-frame. After visual inspection of the confirmatory analysis' similarity indexes presented in Figure 2.11, we fix the cluster number across attributes to be 4. This value shows low WCSS range and norm while serving as an acceptable trade-off between a reduced number of pattern groups and sufficiently high group variety for detailed attribute pattern characterization. The clustering results depicted under Figure 2.12 exhibit close to homogeneous cluster sizes for air temperature alluding to the four seasons of temperate climate zones. Hot and chilled water meter patterns seem to behave in a mirrored fashion with consumption peaks and drops located in either mornings and evenings or evenings and mornings respectively. The electricity meter group size repartition seems closer to hot water consumption for this education building which both seems to testify on the building's operational activity; with three clusters presenting strong daily trends and one close to null consumption, hinting at weekend and holiday-type profiles. As we inspect the temporal

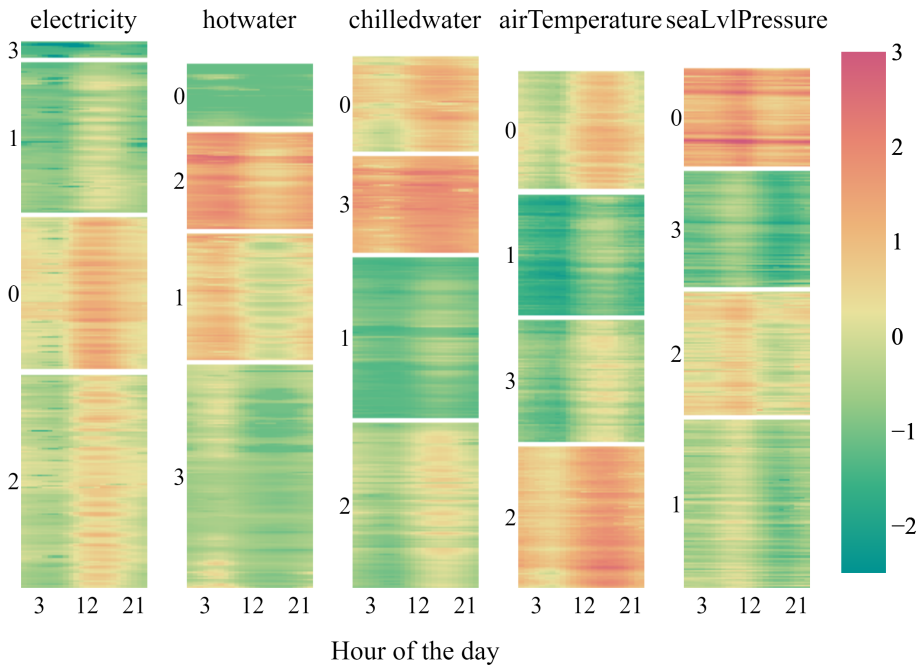


Figure 2.12: Dirunal pattern clusters across Fow_education_Melinda attributes illustrated by daily heatmaps normalized per attribute.

depth of the cuboid and how attribute patterns are cross-distributed, our final cube-space exploration step approaches temporal drill-in analysis, where the complexity of multi-site, multi-attribute dimensions in a set time-space frame is examined.

2.3.3 Temporal drill-in analysis

We investigate the site, attribute dimensions of cuboid C from iterative daily slices of the building data cube. Supported by the temporal cross-attribute exploration of the former cuboid B, we target the most represented cluster group within the temporal dimension, i.e. a typical day within the summer season, below illustrated by the selection of the day 2016-06-07. Selected daily profiles are z-scaled per attribute, across the building stock then SAX transformed, resulting in a singular daily sequence per cuboid dimensions. Given the lack of temporal depth of the sequences, notions of patterns and discords become meaningless along the time dimension. We therefore divert these notions to the site dimension, where buildings would be examined across their attributes as either behaving similarly to other buildings or not, i.e. motifs and discords respectively, given a certain threshold. We enumerate buildings displaying similar cross-attribute sequences and consider motifs for groups larger

Figure 2.13: Grouped building motifs SAX sequences across building stock and attributes on 2016-06-07. Group member counts are presented on the right-hand side by a bar chart, and daily aggregate attribute-specific consumptions are illustrated as heatmaps, echoing the classic OLAP approach..

than 5 members. Figure 2.13 presents aggregated daily attribute values annotated with SAX sequences and building group motifs member counts. From this cross-sectional view, it can be seen that the three most important aggregates possess only electrical meter data with SAX sequences of the three constant aaaa, bbbb and cccc profiles. Discord buildings are filtered out following which clustering can be performed from a weighted average of the daily multi-attribute time series.

Weights were designed to favor resource energy consumption data from weather conditions. In particular, electric meter was weighted as the preponderant attribute accounting for 70% of the time series weighted average, as reported under Table 2.1. The reduced one-dimensional

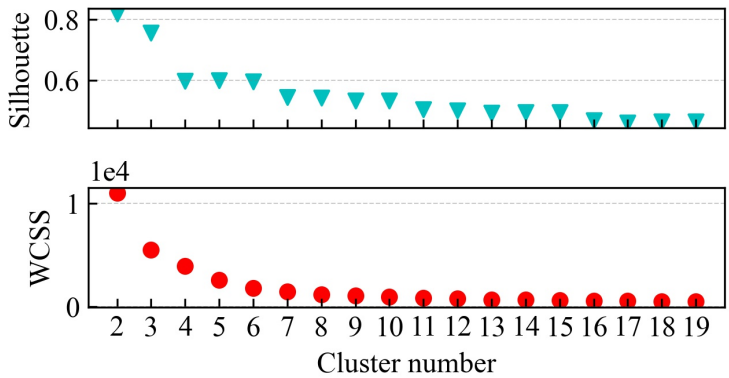


Figure 2.14: Cluster similarity index assessement of cross-building motifs stock from weighted averaged one-dimensional time-series from 2016-06-07.

Table 2.1: Attribute averaging weights.

Electricity	Hot water	Chilled water	Air temperature	Sea level pressure
0.7	0.1	0.1	0.05	0.05

daily-series are then clustered across sites. Selection of the optimal number of cluster is performed from visual inspection of the confirmatory analysis results presented under Figure 2.14. We select this number to be 4, given an over average silhouette score of 0.6 and a flattening WCSS trend. Clustering results are delineated under Figure 2.15 in the form of quantile-profile heatmaps for each attribute across the four obtained building clusters, granting a cross-site, attribute dimensional inspection of patterns. The larger building cluster aggregates a total of 813 buildings together while the smaller one only 19. Electrical patterns across the stock seem to follow overall comparable trends, with consumption increases and drops ranging between 6-10am and 7-9pm for their lower and upper quantiles respectively. This comes as a surprising read given the prevalence of the constant SAX sequences previously mentioned and could appear as a notable pitfall of the quantile heatmap visualizations, which solely show hourly quantiles across the cluster instead of original daily profiles. Quite similarly to the previously observed finding within cuboid B, chilled water appears to be positively correlated to outside air temperature across the building stock, with similar daily-temporal tendencies, i.e., lower morning values increasing from 10am, peaking around 2pm and decreasing in the evening with ranges from 6 to 9pm. Hot water, for the two smaller clusters, behaves in reverse to chilled water, with a prominent daily peak in the early morning, suggesting bathroom hot water consumption, while the larger N=813 building group possess a very flat to null consumption over the day, with faint lower and higher demands in the morning and evening respectively. Finally, the temperatures patterns across this selected day are quite typical of warm summer seasons with steep morning increases and smoother afternoon decays.

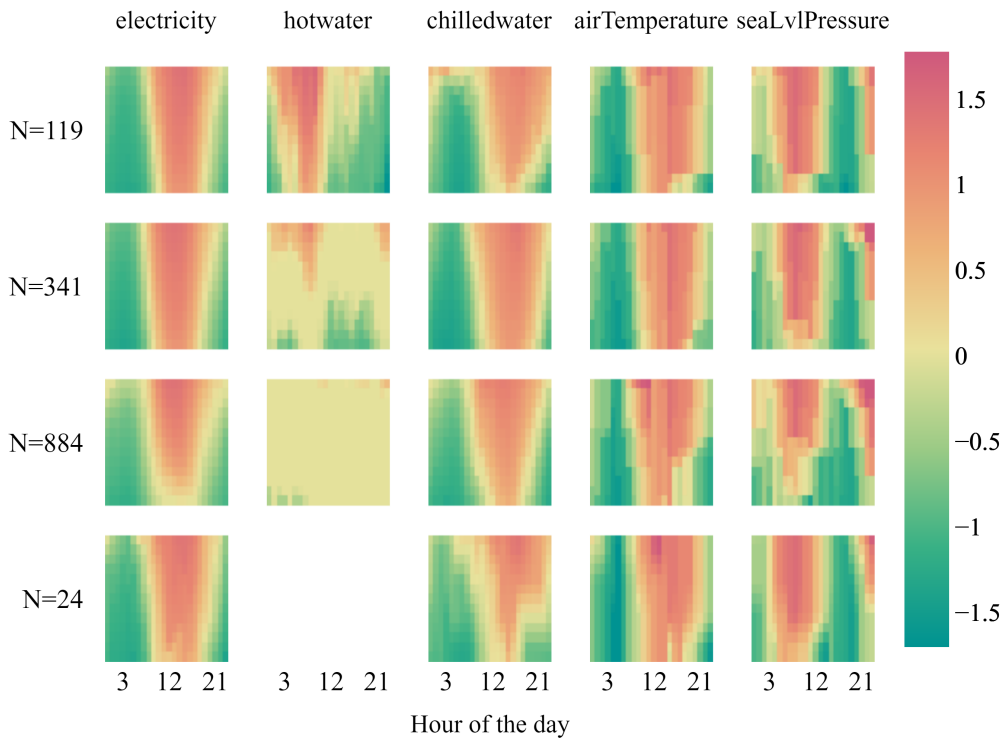


Figure 2.15: Motif-building clusters across attributes from day 2016-06-07 illustrated by quantile heatmaps normalized per attribute. Top and lower horizontal heatmap lines report the upper 75% and lower 25% quantiles per hour respectively, with a range of 10 gradually decreasing quantiles in between.

2.3.4 Towards multi-cube-space visualization

From the examined 2-dimensional lattice of the cube, we reach for a multi-dimensional visual exploration of the highly-dimensional 3D base-cuboid. Daily heatmaps have proven to be powerful visualization tools for 3-dimensional plots, yet the complexity endowed from base-cube visualization needs to be cut-down. To this end we propose combining visual insights from the three afore-examined cuboids to a recomposed, flattened, dimensional visualization of the cube; as if one were studying the cube's pattern rather than the assembled 3-dimensional structure. Figure 2.16 presents this multi-cube space visualization, where cuboid A, grouping site, time dimensions, is presented on the lower left corner, cuboid B with attribute, time in the top right corner and cuboid C, gathering site, attribute links both visuals from aligned dimensional sections. Additional rehashed insights from cuboids including time dimensions were supplemented with aggregated temporal outlooks, here illustrated with barplots resuming the SAX sequences across the temporal study frame. This

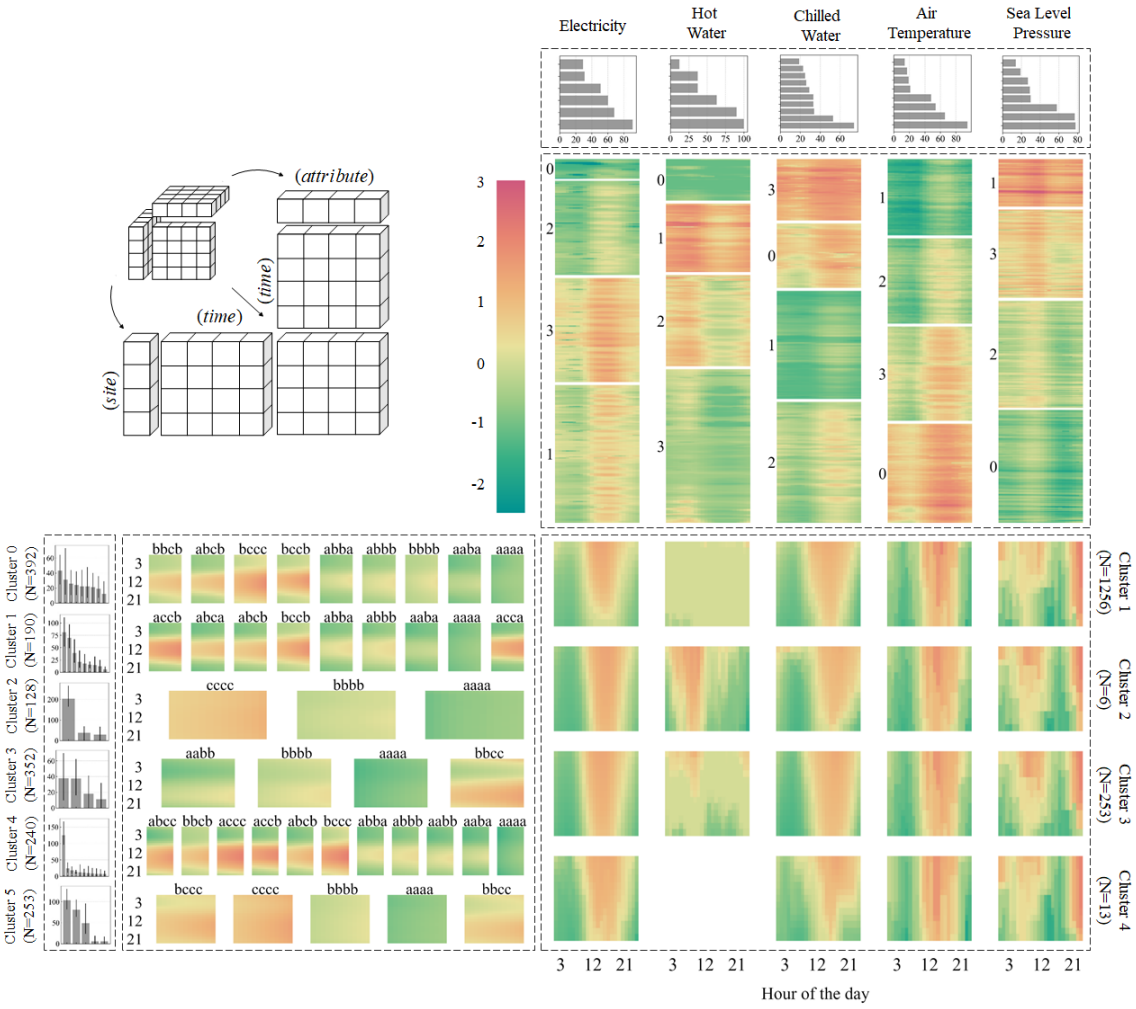


Figure 2.16: Data cube base-cuboid pattern visualization from 2-dimensional cuboid lattice insights, i.e., {site, time} bottom left, {attribute, time} top right and {site, attribute} bottom right..

multi-dimensional view allows distinct knowledge transfer and analytical examination from one dimensional and diagnostic-specific study frame to the next. For instance, while the electricity SAX sequence distribution of cluster 0 within cuboid A should echo that of cuboid B, a building-element subset of cluster 0, the sequence distributions presented are quite different from one another. This stems from the differences in pre-mining normalization frames, where cuboid A scaled electricity consumption over the entire building stock, while cuboid B considered a fixed site selected subset, consequently resulting in different alphabetical ranges and breakpoints during the SAX transformation process. Yet similar heatmap tendencies may be observed from one cuboid view to the next, i.e., both possess clear summer and winter typical consumption trends with a flat weekend-like aaaa consumption profile. This highlights the importance of per cuboid diagnostic focus; as data analytic choices might be relevantly

made for isolated cuboids, mining result comparisons from one cube sub-space to the next ought to be treated cautiously, as a result of different cuboid-specific mining steps. For a common and global multi-dimensional analytical diagnostic, it becomes necessary to follow identical mining tasks at every step of the process. For this work, the importance of framing insight-specific steps was chosen to further highlight the significant role of dimensional-frame determination within the process of cube space mining.

2.4 Discussion

From the definition of a unified multidimensional data mining framework tailored to building analytics, this work intends on bridging the gap between the complexity endowed with big data's high-volume, high-variety and progress towards more interpretable and reproducible research for building analysts. The objective is to link applications to specific diagnostic approaches from dimensionally-reduced cube-space regions. In this context, results of the proposed mining framework implementation are here discussed while considering other possible applications as well as limitations encountered.

2.4.1 Insight-driven mining

On the road towards more interpretable building analytics, definition of the cube dimensional space linked to application-driven insights per cuboid sub-regions has demonstrated great value. From the exploration of the 2D cube lattice, we have covered the established preeminent analytical methods, namely bottom-up site, time and top-down attribute, time approaches, all the while extending them with a temporal drill-in site, attribute analysis. While an identical descriptive pattern filtering mining technique was applied over the lattice, we meet each cuboid with a different analytical angle and diagnostic-objective. It then becomes interestingly relevant to contemplate the more complex analytics that would arise approaching the last, most dimensionally-dense, base-cuboid site, attribute, time region of the cube. Given the previously defined analytical methods, one could imagine tackling this cuboid from three subsequent angles, i.e., including either multiple attributes, sites or temporal-units of interest within the existing analytical frames of cuboids A, B and C respectively. As a conceptual illustration, approaching the base-cube from cuboid A, would involve a classical top-down analytical approach of building benchmarking extended through multi-attribute considerations. While descending from cuboid B, through the antagonist bottom-up approach, would imply in-site diagnostic methods extended to other buildings, e.g., testing a methods' scalability. The temporal drill-in analytical method of cuboid C, lastly, would examine additional days within its frame, adjusting time-specific insights to a larger temporal frame of interest. Additionally, while our implementation depicts a descriptive mining technic, predictive applications share equal benefits from cube space conception. Indeed, how to effectively evaluate and select large number of feature, for example, fit naturally within OLAM supported by explicit cube dimensional-space mapping.

Assessing contributions of feature combinations to the predictive learning performance over the cube-space, allows systemic optimal feature selection in the confirmatory analysis phase. Machine leaning workflows could incorporate such techniques as an a priori mining analysis to improve model performance. Employing pre-trained models within cuboids are another example of how cube space-driven mining can be practiced in predictive applications [108]. Investigation of this application, while outside the scope of this study, reveals a promising future direction for this framework. It subsequently becomes clear that the mining process is fully application- as well as insight-driven. Applications such as energy performance benchmarking and model calibration compel to top-down approaches, while automated fault detection and diagnosis, energy saving management, or rule-base knowledge discovery entail classical bottom-up approaches. Likewise, temporal feature engineering can necessitate temporal drill-in methods for, per time-slice, cross-attribute, -building insights. These connect reduced cube-space regions of interest to undertaken applications.

2.4.2 Visualizing knowledge

The importance of knowledge visualization for effective and impactful result inspection is well established. However, when it comes to high-dimensions it becomes particularly complex, yet crucial to appropriately represent and link insights together. Interactive OLAP visualization tools have already been developed and widely used for data cube exploration, analysis, and pattern extractions in the financial field [121], but, to the best of the authors knowledge, close to none in the building sector. The proposed 3 dimensional data cube-pattern visualization paves the way to the development of OLAM interactive visualization tools, where one could imagine iteratively scrolling through the fixed dimensional items of a cuboid. The building analyst could subsequently employ navigational tools such as drilldown or rollup, through the dimensional hierarchical relationships, e.g. the day slice width considered in cuboid C could be rolled-up to weekly slices or drilled-down to quarter days for SAX sequence analysis.

2.4.3 Limitations

A notable limitation encountered from cube space mining was the iterative need to reformat the data as well as adapting visualization tools to every studied dimensional frame, which are very time-consuming tasks within the data mining process. In then comes into consideration that developing interactive visual tools tailored to OLAM analytics could provide interesting solutions, yet not without challenging limitations. Computational burden resulting from the mining process may render the interactivity of the visual exploration too slow to fully profit from the tool itself. Nonetheless, a priori computation of the visual cube from a set of fixed parameters could be envisioned as a means to initially coarsely characterize the cube and tackle exploratory responsiveness issues.

2.5 Summary

With this work, we have delineated a multi-dimensional, generic data mining framework tailored to big building data, effectively framing which analytical techniques to follow in a step-wise procedure. We appeal to benchmarking methods and apply the proposed DM framework to an automated pattern filtering application using a large building open data set for reproducible, comparable, and empirically validated results. Furthermore, we delineate the existing underlying link between building data dimensional space and building management applications. This pushes further down the existing barriers separating building professionals from effective building data dimensional-space targeting given defined applications and insights of interest. Future research challenges could entail in-depth cube space exploration for comprehensive building management application studies such as multi-automated fault diagnosis and detection. Another interesting research focus emanating from this work could undertake the determination of how dimensional analytical window frames, i.e. data granularities, window frame, and horizon, influence building data analytics and their outcome.

Chapter 3

Scalable building heat dynamics identification

Chapter overview

- Building thermal model identification method
- Automated method extension suitable for building stock scalability
- Case study: 247 Dutch residential buildings smart-thermostat data (2NECO)
- GitHub repositories:
  - Open implementation - /fiftyshadesofgrey
  - Open data set `Grey-brick buildings` - /greybrickbuildings

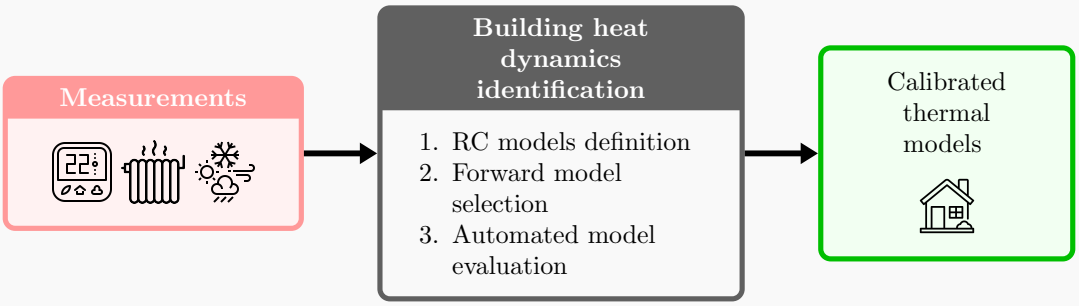


Figure 3.1: Overview of the model identification method

The content of this chapter is building upon material published in [122], [123].

"All models are wrong, but some are useful."
George E. P. Box

3.1 Preface

As briefly introduced in the Introduction, the building sector has embraced data as the new fuel to harvest, at scale, the power of building performance modeling. This grants valuable insights into the dynamics driving the energy demand of buildings. The building sector has investigated multiple strategies over the last decade to reduce, adapt and better anticipate its energy load on the power network. Well-established techniques such as building

retrofitting [13], demand side management [16], energy forecasting [15], and building to grid energy management schemes emerging from model predictive control [17] or reinforcement learning [124], have, and still are, at the center of a considerable amount of attention from both research and industry. All, however, require knowledge of the building thermal dynamics in order to effectively perform, consequently placing our ability to effectively identify building thermal behavior(s) as the backbone of building applications. Yet, despite its momentum, building modeling is still faced with the fundamental challenge of scaling across the heterogeneous building stock, and relies primarily on assumptions rather than field performance data.

3.1.1 Modeling methods

Main existing modeling methods can be divided in three preeminent categories: physics-based methods (white-box), purely data-driven (black-box) and hybrid approaches (grey-box) [125]. The first, physics and knowledge-based models, solves mathematical equations based on physical laws to characterize the energy behavior of buildings. They require exhaustive information on the building and are usually mathematically complex. Yet, they can yield high accuracies if calibrated correctly and are often employed in building performance simulation softwares, e.g., Energy-Plus [126], or using powerful modeling languages, e.g., Modelica [127]. White-box modeling, however, is time-consuming with performances largely depending on accurate energy model calibration consequently making it difficult to scale up. It requires the definition and update of many input parameters along the building's lifetime [128], a process reliant on expert analysis that needs repetition for every considered building. Moreover, their copious amount of parameters makes white-box models non-structurally identifiable, which often becomes problematic when an unknown parameter needs estimating. The second category constitutes data-driven models, often referred to as black-box models. They consist of statistical regressions and machine learning algorithms typically fitted on the input and output time-series data of the system. Its 'black-box' analogy stresses the relationship between model input and output as being hardly transposable to physics-based analysis, making it challenging to produce interpretable models [125]. While significant developments in this field might alleviate the persisting barriers of domain knowledge inclusion or interpretability, progress is still desperately needed for trustworthy and scalable applications within the building sector. Lastly, machine learning approaches require large amounts of quality data to guarantee satisfying accuracies of models from training. This implies data consistency, assessing coherent matching of various attributes, data completeness (no missing values) and accuracy (absence of outliers) [129]. Finally, grey-box models work as a hybrid approach between the aforementioned data-driven and physics-based models. This approach profits from dominant physical properties of the system to build the model structure while employing measurements to fix the model parameters. A common approach to modeling building heat dynamics adopts lumped resistance-capacity models, i.e. RC models, resulting in an electric circuit representation of the thermal conditions of the building

[130]. In this way, grey-box approaches capitalize on the inclusion of physical knowledge in their models. This results in smaller amounts of required experimental data to train the model compared to black-box, thus making grey-box models better at generalization while staying interpretable [131].

3.1.2 Challenges

A common problem with model identification lies in finding a model in agreement with both the physical reality as well as the level of information embedded in the data, meaning the model should avoid both under-fitting and over-fitting the measurements. To tackle this, Bacher and Madsen [132] suggested an extensive stochastic model identification procedure to identify building heat dynamics from numerous RC models of different orders. Models were evaluated based on likelihood ratio tests and selection procedure was carried out through significance improvement evaluation from simpler to more complex models, to avoid over-fits in the model selection phase. It was argued by Yu et. al. [128] that first- and (simple) second-order models are sufficient to capture the thermal dynamics of buildings to fend off the aforementioned problem. The research foundation for this claim is, however, built upon findings emerging either from simulated data sets or fairly simple and isolated single-building measurements. Our intuition would argue that to determine the dynamics of real-world occupied buildings from measurements, larger model orders are not only relevant but necessary to encapsulate the, often different, thermal inertiae of buildings and bring forth most-needed comprehensive thermal behavioral insights. Assuming low-order models without the consideration of higher-order ones is, within our field, a judgmental bias that desperately needs tackling. As of today, there is very little work studying building model identification from large and occupied building stocks. Hossain et. al. [133] notably evaluated the performance of Bayesian neural networks for $n\mathbf{R}n\mathbf{C}$ grey-box thermal model identification from 8'834 Canadian homes with 3 months worth of data. Their study demonstrated the value brought by transfer learning for smaller available building data sets as well as the overall better performance of their Bayesian approach to other black-box models based on root-mean-squared error-metric. Yet, R and C parameters of the fitted models could not be uniquely identified. This, prevents the physical interpretation of parameters in the model evaluation phase, a step most studies do not comment on, along with the identifiability of their assumed model structures as mentioned by Deconinck and Roels [134].

3.1.3 Motivation

This chapter proposes to put forth an automated model selection and evaluation procedure for stochastic model identification of building heat dynamics, providing a much needed scalable method tailored to the existing heterogeneous residential building stock. Leveraging the procedure proposed by Bacher and Madsen [132], RC models of rising complexity are evaluated over 247 Dutch residential buildings. Identified thermal parameters are then

examined and employed to support building envelope performance analysis; providing large-scale, non-intrusive insulation insights into the existing building stock. Important application perspectives to the approach are finally provided illustrating the impact of building model identification from measurements at scale.

3.2 Methodology

This section first describes the formulation of stochastic differential equations for building heat dynamics modeling. Evaluated RC models are then detailed followed by the automated model selection and evaluation procedure.

3.2.1 Grey-box models of a dynamic system

Using prior physical knowledge as well as information embedded in data, grey-box models are established by sets of partially observed first-order stochastic differential equations, also referred to as stochastic linear state-space models in continuous-discrete time. These equations describe lumped RC models of the heat dynamics of the building. Typically building thermal models consider the heat exchanges between inside and outside conditions, i.e., temperature differences, solar radiation gains but can also include long wave radiations, as well as convection and infiltration driven by wind-speed if available data permits it [135]. The building envelope consequently embodies the most crucial component of the model, regulating heat transfers between these two environments. Further, diverse indoor components such as space-heating inertia, measurement errors present in the input variables and additional building zones are modelled using additional temperature state points.

We refer to the work of Bacher and Madsen [132] for the developed models as well as their evaluation and selection procedure. A typical first-order stochastic differential equation can be expressed as

$$dT_i = \frac{1}{R_{ia}C_i}(T_a - T_i)dt + \frac{1}{C_i}\Phi_h dt + \frac{1}{C_i}A_w\Phi_s dt + \sigma_i d\omega_i \ , \tag{3.1}$$

where T, R and C refer to temperature, resistance and capacitor respectively, while Φ is an energy flux and A_w the window area. The subscript i points to the inside temperature, while a refers to the ambient temperature, h to the heater and s to solar. ω describes a standard Wiener process, and σ stands for the incremental variances of the Wiener process which encapsulate model approximations and non-recognized or modeled phenomena [136].

This physical part of the model is coupled to a data-driven one, linking the data observations to the model for parameter estimation. The discrete time measurement equation is

$$Y_t = T_{i,t} + e_t \ , \tag{3.2}$$

where t is the measurement point in time, Y_t the measured interior temperature and e_t is the

measurement error [136], assumed to be Gaussian white noise as the fitted model accurately captures the dynamics of the system. With observations represented by

$$\mathcal{Y}_N = [Y_N, Y_{N-1}, ..., Y_1, Y_0], \tag{3.3}$$

the maximum likelihood estimates of the grey-box model parameters can be identified from the joint probability density [137]

$$L(\theta; \mathcal{Y}_N) = \left(\prod_{k=1}^{N} p(Y_k | \mathcal{Y}_{k-1}, \theta) \right) p(Y_0, \theta), \tag{3.4}$$

where $p(Y_k|\mathcal{Y}_{k-1}, \theta)$ represents the conditional density designating the probability of observing Y_k given the preceding observations and the parameters θ, and where $p(Y_0, \theta)$ is a parameterization of the starting conditions. This is done by introducing an expected value of the initial states and the associated covariance matrix. Maximum likelihood estimates of the parameters can then be found from

$$\hat{\theta} = \underset{\theta}{argmax}\{L(\theta; \mathcal{Y}_N)\}, \tag{3.5}$$

which can be calculated using an optimization algorithm over a Kalman filter. We refer to the work of Kristensen et al. [137] for a detailed description of the approach.

Applied models

This study considers grey-box models ranging from the simple first order T_i model, explicitly described in Eq. (3.1), where the inside temperature state-point T_i and its RC parameters R_{ia} and C_i are solely treated, to 5th order ones, where the addition of sensor T_s, medium T_m, heater T_h and building envelope T_e state points along with their respective RC parameters each add a variety of model extensions to choose from. Additionally, the building envelope component proposes additional parameter extensions modeling direct inside to outside heat exchanges and facade solar gains, which are here considered as a block extension $A_e R_{ia}$.

The full model $T_i T_m T_e T_h T_s A_e R_{ia}$ is visually displayed in Fig. 3.2 and the set of stochastic differential equations describing its heat flows in continuous time are [132]

$$\text{Sensor: } dT_s = \frac{1}{R_{is}C_s}(T_i - T_s)dt + \sigma_s d\omega_s \ , \tag{3.6}$$

$$\begin{aligned} \text{Interior: } dT_i = & \frac{1}{R_{is}C_i}(T_s - T_i)dt + \frac{1}{R_{im}C_i}(T_m - T_i)dt \\ & + \frac{1}{R_{ih}C_i}(T_h - T_i)dt + \frac{1}{R_{ie}C_i}(T_e - T_i)dt \\ & + \underbrace{\frac{1}{R_{ia}C_i}(T_a - T_i)dt}_{R_{ia} \text{ component}} + \frac{1}{C_i}A_w \Phi_s dt + \sigma_i d\omega_i \ , \end{aligned} \tag{3.7}$$

$$\text{Medium: } dT_m = \frac{1}{R_{im}C_m}(T_i - T_m)dt + \sigma_m d\omega_m \ , \tag{3.8}$$

$$\text{Heater: } dT_h = \frac{1}{R_{ih}C_h}(T_i - T_h)dt + \frac{1}{C_h}\Phi_h dt + \sigma_h d\omega_h \ , \tag{3.9}$$

$$\begin{aligned}\text{Envelope: } dT_e = &\frac{1}{R_{ie}C_e}(T_i - T_e)dt + \frac{1}{R_{ea}C_e}(T_a - T_e)dt \\ &+ \underbrace{\frac{1}{C_e}A_e\Phi_s dt}_{A_e \text{ component}} + \sigma_e d\omega_e \ ,\end{aligned} \tag{3.10}$$

where the subscripts s, m, and e point to sensor, medium and envelope components respectively. The discrete time measurement equation is

$$Y_t = T_{s,t} + e_t \ , \tag{3.11}$$

as observed temperature is encumbered with measurement error. We refer to an applied model as a combination of its corresponding state-point(s) T_x component and block model component extension A_eR_{ia}, highlighted in the under-brackets of Eqs. (3.7) and (3.10). For example, the model $T_iT_eT_h$ comprises the first order model T_i with envelope T_e and heater T_h component extensions, but without the inclusion of the A_eR_{ia} block extension of the envelope. For a detailed description of the models, the reader is suggested to refer to the work of Bacher and Madsen [132].

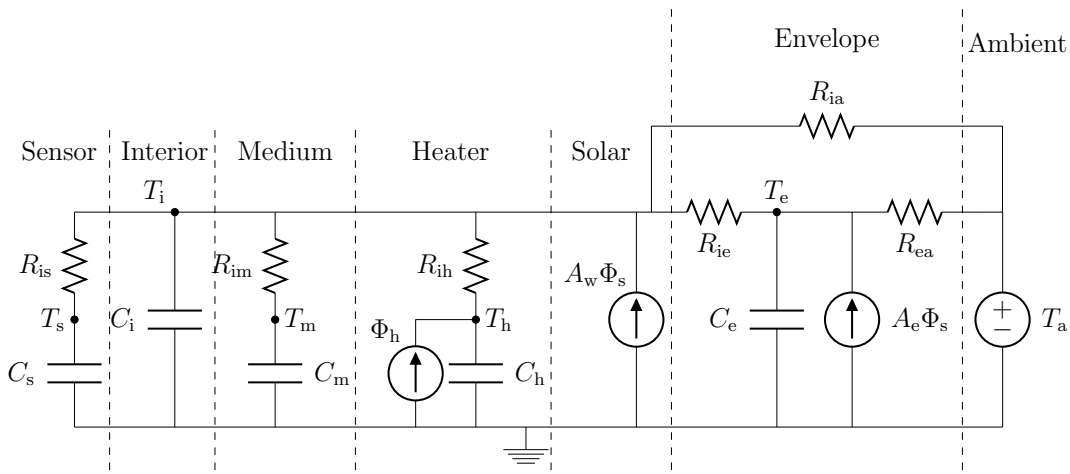


Figure 3.2: The full model $T_iT_mT_eT_hT_sA_eR_{ia}$ with all considered model extension parts presented and individually indicated. The model consequently depicts all parts included in any of the other applied models. Reprinted from the work of Bacher and Madsen [132] with the authors approval.

It should be noted that while our approach proposes to apply the RC models put forward by Ref. [132], our proposed automated model selection and evaluation procedure, describe in the following subsection, can be applied to any set of increasing complexity of grey-box models.

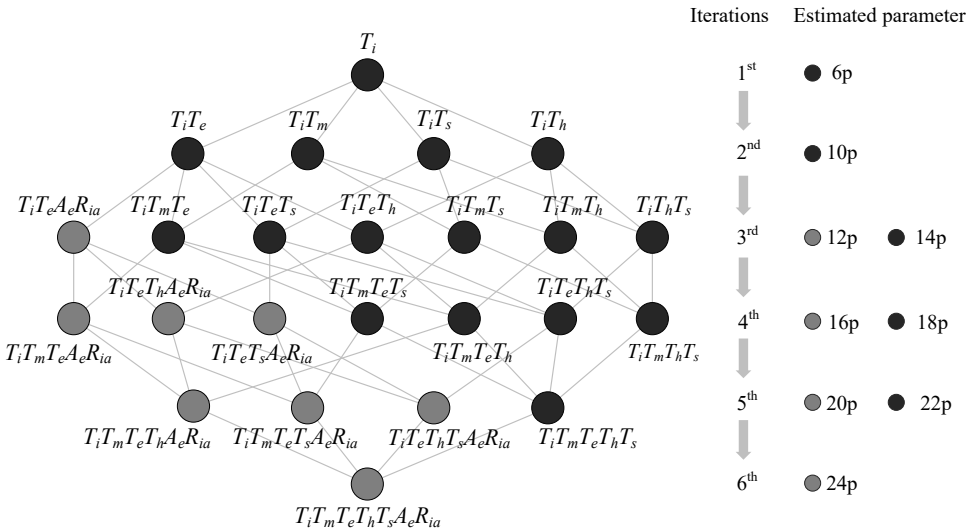


Figure 3.3: Forward model selection scheme

3.2.2 Automated model selection and evaluation procedure

The scaling up of grey-box approaches for automated model selection can be challenging. Initial values of parameters are usually tuned to the case study from expert inspection while model extensions are iteratively built and evaluated. Here we present the proposed model selection and evaluation procedure.

Model selection

The model selection procedure employs a likelihood ratio test to statistically determine whether a more complex model performs significantly better, or not, compared to a simpler, sub-model. Likelihood ratio tests are particularly effective at comparing two competing statistical models with no unknown parameters and have been demonstrated by the Neyman-Pearson lemma to have the highest statistical power amongst all other contestants [138]. This implies that the test is able to make the most efficient use of the available data. A forward selection procedure is proposed beginning with the simplest feasible model, T_i, and extending it iteratively with the component presenting the most significant improvement. The procedure terminates when no model extension yields a p-value below the pre-specified limit, commonly fixed at 5%. Possible candidates for model improvement are selected from a set of predefined extensions, resulting from the combination possibilities of the different considered model components, i.e., T_e, T_h, T_m, T_s, A_eR_{ia}. Figure 3.3 presents the overall model selection scheme. Possible model combinations are mapped and linked, visually exposing the different existing paths of the forward selection procedure. The process has

been adapted from [132] to assure more coherent ranges of evaluated parameters within each iteration phase. We refer to this model selection method as *fms1*.

Capturing occupant-driven heat gains - An additional degree of freedom to the model selection phase was later included to capture aggregated heat gains from hot water usage. Indeed diverse dynamics linked to hot-water demanding appliances, e.g., kitchen or laundry, were suspected of having a significant influence on the thermal dynamics of these buildings. In this model selection variation, the first-order model T_i is estimated under three competing heat input signals:

1. the original *central heating* set-point temperature, comprising space heating demand,
2. the new *hot-water* set-point temperature, representing hot-water heat demands, and
3. the aggregated *boiler* set-point temperature, which englobes both space heating and hot water heat demands.

The heat input signal producing the more likely model is then selected, following which the forward model selection carries on. We refer to this model selection variant as *fms2*.

Model evaluation

Finally, we extend the existing model evaluation phase of [132] to render the process suitable for automated deployment. This last step leverages the commonly employed qualitative appreciation of model fits from cumulated periodograms of the residuals. A periodogram, or sample spectrum, is obtained by Fourier transforming the autocovariance function of a stationary process [139]. Typically, an appropriate model will produce residuals with Gaussian white-noise properties, which in the frequency domain, denotes a theoretical constant periodogram [139]. Observing whether obtained model residuals are located around this straight line, e.g. within a surrounding confidence interval, consequently serves as an appropriate indicator of a model's quality.

By calculating the frequency differences between a selected model's Cumulated Periodogram (CP) and its confidence interval, we obtain boundary excess values which, in turn, can be summed into a unique numerical indicator, i.e., the Cumulated Periodogram Boundary Excess Sum (CPBES). This indicator characterizes the amount of auto-correlation present in the considered input time-series, which implies white noise properties when close to zero. CPBES consequently allows the differentiation of poor, suitable and good models resulting from the previous forward selection procedure. To allow fair comparisons of CPBES between times series of different lengths we normalize it by length and obtain the normalized CPBES (nCPBES).

Quite concretely, the CP is calculated from the normalized sum of the Discrete Fourier

Transform (DFT) of the time series

$$\mathrm{CP}(k) = \frac{1}{K}\sum_{j=1}^{k}\left|\mathrm{DFT}(x_j)\right|^2. \tag{3.12}$$

Here x denotes the input time series, k the considered Fourier frequency of the periodogram, and K is the last Fourier frequency of the domain, which corresponds to the times series length, i.e. N, minus one. The periodogram confidence interval, or boundaries (B), are obtained from the Kolmogorov-Smirnov test for distributions at a given probability $(1-\alpha)$ [139]. The obtained bounds are characterized from the slope and intersection coefficients

$$\mathrm{B}_{intersect} = \sqrt{2}\cdot K_\alpha \cdot \left(\frac{K-1}{2} - 1\right)^{-1/2}, \tag{3.13}$$

$$\mathrm{B}_{slope} = 2T, \tag{3.14}$$

where K_α is equal to 1.358 for confidence intervals of 95%, and T corresponds to the period, or the frequency inverse, of the input time-series [140]. Finally the nCPBES can be determined from:

$$\begin{aligned}\mathrm{CPBE}(k) = &\overbrace{\max\Big(0, \mathrm{CP}(k) - \mathrm{B}_{slope}\cdot k - \mathrm{B}_{intersect}\Big)}^{\text{top boundary excess}}\\ &+\underbrace{\max\Big(0, \mathrm{B}_{slope}\cdot k - \mathrm{B}_{intersect} - \mathrm{CP}(k)\Big)}_{\text{bottom boundary excess}},\end{aligned} \tag{3.15}$$

$$\mathrm{nCPBES} = \underbrace{\frac{1}{K}\sum_{k=1}^{K}\mathrm{CPBE}(k)}_{\mathrm{CPBES}}. \tag{3.16}$$

Figure 3.4, illustrates how nCPBES is obtained from a given cumulated periodogram. It should be noted than while nCPBES is suited for automated model selection as a unique numerical indicator, its cumulated periodogram or derived boundary excess curve still present valuable qualitative information, indicating the frequencies of the dynamics the model is not capturing. These can be employed for in depth model analysis a posteriori to the automated model selection process. For instance, the boundary excess curve of Fig. 3.4 presents two lumps located around frequencies of 0.2 and 0.4 (2/h) which correspond to periods of 10 and 5 hours respectively. This allows the analyst to identify the frequencies of the dynamics still present in the residuals which can later drive the design of model extensions or suggest the need for additional measurements.

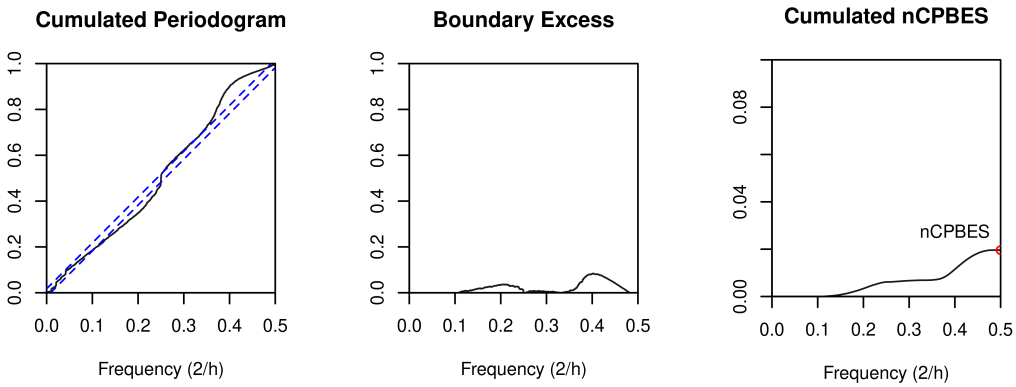


Figure 3.4: Model residual cumulated periodogram transformed to normalized Cumulated Periodogram Boundary Excess Sum (nCPBES) for automated model evaluation.

3.3 Implementation

Our study considers a total of 247 homes located in the Netherlands, a European region under the Köppen climate classification index [141] *Cfb* which describes mild temperate, fully humid, and warm summer regions. Anonymized measurements are collected from the Toon smart thermostats proposed by the energy distributor Eneco. In the remainder of this thesis, we refer to this case study as 2NECO. Measured data include inside temperature and boiler (heater) set-point temperature, at granularities of 15-minute intervals. Few building meta-data are made available by users as self-reported information such as building type, floor surface, and family size.

Boiler set-point temperature can here be considered to act as the centralized space heating input signal of buildings. Indeed, centralized heating systems of dwellings are commonly operated by adjusting delivery temperatures, i.e., measured (boiler) set-point temperature, while a pump maintains a constant pressure across the building's pipelines. This setup implies eventual non-homogeneous power outputs throughout radiators, should their combined valve positions be readjusted, even with fixed boiler set-point temperatures. Commonly, though, radiator valves are set to fix positions by building occupants, and house temperature is adjusted directly from the thermostat. Such variations can therefore be considered negligible and the boiler set-point a good indicator of space heating input signal. The input signal information ϕ_h is included in the models with $\Phi_h = r \cdot \phi_h$, where r is a scaling factor parameter calibrated from maximum likelihood estimates to adjust the input signal information to the space heating needs of the building.

Hourly weather data are collected from the publicly available Royal Netherlands Meteorological Institute (KNMI) weather stations [142]. Stations are paired to each building thanks to a geo-localization process using 4 (over the 6) ZIP code digits; an aggregation level that

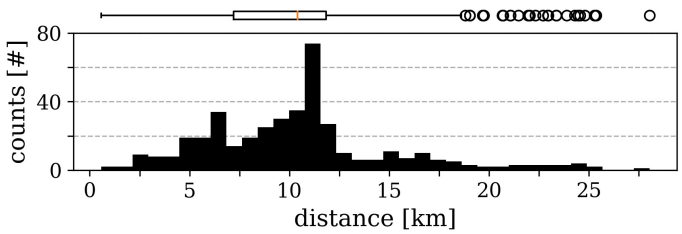


Figure 3.5: Distance distribution between building province and its nearest KNMI weather station, showcasing both a boxplot (top) and histogram (bottom) for a more informed distribution appreciation.

allows no anonymized user to be geographically isolated nor identified. Figure 3.5 shows the distribution of distances between the building's province and its nearest KNMI station. Obtained distances are centered principally between 5 and 12 kilometers, while a few larger distances can be found above 20 kilometers. While these measurements cannot encapsulate local weather conditions particular to micro urban surroundings, they provide a sufficient approximation of building outside conditions.

To capture the thermal dynamics of a building, we consider minimum measurement periods of 2 months and limit the maximum times-series horizon to 4 months. We filter available data to obtain the most recent continuous measurement period for each building resulting in periods ranging from February 1st to the end of May 2021, which holds a notably cold start of spring season at the beginning of April. Weather data are re-sampled to 15-minute resolutions to fit thermostat measurements. While finer granularities, typically 1 or 5 minutes, are better suited to capture the thermal dynamics of building systems, 15-minute resolutions are sufficient to do so. Cumulative missing values larger than 2 hours are imputed while smaller gaps are interpolated via a moving average using a window size of 8 hours. Table 3.1 summarizes the resulting data set characteristics.

Table 3.1: Data set characteristics

Input measurements	T_i, Φ_h, T_a, Φ_s	
Resolution	15 minutes	
Period of measurement	*fms1*	01.02.2021 - 31.05.2021
	fms2	01.02.2021 - 30.11.2022
Input time series length	2 months $\leq$ input $\leq$ 4 months	

Grey-box models are implemented using the computer software CTSM-R [143] developed at the Technical University of Denmark. It produces maximum likelihood estimates of model parameters thanks to an optimization algorithm performed over a Kalman filter. The code developed for this study is made available under an open-access github repository, i.e. https://github.com/JulienLeprince/fiftyshadesofgrey, to encourage dissemination and

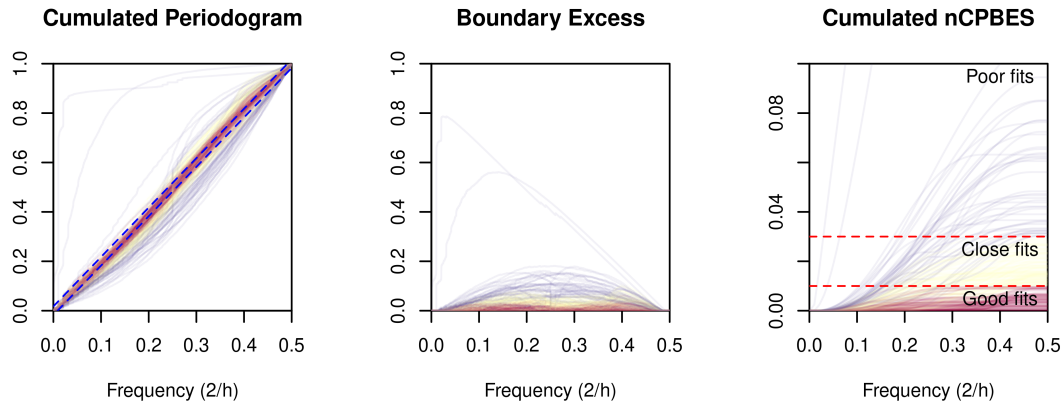


Figure 3.6: Models residual cumulated periodograms, boundary excess curves, and cumulated nCPBES of *fms1*

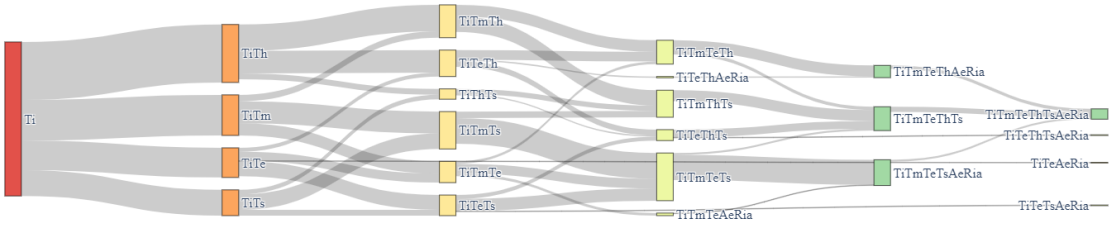
support its reproducibility.

3.4 Results and Discussion

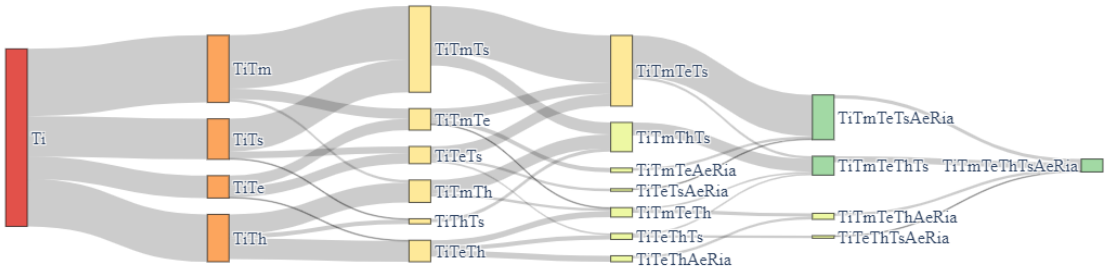
We describe the outcome of the implementation section here with a first model evaluation subsection. Estimated model parameters and highlighted links to building characteristics are then presented, followed by a final illustrative building performance application, leveraging model parameter estimates for city-scale building envelope insights.

3.4.1 Model evaluation

The normalized cumulated periodogram boundary excess sum indicators are used to differentiate good from close and poor model fits. Figure 3.6 presents the cumulated periodogram, boundary excess, and nCPBES of all final models obtained from the forward selection procedure *fms1*. After an attentive inspection of the CPs and their respective nCPBES, threshold values of 0.03 and 0.01 nCPBES were fixed to differentiate regions of poor, close and good quality fits as illustrated in Fig. 3.6. In total, 93 models are determined as good fits, 95 as close fits and 59 are categorized as poor fits. It can also be noted that most close fits present boundary excess lumps located around frequencies of 0.15 and 0.4, which indicates that these models are not capturing dynamics occurring at periods of 13 and 5 hours respectively. Both these dynamics might be caused by occupant usage of appliances generating heat inputs not covered in the measurements, e.g., kitchen appliances. A 13-hour period for instance typically represents daily unoccupied residential kitchen-usage schedules with morning to late evening activities, i.e., 7:00-20:00, while a 5 hour period corresponds better to an occupied daily Dutch kitchen-usage schedule, where dinner is prepared rather early, i.e., 7:00-12:00-17:00.



(a) Forward model selection *fms1*



(b) Forward model selection *fms2*

Figure 3.7: Flow diagram of forward model selection paths. The figure presents flows and nodes to illustrate the selection path of models, with their width being proportional to the number of final models using this selection path. A final selected model $T_iT_eT_h$ for instance, could be obtained by a selection path either following $T_i > T_iT_h > T_iT_eT_h$ or $T_i > T_iT_e > T_iT_eT_h$. The width of the flows entering the designated $T_iT_eT_h$ node is therefore proportional to the number of final models considering the $T_iT_eT_h$ model in their forward selection path. Nodes may possess fewer number of flows exiting it (right-hand side) than entering it (left-hand side). Taking the same example of the $T_iT_eT_h$ node, this means the model $T_iT_eT_h$ was considered final for a number of cases proportional to the missing existing flow of the node. As the selection procedure moves forward (from left to right), fewer flows are represented as more and more models are identified as final in earlier stages, consequently making the presented selection path flows non-conservative over the flow diagram.

In fact, findings from related work leveraging symbolic regression knowledge discovery on a similar data set, reported under Annex A, revealed the preferred use of gas-meter measurements over heat input signal for the heat dynamics identification of 24 of these buildings [144]. Indeed, Dutch homes typically employ gas to supply both space heating and kitchen appliance needs. This underlines the impact and importance of these appliances on building heat inputs.

The forward selection paths adopted in both model selection phases are illustrated in Figure 3.7.

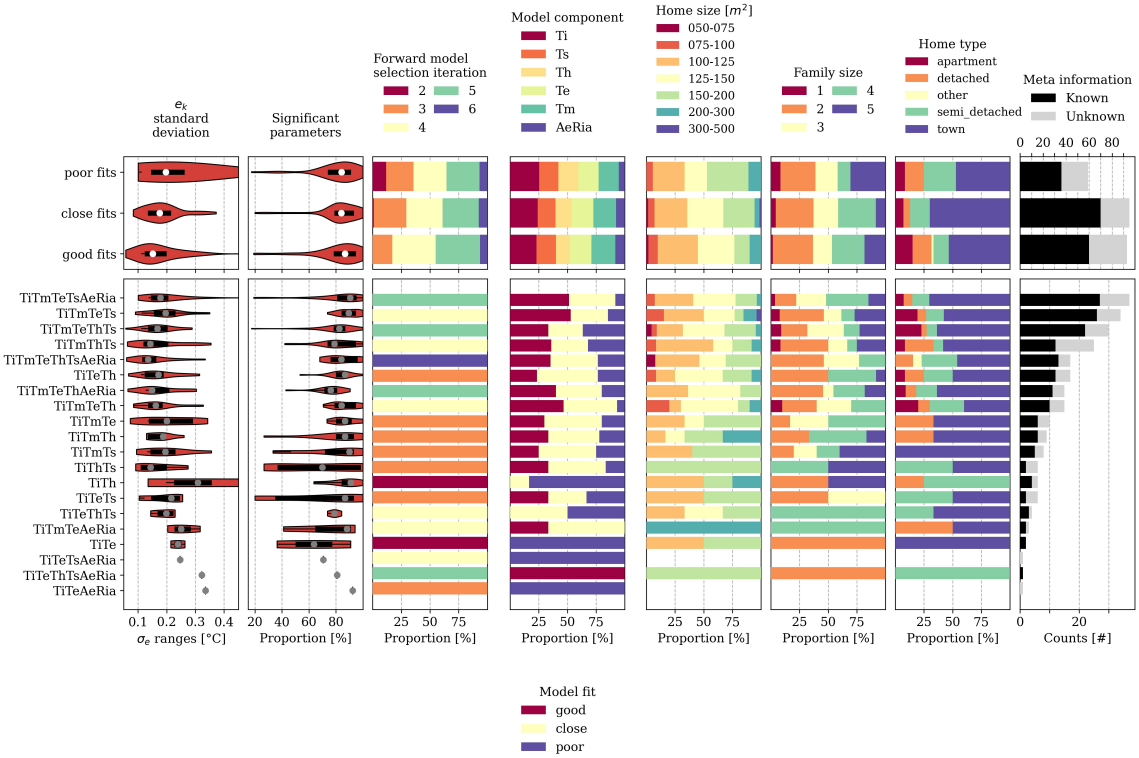
The forward model selection *fms1* displays an overall even distribution of models amongst the 2nd iteration phase, with a slight preference for T_iT_h models, while in the 3rd iteration phase, $T_iT_mT_s$ comes out as the most likely choice. This seems to indicate favored additional

degrees of freedom around the inside temperature in the forward selection procedure. It can be noted that only one A_eR_{ia} envelope extension is selected out of the 2nd iteration phase, yet more envelope model extensions are preferred in the later phases of the selection. The 4th iteration distinctly denotes $T_iT_mT_eT_s$ as the most preferable choice, yielding a consequent share of final models. Finally, the 5th and 6th iterations largely compose final selected models with little to be noted from their selection paths.

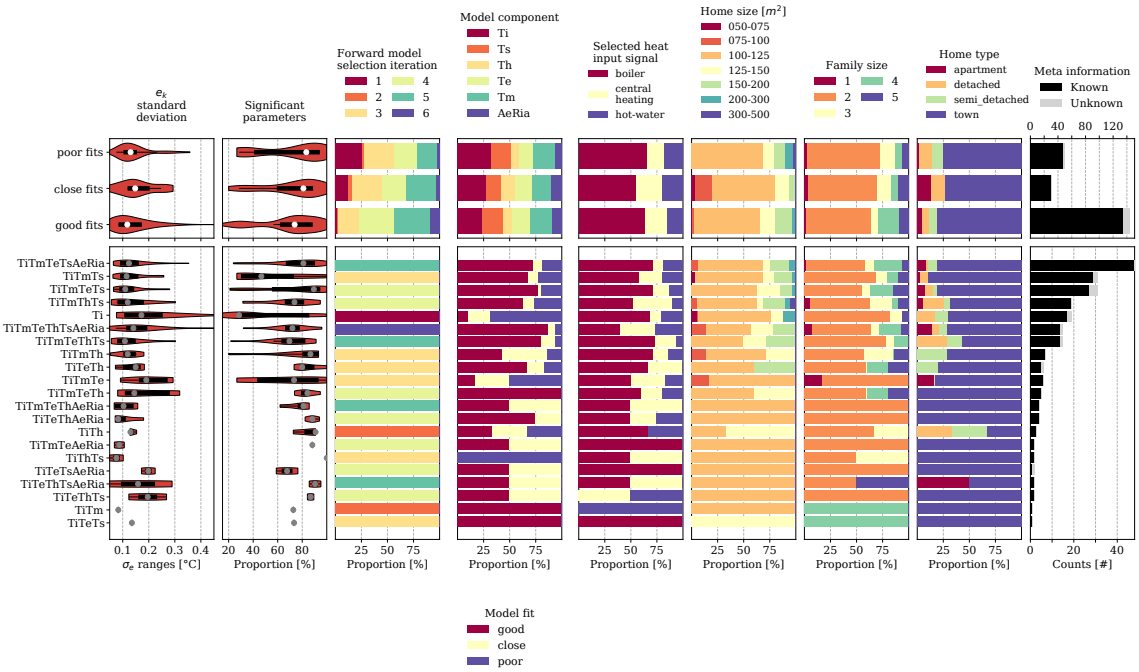
On the other hand, the forward model selection *fms2* portrays a close to even distribution of models amongst the 2nd iteration phase, with an under-representation of T_iT_e models, while in the 3rd iteration phase, $T_iT_mT_s$ comes out as the clear most likely choice, similarly to *fms1*. The 4th iteration distinctly denotes $T_iT_mT_eT_s$ as the most preferable choice, yielding a consequent share of final models. It can be noted that A_eR_{ia} envelope extensions only appear from the 4th iteration phase on, while the 5th iteration portraits $T_iT_mT_eT_sA_eR_{ia}$ as the preferred (final) model. The forward selection path logically terminates at iteration 6.

Identified models are presented in Figure 3.8 along with their residual standard deviation, parameter significance, proportions of fit quality as well as available building meta-information, i.e. home size, home type and family size, which will be discussed in the following sub-section. Residuals' standard deviation values are displayed from violin plots in panel one, while panel two depicts in a similar manner the distribution of parameter significance proportions per group, i.e., the number of estimated parameters possessing a significant p-value over the total number of estimated parameters per model. Panel three shows the iteration phase in which the forward model selection procedure stopped. These are logically homogeneous for any given individual model (bottom panel). Panel four presents either the proportion of model components across the groups (top) or model fit quality (bottom), as these are insights specific to their opposing panel. Panels five to seven offer the proportions of available meta-data information as self-reported by occupants, namely home size, family size, and home type. Lastly, panel eight displays the number of members per group while highlighting the portion of meta-information self-reported, or not, within them.

fms1 - First, the standard deviations of obtained model residuals serve to illustrate the amplitude of forecasting errors produced. These range between values of 0.05 and 2°C for the most extreme cases, and possess central tendencies grouped between values of 0.1 and 0.2°C. The grouping of residual standard deviations per model fit quality (top Figure 3.8) clearly demonstrates a decrease in residual amplitude as the fit increases in quality, although a larger tail persists amongst good model fits, compared to close ones. It should be noted, however, that residual amplitudes are an indication of the degree of noise variations present in the measurements, and are independent from residual covariance, i.e., model fit quality. A residential home with 4 occupants might result in a larger amount of noise, i.e., higher standard deviations, while possessing a thermal model accurately capturing all of its dynamics, with residuals demonstrating white noise properties, i.e., nCPBES close to zero. The aforementioned observation consequently comes rather as a fortuity than the



(a) *fms1*



(b) *fms2*

Figure 3.8: Identified models grouped per fit quality (top) and RC model (bottom) with their respective (left to right) residual standard deviation distribution, parameter significance, iteration phase, modeling components (top), or model fit quality (bottom), selected heat input signal (*fms2* only), meta-data proportions, and group size.

result of a correlation between these two features. Secondly, parameter significance reveals whether estimated model parameters exhibit substantially likely probabilities and support the evaluation of model robustness. The comparison of parameter significance grouped by model fit quality shows that good fits present slightly larger central tendencies of significant parameter proportions compared to poor and close fits, although all three groups present similar distribution tails reaching down to 20%. Thirdly, the forward model selection iteration phase proportions show no 1st order model selected as best fit in either of the categories, and no 2nd order models are either present in close and good model fits. This finding confirms our initial assumptions that first- and second-order models would not be sufficient to capture thermal dynamics of buildings. The iteration proportions clearly display larger iteration phases becoming more important from poor to good model fits, suggesting that to obtain a good fit it is likely the model will be more complex. Lastly, it is found that the four most considered models are $T_iT_mT_eT_sA_eR_{ia}$, $T_iT_mT_eT_s$, $T_iT_mT_eT_hT_s$ and $T_iT_mT_hT_s$, which all employ the sensor, T_s, and medium, T_m, model components. The two largest of them notably possess the greatest proportion of good fits, reaching 50%, while being sensibly similar models with common components T_i, T_m, T_e and T_s.

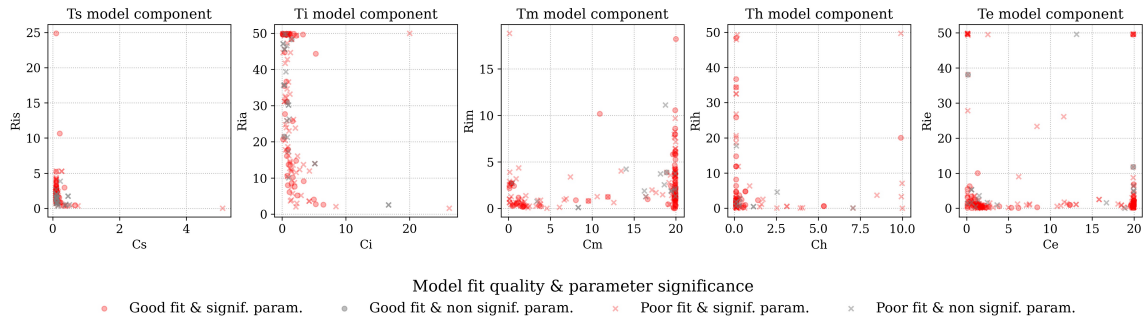
fms2 - The forward model selection variant displays a much larger share of good model fits than *fms1*, with 64% of good fits (144 counts), 14% of close (31 counts) and 22% of poor quality model fits (50 counts). This significant increase in good model fit quality can be attributed to either the different and more recent data set employed in *fms2* or to the additional degree of freedom given to the forward selection procedure *fms2* in selecting a more likely heat input signal. The aggregated central heating and hot water demand heat input signal (boiler set-point) shows to be the dominant choice favored in the forward selection process, with 142 models over 46 selecting central heating and 37 hot water. This finding confirms the assumption that hot-water demands link to a significant heat input contribution to residential buildings from either kitchen, laundry, or bathroom appliances.

3.4.2 Parameters and building characteristics

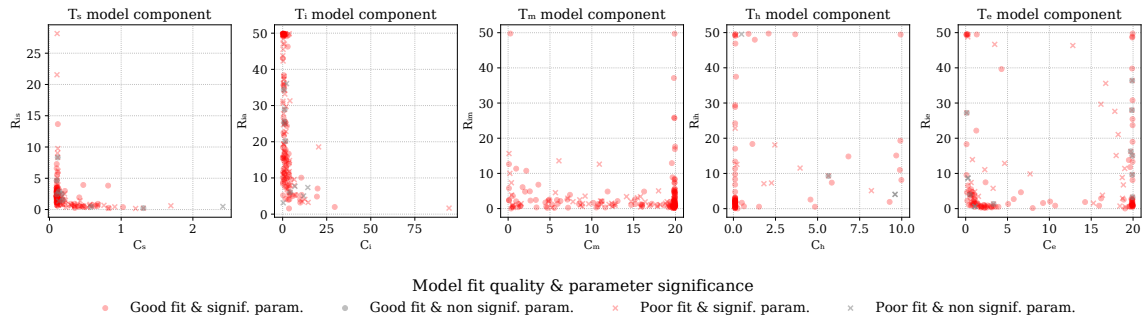
Identified parameters and models coupled with available building characteristics constitute a valuable examination analysis that has the potential to unveil meta-data links to exposed building thermal dynamics. The building meta-data distributions of Figure 3.8 reveal that poor to good model fits possess increasing proportions of smaller home sizes along with larger shares of family sizes of 2. This supports the intuitive thought that smaller, thus simpler, residential homes are more likely to result in good model fit.

Estimated thermal capacities, C, and resistance, R, parameters are presented per model component under Figure 3.9 with highlighted model quality fit and parameter significance to allow visual appreciation of their coupled distributions.

fms1 - Results display thermal capacities of sensor and heater components as relatively



(a) *fms1*



(b) *fms2*

Figure 3.9: Scatter plots of RC parameter estimated per model component, with model fit quality and parameter significance. Good from poor model fits are differentiated, where close fits are here grouped together with poor fits, while the scatter point crossover between two parameters will be considered significant only if both represented parameters are so.

aggregated around 0 kWh/K, while medium and envelope capacities are relatively split between their lower upper bound values, 0 and 20 kWh/K respectively. On the other hand, resistance estimates are mostly comprised between values of 0 to 5 K/kW at the exception of R_{ia} which spreads quite evenly from 0 to 50 K/kW. The scatters show no apparent correlation between estimated parameter values and their relationship to fit quality or significance.

fms2 - The scatters show no apparent correlation between estimated parameter values and their relationship to fit quality or significance, yet components display varying distribution trends. Component capacities, notably the *sensor* component, typically exhibit concentration around values of 0 kWh/K, while, on the contrary, *medium* and *envelope* capacities both spread till upper bound values of 20 kWh/K. Resistance distributions are most dispersed for the *inside* component while the *sensor* one, again, presents skewed characteristics towards values of 0 K/kW.

3.4.3 Building envelope performance

Identified building heat dynamics can be leveraged to calculate building envelope insulation properties such as the Heat Transfer Coefficient (HTC). HTC is defined in ISO 13789:2017 [145] as the heat flow rate from the internal air mass to the surrounding external environment divided by the indoor-outdoor air temperature difference [146]. Its estimation is obtained from the sum of heat flow rates due to ventilation UA_V, and transmission UA_T, which includes plane building envelope elements as well as thermal bridges. Linking these elements to the identified thermal resistances R of the model, the HTC can be derived from [128]

$$HTC = \overbrace{\frac{1}{\frac{1}{UA_{ie}} + \frac{1}{UA_{ea}}}}^{UA_T} + \overbrace{UA_{ia}}^{UA_V} \qquad (\mathrm{WK}^{-1}), \tag{3.17}$$

$$HTC = \underbrace{\frac{1}{R_{ie} + R_{ea}}}_{\forall T_e \in \mathbb{M}} + \underbrace{\frac{1}{R_{ia}}}_{\forall T_e R_{ia} \in \mathbb{M} \text{ or } T_e \notin \mathbb{M}} \qquad (\mathrm{WK}^{-1}), \tag{3.18}$$

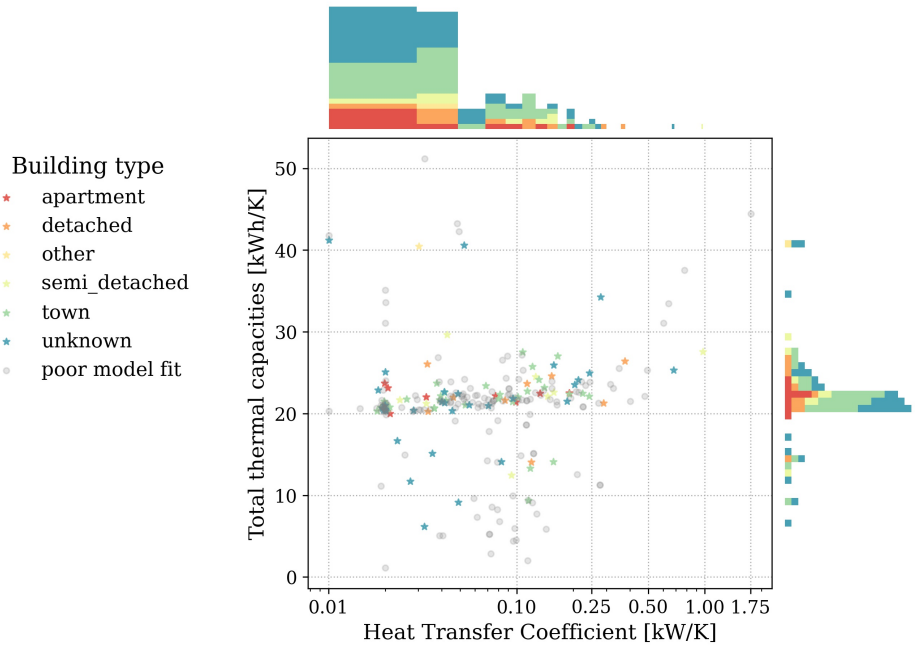
where $\mathbb{M}$ stands for the model component ensemble of the final selected model. HTC can be expressed in absolute units, i.e., W/K as defined above, or in useful floor area relative units, i.e., W/(K·m^2) as recalled in ISO 52016:2017 [147] with

$$HTC_{norm} = \frac{HTC}{A_{use}} \ (\mathrm{WK}^{-1}\mathrm{m}^{-2}), \tag{3.19}$$

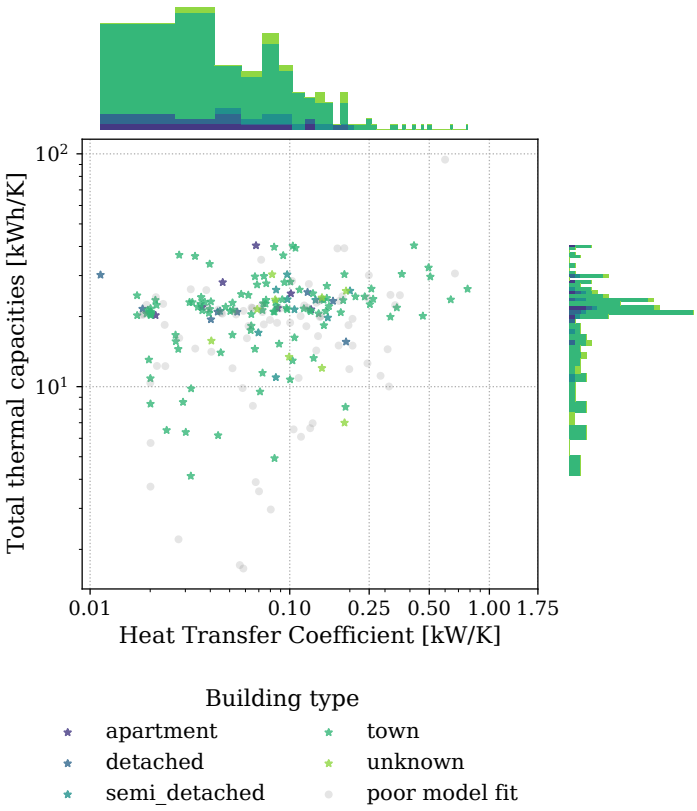
where HTC_{norm} refers to the area normalized HTC and A_{use} to the total useful floor area of the considered building. The latter naturally being better suited to benchmark insulation performances thanks to building floor surface independence.

With useful floor area unavailable across this study's building stock however, we proceed to identify absolute building HTCs and total thermal capacities jointly. This allows a relatively fair comparison of building thermal properties together. Indeed, investigating these parameters by pair presents the advantage of providing a complete overview of a building stocks' thermal properties. The inclusion of total thermal masses in this frame puts the absolute HTC into perspective while providing an efficient way to identify groups of similar buildings, as well as singling-out poorly insulated home envelopes for instance. Figure 3.10 presents these identified characteristics along with highlighted building type and poor quality fits.

fms1 - results show a strong concentration of total thermal capacities between values of 20 and 24 kWh/K with few values reaching above 40 kWh/K and below 10 kWh/K. HTCs present a strong positively skewed distribution, mostly concentrated between 0.02 to 0.05 kW/K, with a smaller peak centered around 0.1 kW/K. The scatter reveals a main cluster of points originating from the strong concentrations of both thermal parameters around their central distributions. Building types do not appear correlated to the presented thermal



(a) *fms1*



(b) *fms2*

Figure 3.10: Scatter plot of building envelope Heat Transfer Coefficient (HTC) versus sum of total heat capacities, with highlighted building types. Poor model fits here englobe both close and poor fits.

properties. A larger share of isolated dwellings (*town* building type) are present at the center of the cluster, yet their vast larger proportion within the data set biases this observation. Identified good model fits presenting HTC values above 0.20 kW/K can thus be considered of poor thermal insulation. This constitutes a non-intrusive, scalable and quite simple building envelope characterization which can support city-scale building stock analytics. Retrofit potentials could consequently be evaluated from the granted insights and provide users with impactful energy saving opportunities from recommended insulation upgrades.

fms2 - results display a strong concentration of total thermal capacities between values of 20 and 30 kWh/K, with few values dropping below 10 kWh/K and none reaching above 40 kWh/K. HTCs present a strong, positively skewed distribution, mostly concentrated between 0.02 and 0.25 kW/K. The scatter reveals a central cluster, although building types do not appear correlated to the presented thermal properties.

3.5 Applications and Outlooks

The development of methods that focus on the scalability of analysis across the building stock unlocks the potential of several important applications. The results of this method illustrate its effectiveness on a set of buildings that could be replicated in other contexts. This section outlines a review of those results related to several applications.

3.5.1 Building performance benchmarking

Benchmarking the building stock enables the owners and occupants to understand how their building compares to its *peers*. Defining just what a peer is for a certain building is a challenge in itself. A certain amount of metadata, or characteristic attributes are necessary to undertake fair benchmarks between buildings. Notable building geometry-related characteristics proportionally influencing heating and cooling demands encompass area-to-volume ratios such as shape factor, i.e., an envelope surface to heated volume [148] or surface [149] ratio. Yet the collection of these information at scale is tedious [150]. The automated creation of dynamic models could enhance this process by enriching existing metadata with identified building thermal properties [151], ranging from model structures to parameter estimates which could be employed within the benchmarking process.

For instance, building performances are typically evaluated from key performance indicators (KPI) such as CO_2 emissions reductions, energy costs savings, energy balance, thermal/light comfort, system efficiency or peak demand reduction [152]. A common approach employed to compare the energy performance of a heterogeneous building set consists in area-normalizing their respective energy consumption [153]. However, floor surface does not provide a complete characterization of a building's thermal mass. Knowledge of buildings thermal capacities can provide a much richer description of their heating and cooling inertiae for fairer thermal load comparisons. Building thermal capacities C can either encapsulate their internal environment

with units in J/K, or air and furniture areal capacity considering units of J/(K·m^2) [147]. The latter, similarly to HTC_{norm}, includes useful floor area information and can serve in benchmarks for the consideration of not only building thermo-physical properties, but also geometry. When undertaking such energy performance benchmarks, heating and cooling demands Φ_d could consequently be normalized as follows,

$$\Phi_{d,norm} = \frac{\Phi_d}{\sum_j C_j} \tag{3.20}$$

where $\Phi_{d,norm}$ here either stands for the internal environment capacity normalized heat load of a building, with C in J/K and $\Phi_{d,norm}$ in K/s, or the air and furniture areal capacity normalized heat load, with C in J/(K·m^2) and $\Phi_{d,norm}$ in Km^2/s. The subscript j refers to the components of the fitted grey-box model.

This scaling approach makes it possible to compare design assumption, i.e., useful floor area, with field performance, i.e., heating/cooling demands, at scale. Indeed, through the inclusion of building geometry information in the scaling process of identified thermal characteristics, it becomes possible to compare actual building thermo-physical properties to design parameters from technical standards on a building stock level.

3.5.2 Building stock scenario modeling

One of the key benefits of white box modeling is the ability to test possible future scenarios of performance enhancement [154]. Undertaking this effort on a large building stock is a significant challenge [155]. The automated creation of physics-informed models enables this in a scalable and effective way, allowing renovation influences or policy decisions impact assessments up to a city or district-scale. This type of effort has been explored on non-residential buildings in the direction of inference of higher granularity data from annual and district-level public data [156]. Quite concretely, one could imagine varying identified HTC distributions within the building stock to predict insulation renovations' impact on thermal loads.

3.5.3 Decomposition of energy meter into end-use loads at a city-level

The integration of more dynamic influences on energy grids from renewable sources such as wind and solar have a significant impact on their operation. The ability to characterize, at scale, energy-consuming dynamics of large numbers of buildings can improve optimal grid operations. Policies aiming at enhancing grid stability using technologies like storage unavoidably profit from more accurate energy demand characterizations [157]. Additionally, the decomposition of energy-meter information into load-influencing factors sets the foundations for factor-dependent energy predictions, which support scenario-specific predictions at city level. For example, with identified climate's influence on overall building stock load, one could predict how a city's energy demand might change under varying weather scenarios, or

account for per-factor uncertainties in forecasted values for a resilient and optimal operation of the energy system. Further innovations in this direction can support efforts from the literature focusing on model development using large, open data sets from energy disclosure programs [158], [159] or from geospatial sources [160] for instance.

3.5.4 Demand side management

Identified thermal dynamics may also be leveraged to evaluate the energy flexibility potential of buildings. Indeed, the derivation of a system's time constants, which characterize the dynamic response of the considered system, can be determined based on estimated parameters [136]. These time constants can notably provide information about the building's reaction to affecting variable changes, namely weather conditions and heat inputs. The work of [151] exemplifies this process on a data set of 39 Danish residential buildings. From identified thermal dynamics it formally links estimated time constants to each building's energy flexibility potential. Proposing scalable methods to evaluate the dynamic response of the building stock is a crucial step that our work opens the door to for developing more effective demand-side management strategies.

3.5.5 Data augmentation

Additionally, calibrated RC models may be exploited to generate large, augmented data sets that could serve multiple end-use applications, e.g., supporting surrogate black-box model training either for predictive- or reinforcement-learning-based applications.

3.6 Summary

This work puts forward an automated stochastic model identification approach for building heat dynamics, suited for scalable deployment. It proposes a forward model selection procedure adapted from [132] and extended with a novel residual auto-correlation indicator, i.e. the normalized Cumulated Periodogram Boundary Excess Sum (nCPBES). This indicator allows automated classification of identified models into groups of fit quality.

Out of the 247 buildings the approach was tested on, two sets of results were gathered and analyzed.

fms1 - 93 model fits were identified as good, 95 were classified as close while 59 were designed as poor. Good model fits presented overall larger shares of model complexities and parameter significance along with smaller reported building surfaces and family sizes compared to poor and close model fits. Estimated RC parameters presented no notable tendencies when compared to model fit qualities or their significance.

fms2 - close to two-thirds (144) of all identified models were evaluated to be of good quality, providing stable, accurate, and valuable thermal descriptions of residential buildings.

We examined the thermal properties of the building stock by visualizing their total thermal capacities and respective HTC. A main cluster of buildings with similar properties is clearly observed suggesting a large share of buildings possessing similar thermal characteristics.

Finally, we open source the results of *fms2* as an open data set entitled `Grey-brick buildings` [123] and discussed how the proposed approach and open set is valuable to the building sector. In particular, how automation and scalable solutions for building stock model identification can support in an unprecedented way applications such as building performance benchmarks, city-scale scenario modeling, energy disaggregation to building end-loads, large-scale demand-side management, and data augmentation.

Chapter 4

Hierarchical building load forecasting

Chapter overview

- Extension of uni-dimensional hierarchical structures to multi-dimensional ones
- Developing a machine learning method for forecasting coherent hierarchies
- Performance examination of tailored machine learning architectures
- Application to building electrical load forecasting
- Case studies: 2NECO and BDG2
- GitHub repositories: /hierarchicallearning and /structuralhierarchicallearning

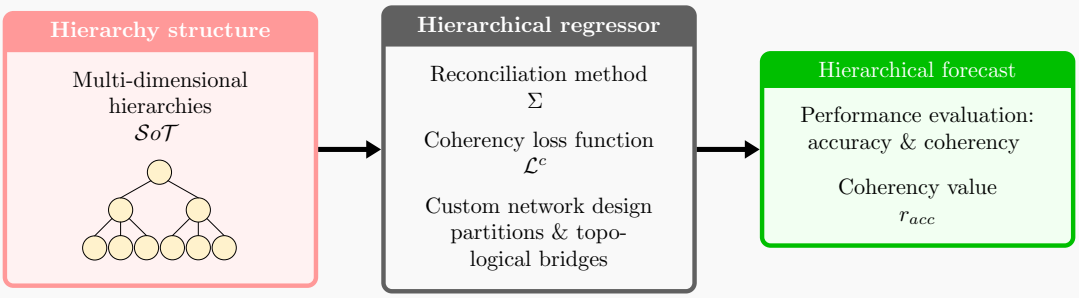


Figure 4.1: Overview of the hierarchical learning method

The content of this chapter builds upon materials published in [161], [162].

"The whole is greater than the sum of its parts."
Aristotle

4.1 Preface

A better anticipation of the future supports better decision-making. This is true across all sectors. Yet, more accurate forecasts alone often do not suffice. When dealing with different abstraction levels across a system or organization, it is commonly more important to obtain coherent predictions across all considered layers and horizons, not to result in unaligned decisions or possibly even conflicting ones [163]. This obstacle arises in multiple

domains, including tourism [164], [165], retail [166], stock management [167] and smart grid management [168], which showcases this matter quite adequately.

Traditionally, smart grid operators focused on forecasting the system's total demand. However, with the increasing adoption of smart meters at grid edges and substations, the focus is shifting. Grid management now benefits from high-frequency measurements available at multiple levels of aggregation allowing accurate forecast estimations across both spatial and temporal scales, i.e., from sub-meters to regional-level, with per seconds to monthly aggregated information [168]–[170]. Yet, the pluralities and independence of models and their consequent forecasts inevitably produce inconsistencies across aggregation levels, i.e., lower-level predictions might not sum up to higher-level ones and vice-versa [171]. The consequent challenge decision-makers are now faced with is to obtain coherent predictions across the different horizons and scales of the system. Hierarchical structures (or trees) are said to be coherent when their values at the disaggregate and aggregate scales are equal when brought to the same level [164]. Should forecasts not be coherent, decision-making units would be planning using diverging views of the future. Optimal decision-making consequently requires forecasts to be coherent across all considered dimensional hierarchies.

4.1.1 State-of-the-art: hierarchical forecasting

Enforcing coherency in hierarchical structures is a concept that dates back to 1942 [172] and was first defined in 1988 as *reconciliation* [173]. It leverages linear balancing equations from covariance compositions inherent to hierarchical structures to optimally re-adjust coherency mismatches. Hyndman et al. [174] later reformulated the approach with a unifying statistical method, independent of prediction models, along with notations more appropriate to hierarchical forecasting.

Hierarchical forecasting can thus be defined as the process in which coherent predictions need to be made within a fixed hierarchical structure. Commonly, forecasts are first estimated separately considering each series of the hierarchy in a disjointed manner. These forecasts are designated as independent *base* forecasts [165]. Generating base forecasts for each series implies that specialized models can be developed for each part of the hierarchy, incorporating node-specific available information [164]. Base forecasts are then linearly combined (reconciled) leveraging available information across the hierarchy to ensure coherency; a process employed by all hierarchical forecasting approaches as of to date [163], [165], [167], [168], [173]–[183].

Reconciliation approaches

Predominant reconciliation techniques comprise traditional bottom-up and top-down approaches, trace minimization, optimal combinations, and recently developed machine-learning methods.

Bottom-up hierarchical forecasting consists in generating base forecasts at the very bottom level of the hierarchy and enforce coherency through their direct aggregation across the tree [184]. The greatest advantage of this approach is that it can draw information from the most disaggregated levels of the tree, consequently avoiding any information loss from aggregation [165]. However, series located at tree leaves tend to possess low signal-to-noise ratios making them more difficult to predict. This is particularly true when dealing with smart-meter electrical demands which are notoriously volatile. Consumption peaks are indeed driven by often highly stochastic occupant behaviors that are close to intractable, consequently making bottom-up aggregation unlikely to provide accurate forecasts across the upper levels of the tree [174].

Top-down hierarchical forecasting on the other hand only generates forecasts for the top level of the hierarchy (tree-root) and proceeds to disaggregate and distribute it down the hierarchy from either historical [185] or forecasted [165] proportions of the data. The approach commonly favors higher aggregation levels of the tree with more accurate predictions and is notably valuable for low-count data. However, aggregation is not without a large loss of information as temporal dynamics and other individual series characteristics cannot be exploited [165]. Additionally, as the success of this approach depends solely on one top-level model, it possesses a higher degree of risk from model misspecifications or inaccuracies [186]. Given both bottom-up and top-down approaches inadequate to profit from the richness of information across a given hierarchy, optimal combination techniques emerged. Linearly reconciling base forecasts towards coherency, these approaches allowed interactions between different levels of the hierarchy, leveraging in particular correlations and covariances present in such structures [174].

However, estimating the covariance structure of a hierarchy from base forecasts is challenging. Indeed, Wickramasuriya et al. [181] declared that the covariance matrix of the coherency errors is "impossible to estimate in practice due to identifiability conditions" such that even with high-frequency data available, assumptions on its form must be made [187]. The ordinary least-square (OLS) estimator was particularly developed by Hyndman et al. [174] and Athanasopoulos et al. [165] to avoid this problem. Their approach demonstrated improved results compared to other commonly adopted techniques. A weighted least squares (WLS) approach, considering variances from the variance-covariance matrix diagonal but ignoring the off-diagonal covariance elements, was put forward by Hyndman et al. [188]. Wickramasuriya et al. [181] later provided the theoretical justification for estimating variances from base forecast error variances. They proposed a generalized least-squares (GLS) estimator and found the incorporation of correlation information into the reconciliation process to benefit forecasting accuracy, with resulting reconciled forecasts guaranteed to be, in mean or in sample, at least, as good as their base forecasts, given a particular covariance structure.

Finally, in recent years, machine learning approaches have made their way into hierarchical forecasting. Relying on powerful statistical regressors and the availability of larger and richer

data sets, machine learning emerges as an appealing and suitable tool for estimating the persistently challenging covariance matrix. Spiliotis et al. [167] put forward such an approach employing a bottom-up method to reconcile predictions from Random Forest and XGBoost regressors. Taking as input the base forecasts of all the series of the hierarchy, the reconciled tree is then obtained from bottom-up aggregation. It allows non-linear combinations of the base forecasts, extending conventional linear approaches thanks to its machine-learning nature. Sagheer et al. [189] proceeded to obtain coherent hierarchies from deep long-short term memory (DLSTM) recurrent neural networks by applying transfer learning across their hierarchies in a bottom-up fashion. They evaluated their approach on national-scale Brazilian electrical power production as well as Australian domestic tourism data. In another work, Mancuso et al. [190] proposed a method to unify the two prevailing processes that are forecasting and reconciliation. By including hierarchical information in the forecasting process through a customized loss function, they allow the network to train towards reconciled forecasts using a top-down disaggregation process.

None of these approaches, however, include the general formulation of hierarchical forecasting within their learning framework. This limits their reconciliation approaches to encompass solely traditional approaches, i.e., bottom-up or top-down, which, as has been mentioned, only exploit a fraction of the available information of hierarchical structures.

Dimensional considerations

While numerous works have first approached the reconciliation of hierarchical structures from a spatial (cross-sectional) dimensional frame perspective [165], [167], [168], [173]–[175], [181]–[183], temporal hierarchies have also been the center of recent attention within the field [163], [176]–[180].

Athanasopoulos et al. [176] first introduced the notion of temporal hierarchies with forecasting reconciliation performed in the temporal dimension. Quite similarly to spatial reconciliation, base forecasts are independently produced across a defined set of temporal aggregation levels, e.g., weekly, daily, quarter-daily, hourly to per-minute or seconds granularities. This allows models to capture temporal-specific characteristics of the times series across the hierarchical-structure, e.g., trends or seasonality possessing particular time-frames. Base temporal-forecasts are then reconciled across all forecasting horizons and temporal tree-structure, allowing aligned decisions across multiple planning horizons [164]. Nystrup et al. [163] notably proposed temporal estimators accounting for autocorrelation structures to reconcile electric grid load forecasts. It was found that auto- and cross-covariances significantly improved forecast accuracy uniformly across all temporal aggregation levels.

It thus becomes clear that both spatial and temporal hierarchical forecasts produce substantial empirical accuracy improvements. By dealing with parameter estimation errors and model misspecifications, forecast combinations have demonstrated significant error variance reduction across numerous works [186], [191], [192]. Exploiting both available hierarchical di-

mensions to further improve prediction accuracies consequently emerges as not only appealing but quite evident. Kourentzes and Athanasopoulos [164] notably advanced a framework to produce spatial- and temporal-coherent forecasts (designated as cross-temporal), supporting all hierarchical levels with short- to long-term forecasts. Their work demonstrated empirical evidence that leveraging both dimensions in reconciliation offered improved accuracies compared to uni-dimensional reconciliation, i.e., spatial or temporal. A finding certainly due to the complete information exposure the approach provides. Spiliotis et al. [171] later proposed a cross-aggregation process to iteratively generate coherency across spatial hierarchies from multiple temporal aggregations applied to electricity consumption forecasting. Punia et al. [193] introduced a similar framework leveraging deep learning algorithms applied to supply chain base forecasts. Their approach, however, produced coherency solely from bottom-up approaches.

While the advantage of multi-dimensional hierarchical forecast has become evident, there exists, as of today, no generic formulation of these approaches. Indeed, while Spiliotis et al. [171] stated that it is possible, in principle, to design a summing matrix S that accounts for both considered dimensions of reconciliation, a theoretical formulation of S and its subsequent reconciliation approaches was not put forward. Indeed, the design of a reconciliation estimator that fully captures scaling issues and cross-sectional interdependencies is not straightforward. Yet, this deprives multi-dimensional reconciliations of exploiting custom dimensional considerations. The principal counterargument to undertaking such formulations is grounded on the fact that multi-dimensional hierarchies generate increasingly large tree structures that could soon become intractable to estimate. Recently, however, the work of Nystrup et al. [179] proposed a dimensionality reduction technique to counter this problem. Using eigendecomposition when reconciling forecasts, maximum information can be extracted from the error structure using available data. They find that uniformly improved predictions can be obtained across all aggregation levels, with the estimator achieving state-of-the-art accuracy all the while being applicable to hierarchies of all sizes.

4.1.2 Motivation

This comprehensive state-of-the-art overview underlines the following shortcomings;

(i) Base forecasts are typically produced separately, considering each series of the hierarchy in a disjointed manner. While this procedure allows the independence and hierarchically-tailored design of these models, it is inherently deprived from the benefits of data information (learning) transfer across models.

(ii) Machine-learning reconciliation approaches have exhibited clear forecast improvements potential. Yet, developed approaches have, so far, not proceeded to put forward a unified method for machine-learning based hierarchical-forecasting. This limits considered reconciliations approaches to the more information-limited bottom-up and

top-down approaches [167], [193]. Embedding advanced reconciliation techniques, e.g., optimal combinations, in the learning process of machine learning regressors is, as of today, still missing.

(iii) Although advantages of leveraging multi-dimensional hierarchies in forecasting has become evident, a generic formulation of such hierarchical-combinations is still needed. Existing tools have demonstrated effective dimensionality reductions of large hierarchies [179], presenting promising solutions to the problem of dimension intractability.

This study proposes a response to this appeal and puts forward a generic multi-dimensional formulation for hierarchical forecasting with machine-learning. We put together a unified and adaptable forecasting and reconciliation method founded on native multi-task machine-learning regressors while framing multi-dimensional hierarchical-forecasting approaches in a generic way. Contributions of this work can be summarized as five-fold;

1. To best exploit available information embedded within multi-dimensional data, we formulate a generic multi-dimensional extension of conventional hierarchical forecasting methods. In particular, we address the problem of diverging reconciliation considerations in a multi-dimensional setting with uni-dimensional couplings of the covariance estimator. This allows the unification of multi-hierarchical structures under a common frame, fueling both traditional and machine-learning approaches with ever-richer and transferable (learning) information.

2. We develop a unified machine-learning-based hierarchical forecasting approach. This grants (*i*) a unique forecasting model the benefit of a complete information overview across its hierarchy, while (*ii*) including coherency constraints within its learning process as well as (*iii*) being adaptable to either independent or combined forecasting and reconciliation processes. It establishes a unified method generating accurate and coherent forecasts at all levels of the hierarchy thanks to a custom hierarchical loss function leveraging coherency information from established field-taxonomy.

3. In the interest of addressing the dimensional tractability of our approach, we put forward dimensionality reduction prospects and illustrate them both theoretically and in practice with an applied demonstration.

4. Further, we expand the approach by cutting down the complexities of models originating from large hierarchical structures with tailored network designs. These exploit topological hierarchical information from trees targeted to support a resourceful, data-efficient, and information-rich learning process.

5. We formally evaluate the relative performance brought by the addition of the coherency requirement across all examined model designs, thus clearly establishing the realized value of coherent hierarchical learning.

6. Our study considers two substantial smart-meter data sets including an established

open source, i.e., the Building Data Genome project 2 (BDG2) [20]. This allows the grounding of our approach thanks to a first-of-its-kind performance benchmark in the field of electric-meter hierarchical predictions, which we render fully replicable.

The greatest advantage of this approach is granting access to the regressor a complete information overview of the considered (multi-dimensional) hierarchy. This permits both a cross-dimensional, data-rich learning process as well as a hierarchically-informed training for hierarchical forecasting. Additionally, by putting forward tailored, ingenious architectures of neural networks we effectively reduce hierarchical model complexities while serving improved and coherency-aware hierarchical forecasts. The outcome is a unified and coherent forecast across all examined dimensions, granting decision-makers a common view of the future serving aligned and better decisions.

In a smart grid operation setting, for example, by inherently leveraging varying hierarchical time-series characteristics, one could imagine the approach to perform above traditional independent forecasts, all the while providing a cohesive view of the energy network to operators. This actively supports the optimal energy management of energy networks; by providing more reliable and coherent predictions, bidding and scheduling algorithms can enhance their performances, thus reducing associated energy costs as well as carbon emissions and accelerating the energy transition. These desirable qualities can similarly benefit sectors such as retail, stock management, and tourism.

4.2 Hierarchical forecasting

In this section, we present the foundations of hierarchical forecasting as defined by Athanasopoulos et al. [165] and Wickramasuriya et al. [181] and extend them to multi-dimensional frames with a generic formulation. We discuss dimensionality tractability limitations and offer dimensionality reduction considerations to address them.

4.2.1 Hierarchical structures

Let us refer to the simple hierarchy of Fig. 4.2 to demonstrate the methodology. Every element (node) of the hierarchy (tree) can be labeled as y_{kj}, where the subscripts k and j stand for the aggregation-level and node observations respectively. We define k_1 as the most aggregate level of the hierarchy (tree root), i.e., node y_{11}, and k_K as the most disaggregate level (tree leaves), i.e., nodes y_{Kj} where $j \in [1 : m]$ and $K = 3$. In such a setting, two important components must be considered; the number of nodes in the bottom level of the hierarchical tree, which is denoted as m, and the total number of nodes on the tree n. Here $n = 9$ and $m = 6$.

Stacking all tree elements in a n-dimensional vector $\boldsymbol{y} = (y_{11}, y_{21}, y_{22}, y_{31}, y_{32}, y_{33}, y_{34}, y_{35}, y_{36})^T$, and bottom-level observations in an m-dimensional vector $\boldsymbol{b} = (y_{31}, y_{32}, y_{33}, y_{34}, y_{35}, y_{36})^T$,

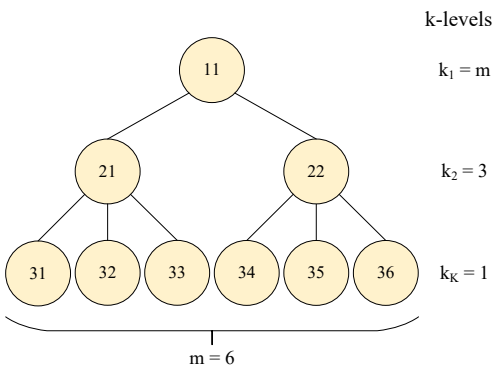


Figure 4.2: A two-level hierarchical tree diagram.

we can write

$$\boldsymbol{y} = S\boldsymbol{b}, \tag{4.1}$$

where S is the summation matrix, here expressed as

$$S = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ & & & I_m & & \end{bmatrix}, \tag{4.2}$$

which is of dimension $n \times m$, and I_m is an identity matrix of size m. S maps the hierarchical structure of the tree, where from the tree leaves $\boldsymbol{b}$ the complete hierarchy $\boldsymbol{y}$ can be reproduced. Notice how S captures the coherency requirements within the hierarchy, integrated here as the linear summations of the bottom-level observations.

Uni-dimensional

Hierarchical structures encompassed within hierarchical forecasting have, as of today, treated either one of the two following dimensional frames, namely, temporal $\mathcal{T}$ or spatial $\mathcal{S}$ (sectional).

We define *spatial* dimensional perspectives as a unique inter-element dimension, which places itself in opposition to the previously-defined *cross-sectional* dimensions [164], [171], [176], [194], which aggregated elements from very different entities together, e.g., stock management, resulting in considerable heterogeneity within "one" (but in fact, multi-) dimension. It is our proposal to re-frame these *cross-sectional* considerations into separate dimensions to allow clear delineations of multi-dimensional frames, as we later detail in Sect. 4.2.3.

Although structures of any shape or form can be designed in both dimensions, it is common for

temporal hierarchies to adopt symmetrical structures, with k-level values being homogeneous across the trees' aggregation levels. Taking the exemplified symmetrical hierarchy of Fig. 4.2, one could consider removing nodes y_{32} and y_{33}; resulting in a hierarchy where $m = 4$, $n = 7$ and node y_{21} being consequently removed as a redundant element of y_{31}. This would result in a non-symmetrical tree which, in the temporal domain, implies non-equally spaced measurement points (or sampling rate) across the considered aggregation level and the ones above it.

Typically, for symmetrical trees, there are $k \in \{k_1, ..., k_K\}$ aggregation levels, where k is a factor of m, with $k_1 = m$, $k_K = 1$, and m/k is the number of observations at aggregation level k. The summation matrix of temporal hierarchies can therefore be expressed as [163]

$$S_{\mathcal{T}} = \begin{bmatrix} I_{m/k_1} \otimes \mathbf{1}_{k_1} \\ \vdots \\ I_{m/k_K} \otimes \mathbf{1}_{k_K} \end{bmatrix}, \tag{4.3}$$

where $\otimes$ is the Kronecker product and $\mathbf{1}_k$ is a k-vector of ones.

To generically define the formulation of the summation matrix of any uni-dimensional hierarchy $\mathcal{H}$, however, one needs to consider the eventuality of non-homogeneous k-level values across aggregation levels as well as uneven tree-depths. To this end, we define

$$s_{ij} = \begin{cases} 1, & \text{if } y_i \text{ is ancestor of } y_{Kj}, \\ 0, & \text{if } y_i \text{ is not ancestor of } y_{Kj}, \end{cases} \tag{4.4}$$

where s_{ij} is a matrix element of the summation matrix $S_{\mathcal{H}}$ given a fixed hierarchical structure $\mathcal{H}$ and y_i here refers to the i-th element of $\boldsymbol{y}$. The subscripts i and j go from 1 to $n - m$ and m respectively. They refer to the considered tree node element i and tree leaf element j. This sets the matrix element of a given node i to either 1 or 0 if it is an ancestor of the leaf element j, or, in other words, whether it is a result of the aggregation of the corresponding tree-leaf element y_{Kj} or not respectively. The summation matrix can then be expressed as

$$S_{\mathcal{H}} = \begin{bmatrix} s_{11} & \dots & s_{1j} & \dots & s_{1m} \\ \vdots & & \vdots & & \vdots \\ s_{i1} & \dots & s_{ij} & \dots & s_{im} \\ \vdots & & \vdots & & \vdots \\ s_{(n-m)1} & \dots & s_{(n-m)j} & \dots & s_{(n-m)m} \\ & & I_m & & \end{bmatrix}. \tag{4.5}$$

This enables the formulation of any hierarchical structure to a summation matrix, e.g., from event-based or equally spaced time-series measurements for temporal hierarchies $\mathcal{T}$, to non-symmetrical or homogeneous aggregation structures for spatial hierarchies $\mathcal{S}$.

Multi-dimensional

Multi-dimensional hierarchies are the product of two uni-dimensional structures and can be obtained from function composition of separate hierarchical structures over another one. Fig. 4.3 illustrates the derivation of a spatio-temporal $\mathcal{ST}$ hierarchy from two disjointed spatial $\mathcal{S}$ and temporal $\mathcal{T}$ structure compositions, i.e., $\mathcal{S}o\mathcal{T}$ and $\mathcal{T}o\mathcal{S}$. The resulting tree structures demonstrate fundamental equivalences, with all tree nodes possessing identical bonds linking one element to the other, and consequently producing a unique hierarchical structure $\mathcal{ST}$.

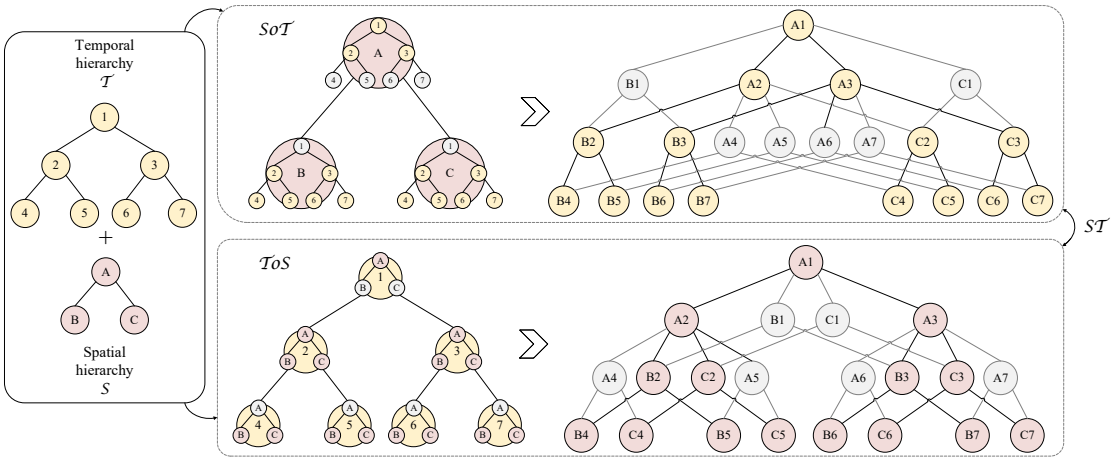


Figure 4.3: Schematic of spatio-temporal $\mathcal{ST}$ hierarchical structure conception from either $\mathcal{S}o\mathcal{T}$ or $\mathcal{T}o\mathcal{S}$ structure composition, both producing an equivalent $\mathcal{ST}$ tree structures. Highlighted nodes (in grey) reveal opportunities for dimensionality reduction by dropping nodes of little dimensional interest, i.e., high temporal granularity in high spatial aggregation levels, and low temporal frequencies in high spatial granularities.

The formulation of the multi-dimensional summation matrix in a generic way, can thus be expressed as a Kronecker product, where

$$S_{\mathcal{ST}} \equiv \begin{cases} S_{\mathcal{S}o\mathcal{T}} = & S_{\mathcal{S}} \otimes S_{\mathcal{T}}, \\ S_{\mathcal{T}o\mathcal{S}} = & S_{\mathcal{T}} \otimes S_{\mathcal{S}}, \end{cases} \tag{4.6}$$

from which the resulting spatio-temporal summation matrix $S_{\mathcal{ST}}$ is of dimension $n_{\mathcal{S}}n_{\mathcal{T}} \times m_{\mathcal{S}}m_{\mathcal{T}}$, which, in the example of Fig. 4.3, yields $3 \cdot 7 \times 2 \cdot 4 = 21 \times 8$. The equivalence of $\mathcal{S}o\mathcal{T}$ and $\mathcal{T}o\mathcal{S}$ is attained via varying orderings of the $n_{\mathcal{S}}n_{\mathcal{T}}$-dimensional vector $\boldsymbol{y}_{\mathcal{ST}}$. These are derived from alternative transpose definitions of the observation matrix $Y_{\mathcal{S}o\mathcal{T}}$ such that

$$Y_{\mathcal{S}o\mathcal{T}} = Y_{\mathcal{T}o\mathcal{S}}^{T} = \begin{bmatrix} y_{11} & \cdots & y_{1n_{\mathcal{T}}} \\ \vdots & \ddots & \vdots \\ y_{n_{\mathcal{S}}1} & \cdots & y_{n_{\mathcal{S}}n_{\mathcal{T}}} \end{bmatrix}, \tag{4.7}$$

where uni-dimensional vectors $\boldsymbol{y}_{\mathcal{S}}$ and $\boldsymbol{y}_{\mathcal{T}}$ are stacked together to form an observation matrix

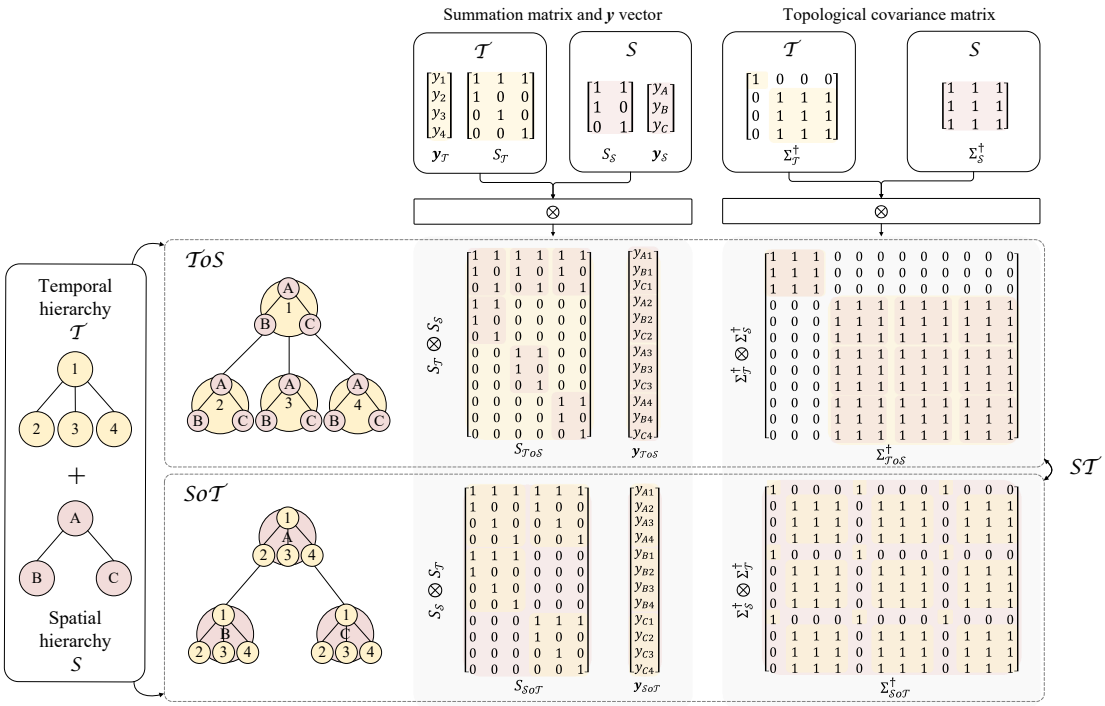


Figure 4.4: Exemplified illustrations of hierarchical derivations of summation matrix, $\boldsymbol{y}$ vector and topological covariance matrix from spatio-temporal $\mathcal{S}o\mathcal{T}$ or $\mathcal{T}o\mathcal{S}$ function composition.

$Y_{\mathcal{S}o\mathcal{T}}$ of dimension $(n_{\mathcal{S}}, n_{\mathcal{T}})$. The $\boldsymbol{y}_{\mathcal{ST}}$ equivalent vectors can then obtained with

$$\boldsymbol{y}_{\mathcal{ST}} \equiv \begin{cases} \boldsymbol{y}_{\mathcal{S}o\mathcal{T}} = & \text{vec}(Y^T_{\mathcal{S}o\mathcal{T}}), \\ \boldsymbol{y}_{\mathcal{T}o\mathcal{S}} = & \text{vec}(Y^T_{\mathcal{T}o\mathcal{S}}). \end{cases} \tag{4.8}$$

In the exemplified structures of Fig. 4.3, we obtain $\boldsymbol{y}_{\mathcal{S}o\mathcal{T}} = (y_{A1}, y_{B1}, y_{C1}, ..., y_{A7}, y_{B7}, y_{C7})^T$ and $\boldsymbol{y}_{\mathcal{T}o\mathcal{S}} = (y_{A1}, ..., y_{A7}, y_{B1}, ..., y_{B7}, y_{C1}, ..., y_{C7})^T$.

With structural combinations of two disjointed dimensional hierarchies producing a unique bi-dimensional structure, it consequently follows that multi-dimensional combinations can be exploited in a similar manner. By chaining function compositions of considered *singular* dimensions over summation matrices and $\boldsymbol{y}$ vectors, any combination of dimensional frames can be considered.

Dimensionality reduction

Multi-dimensional trees, however, introduce a key limitation: the dimensional explosion of hierarchical structures from function composition. With the multiplication of dimensions from summation matrices, what was then considered a tractability shortcoming has now

become an inevitable obstacle needing overcoming.

However, multi-dimensional hierarchies bring with them a consequential consideration: multi-dimensional aggregation levels. Indeed, such trees encompass more than former uni-dimensional high- or low-aggregation levels, they consist of deep structures where multi-dimensional aggregation combinations demand investigation. Spatio-temporal hierarchies, for example, display dissimilar insights from high-temporal-low-spatial aggregation levels, low-temporal-low-spatial or high-temporal-high-spatial ones.

It thus comes to light that, given a defined insight-driven application, subsets of certain multi-dimensional aggregation regions can be of limited use. High-frequency forecasts at very aggregate geographical levels might be of great value to grid operators contemplating frequency control in power systems, but not so much when forecasting tourism flows for instance [164]. Considering the end-goal application of optimal smart-grid control from electric load forecasting of grid edges (smart-building meter), low temporal frequencies and low spatial aggregations would be of little interest. Indeed, frequency control focuses on rather high-frequency samplings at medium-high spatial aggregation levels. However, should the end-goal application be optimal cooperative control of smart-building neighborhoods, then low temporal frequencies and high spatial aggregations would become the dimensional frame of lesser concern. Fig. 4.3 highlights these bi-dimensional nodes over the hierarchical structure conception, i.e., in grey, revealing the potential of dimensionality reduction within multi-dimensional hierarchies.

Therefore, while using spatio-temporal coherent forecasts offer benefits to decision-making, not all outputs from these hierarchies are effectively useful, opening the door to dimensionality reduction.

4.2.2 Reconciliation methods

Traditionally, forecast reconciliation starts by generating an initial forecast of the tree independently for each node, referred to as *base* forecasts $\hat{\boldsymbol{y}}$. This set of hierarchical forecasts is stacked in the same manner as the $\boldsymbol{y}$ vector. Because of the independence of the base forecasts, in most cases, they do not exhibit coherency properties throughout their hierarchical structures. By introducing a matrix

$$G = \left[0_{m \times (n-m)} \,|\, I_m\right], \tag{4.9}$$

of order $m \times n$ that extracts the m bottom-level forecasts, the reconciliation constraint is formulated as

$$\tilde{\boldsymbol{y}} = SG\tilde{\boldsymbol{y}}. \tag{4.10}$$

Reconciliation is necessary when base forecasts $\hat{\boldsymbol{y}}$ do not satisfy this constraint [163]. In such situations, Eq. (4.10) becomes $\tilde{\boldsymbol{y}} = SG\hat{\boldsymbol{y}}$, where G maps the base forecasts into the reconciled tree-leaves and S sums these up to a set of coherent forecasts $\tilde{\boldsymbol{y}}$. SG can thus be thought of as a reconciliation matrix taking the incoherent base forecasts as input and reconciling them to $\tilde{\boldsymbol{y}}$. A major drawback of traditional approaches is that G, as defined in Eq. (4.9), only considers information from a single level.

Optimal reconciliation

To include the exploitation of all aggregation levels in an optimal manner, Hyndman et al. [174] and later, Van Erven and Cugliari [182] and Athanasopoulos et al. [176] formulated the reconciliation problem, as linear regression models. Exploiting either spatial or temporal hierarchical structures, reconciled forecasts are found employing the generalized least-squares estimate:

$$\begin{aligned} \text{minimize} \quad & (\tilde{\boldsymbol{y}} - \hat{\boldsymbol{y}})^T \Sigma^{-1} (\tilde{\boldsymbol{y}} - \hat{\boldsymbol{y}}), \\ \text{subject to} \quad & \tilde{\boldsymbol{y}} = SG\tilde{\boldsymbol{y}}, \end{aligned} \tag{4.11}$$

where $\tilde{\boldsymbol{y}} \in \mathbb{R}^n$ is the decision variable of the optimization problem and $S \in \mathbb{R}^{n \times m}$ and $G \in \mathbb{R}^{m \times n}$ are constant matrices defined by the structure of the hierarchy. The parameter $\Sigma \in \mathbb{R}^{n \times n}$ is the positive definite covariance matrix of the coherency errors $\varepsilon = \tilde{\boldsymbol{y}} - \hat{\boldsymbol{y}}$, which are assumed to be multivariate Gaussian and unbiased, i.e., with zero mean.

If Σ were known, the solution to (4.11) would be given by the generalized least-squares (GLS) estimator

$$\tilde{\boldsymbol{y}} = S\left(S^T \Sigma^{-1} S\right)^{-1} S^T \Sigma^{-1} \hat{\boldsymbol{y}}, \tag{4.12}$$

which has been employed in close to all notable hierarchical forecasting works over the last years [163]–[165], [168], [171], [174], [176], [181]. The precision matrix Σ^{-1} is used to scale discrepancies from the base forecasts, hence, is often referred to as a weight matrix.

The recurrent challenge in estimating Σ^{-1} stems from its dimension $n \times n$ which can potentially become very large.

4.2.3 Multi-dimensional reconciliation

Traditional uni-dimensional estimators can be coupled together topologically to form multi-dimensional ones in a similar manner to the summation matrix, with

$$\Sigma^{\dagger}_{\mathcal{ST}} \equiv \begin{cases} \Sigma^{\dagger}_{\mathcal{S}o\mathcal{T}} = & \Sigma^{\dagger}_{\mathcal{S}} \otimes \Sigma^{\dagger}_{\mathcal{T}}, \\ \Sigma^{\dagger}_{\mathcal{T}o\mathcal{S}} = & \Sigma^{\dagger}_{\mathcal{T}} \otimes \Sigma^{\dagger}_{\mathcal{S}}, \end{cases} \tag{4.13}$$

where $\Sigma^{\dagger}$ refers to the topological covariance matrix of a given covariance matrix Σ. This allows uni-dimensional estimators $\Sigma_{\mathcal{S}}$ and $\Sigma_{\mathcal{T}}$ to incorporate dimension-specific topological

considerations and produce a suitable multi-dimensional estimator $\Sigma_{\mathcal{ST}}$.

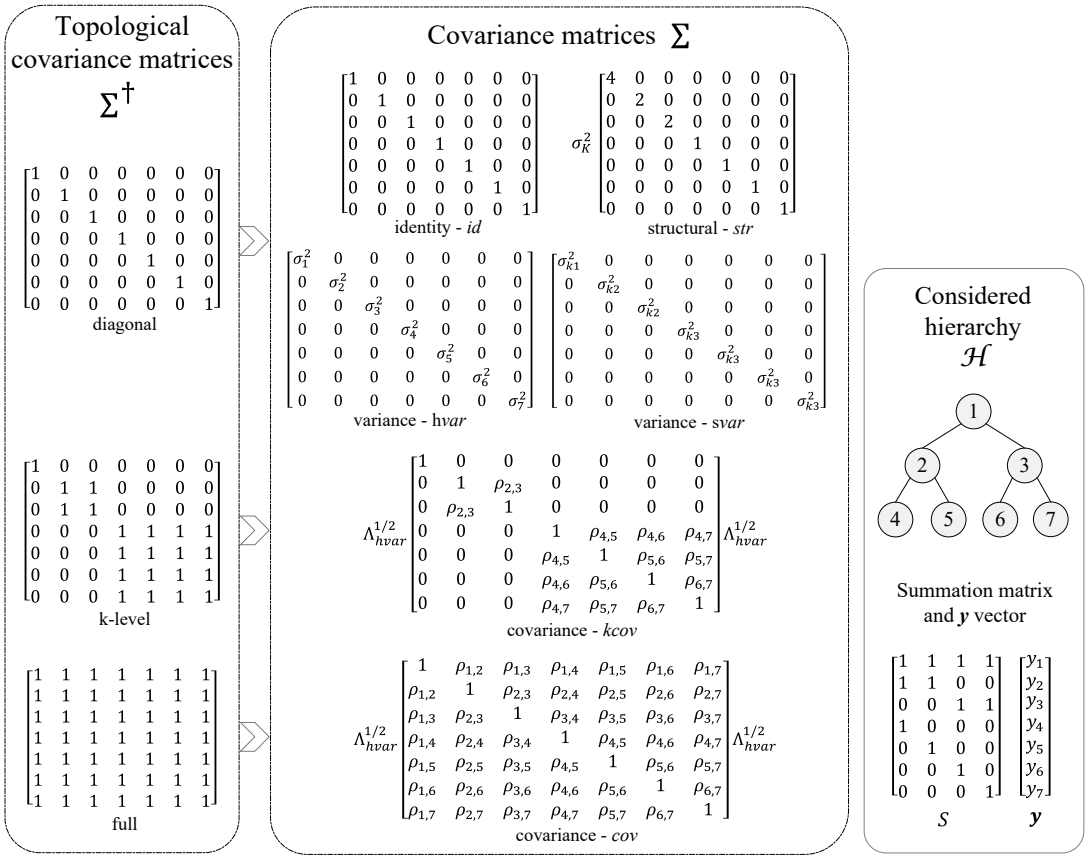


Figure 4.5: Example illustration of the covariance matrices considered in this work along with their associated topological covariance matrices.

The topological covariance matrix is characterized by elements of either 0 or 1 that indicate the mapping form assumption of the considered covariance matrix. Once the topological covariance matrix is identified, we simply populate it with the scaling parameters dictated by the reconciliation approach considered to obtain the covariance matrix. Figure 4.4 exemplifies the identification of multidimensional topological covariance matrices from both $\mathcal{S}o\mathcal{T}$ and $\mathcal{T}o\mathcal{S}$ dimensional-derivations.

To address dimensional considerations in traditional estimators of the covariance matrix applied in reconciliation, we present four state-of-the-art estimators, namely, identity, structural, variance, and covariance scaling with shrinkage, while detailing dimensional deliberations individually.

Identity

A simplifying assumption proposed by Hyndman et al. [174] puts the following identity approximation forward

$$\Sigma_{id} = I_n. \tag{4.14}$$

This simplistic approach has been shown to work well in practice [165] and allows to bypass the estimation of the covariance matrix. It ignores scale differences (captured by the variances) and interrelations (captured by the covariances) information of the observations within the hierarchical structure, which makes it independent of dimensional frame considerations.

Deep neural networks can be expected to build upon such simple relationships and approximate the more complex dependencies of the hierarchy thanks to its automated feature selection, as we later detail in Section 4.3. We refer to this approach with the subscript *id*.

Structural scaling

Structural scaling was proposed by Athanasopoulos et al. [176] as a solution to cases where forecast errors are not available for some aggregation levels. It assumes the variance of each bottom-level base forecast error σ_K^2 is equal and that these are uncorrelated between nodes. Therefore, higher-level error variances are the sum of the error variances of tree leaves series connected to them. By introducing a diagonal matrix Λ_{str} with each element containing the number of forecast errors contributing to that aggregation level, they define

$$\Sigma_{str} = \sigma_K^2 \Lambda_{str}, \tag{4.15}$$

$$\Lambda_{str} = \text{diag}(S\mathbf{1}_m), \tag{4.16}$$

where $\mathbf{1} \in \mathbb{R}^m$ is a column vector. The hierarchy illustrated in Fig. 4.2, for instance, gives $\Lambda_{str} = \text{diag}(6, 3, 3, 1, 1, 1, 1, 1, 1)$. The estimator is independent from the considered forecasting method, since no estimation of the variance of the forecast errors is needed, making it computationally efficient [163].

If considering a temporal dimensional frame, the estimator depends only on the seasonal period m of the tree leaves. While with spatial perspectives, the estimator can suffer from heterogeneity within aggregation levels, e.g., residential and commercial buildings typically have heterogeneous electricity demand patterns and scale. Hence, assuming a common forecasting error variance across all leaves-series is not suitable [164]. We refer to this approach as *str*.

Variance scaling

Another estimator proposed by Athanasopoulos et al. [176], referred to as variance scaling, scales the base forecasts using the variance of the residuals. It includes separate variance

estimates for each aggregation level and assumes either homogeneous or non-homogeneous error variance within, but not across, a level. Given the hierarchy presented in Fig. 4.2, this gives

$$\Sigma_{svar} = \Lambda_{svar} = \text{diag}(\sigma^2_{k_1}, \sigma^2_{k_2}, \sigma^2_{k_2}, \sigma^2_{k_3}, \ldots, \sigma^2_{k_3}), \tag{4.17}$$

$$\Sigma_{hvar} = \Lambda_{hvar} = \text{diag}(\sigma^2_{11}, \sigma^2_{21}, \sigma^2_{22}, \sigma^2_{31}, \ldots, \sigma^2_{37}), \tag{4.18}$$

for homogeneous and heterogeneous variances respectively. By definition, this scaling ignores correlations across and within aggregation levels and can be considered as an alternative weighted least-squares estimator.

Similarly to structural scaling, spatial and temporal dimensional scaling can differ due to their intrinsic heterogeneous and homogeneous error variances respectively; we refer to these estimators as *hvar* and *svar*. It follows, that Σ_{svar} and Σ_{hvar} are appropriate to temporal $\mathcal{T}$ and spatial $\mathcal{S}$ dimensional scalings respectively.

Covariance scaling

To exploit important information about a time series at different frequencies (temporal dimension) or inter-scale differences (spatial dimension), Nystrup et al. [163] argue that potential information in the autocorrelation structure should be included. They consequently proposed a *covariance* scaling for temporal hierarchies estimating the full covariance matrix within each aggregation level, while ignoring correlations between them.

Following these footsteps, we explore both full and k-level, or so-called block, covariance estimates such that, for the hierarchy illustrated in Fig. 4.2, the estimator is either

$$\Sigma_{cov} = \Lambda^{1/2}_{hvar} R \Lambda^{1/2}_{hvar}, \text{ or} \tag{4.19}$$

$$\Sigma_{kcov} = \Lambda^{1/2}_{hvar} R_k \Lambda^{1/2}_{hvar}, \tag{4.20}$$

where R and R_k refer to the full and k-level empirical cross-correlation matrix respectively,

$$R = \begin{bmatrix} 1 & \ldots & \rho_{11,36} \\ \vdots & \ddots & \vdots \\ \rho_{11,36} & \ldots & 1 \end{bmatrix}, \tag{4.21}$$

$$R_k = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & \rho_{21,22} & 0 & \dots & 0 \\ 0 & \rho_{22,21} & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & \rho_{31,36} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \rho_{31,36} & \dots & 1 \end{bmatrix}. \tag{4.22}$$

With increasing difficulties in estimating the full covariance matrix from high-dimensional hierarchies, even with high-frequency data available, special forms are commonly assumed. To alleviate this burden, Ledoit and Wolf [195] proposed a Stein-type shrinkage estimator of the sample covariance matrix. Following these footsteps, Nystrup et al. [163] considered a shrinkage estimator of the cross-correlation rather than the cross-covariance matrix to avoid problems with heteroscedasticity. Their estimator is based on decomposing the cross-covariance matrix into two diagonal (heterogeneous) variance matrices $\Lambda_{hvar}^{1/2}$ and a shrunk cross-correlation matrix R_{srk}.

The estimator is defined as

$$\Sigma_{srk} = \Lambda_{hvar}^{1/2} R_{srk} \Lambda_{hvar}^{1/2}, \tag{4.23}$$

$$R_{srk} = (1-\lambda)R + \lambda I_n, \tag{4.24}$$

where $0 \leq \lambda \leq 1$ is a regularization parameter to control the degree of shrinkage towards the identity matrix.

When $\lambda = 1$, shrinkage scaling is equivalent to scaling by the diagonal variance matrix Λ_{hvar}. When $\lambda = 0$, it is equivalent to scaling by the sample covariance matrix. A closed-form solution for the optimal value of λ was derived by Ledoit and Wolf [195] by minimizing the mean squared error. This shrinkage estimator is ideal for a small number of data points with a large number of parameters. With an assumed constant variance, the optimal shrinkage parameter is expressed by,

$$\lambda = \frac{\Sigma_{i \neq j} \mathrm{Var}(\sigma_{ij})}{\Sigma_{i \neq j} \sigma_{ij}^2}, \tag{4.25}$$

where σ_{ij} is the ijth element of the covariance matrix from the base forecast errors. The variance of the estimated covariance, $\mathrm{Var}(\sigma_{ij})$, is computed as depicted in Appendix A of Schäfer and Strimmer [196].

Therefore, in contrast to the preceding variance and structural scaling estimators, this allows strong interrelations between time series in the hierarchy to be captured, while shrinkage alleviates the complexities of the estimation of Σ_{srk} due to its size.

We refer to the shrunken estimators of Eqs. (4.19) and (4.20) as *cov* and *kcov* respectively.

It should be noted that a variety of other well-performing estimators remain, including, but not limited to, Markov [163] or spectral scaling [179] supported by alternative inverse covariance shrinkage GLASSO method [163]. In the intent of limiting the scope of this work to the evaluation of a novel hierarchical regressor, however, the afore-presented prevailing covariance approximation methods are favored. Figure 4.5 provides a visual illustration of the encompassed techniques along with their associated topological covariance matrices.

4.2.4 Evaluation method

The accuracy evaluation of hierarchical forecasting performances requires the consideration of an important principle that common forecasting methods are exempt from, i.e., the structural scale differences inherent to hierarchical structures. Indeed, by its nature, hierarchical forecasting creates outputs of increasing orders of magnitudes, typically characterized by the aggregation levels of the tree, i.e., k-levels. It consequently becomes crucial to take these hierarchically-impended scale differences into account when undertaking the accuracy performance evaluation of hierarchical forecasts, else these would consistently produce poorer performances for the top levels of the aggregation, where predicted values possess larger magnitudes.

This is commonly done by treating each aggregation level of the tree separately first, then evaluating the *relative* per-level performance of the reconciliation phase over the base forecast, allowing the removal of scale differences between aggregation levels. However, a relative performance evaluation does not allow the comparison of approaches across case studies nor the distinctive performances of forecasting and reconciliation phases, which is why we propose to complement relative performance evaluations with measures based on structurally-scaled errors to provide an evaluation method more suited to the evaluation of hierarchical learning regressors.

Relative measures

The prevailing approach employed to evaluate hierarchical forecasting accuracy consists in scaling the accuracy performance of the reconciliation phase over a reference base forecast. This can be done by exploiting either Relative Mean Squared Error (RelMSE) [164] or Relative Root Mean Squared Error (RRMSE) [163], [176], [178]. Both depict the improvement of a given reconciliation approach compared to base forecast. We favor RelMSE over RRMSE to align with the commonly employed Mean Squared Error (MSE) loss function of machine learning models. The RelMSE can be expressed as

$$\text{RelMSE}_k = \frac{\text{MSE}_k}{\text{MSE}_k^{base}} - 1, \tag{4.26}$$

where the RelMSE_k is computed for each aggregation level k and $k \in \{1, 2, ..., K\}$. A negative entry describes a percentage improvement of the reconciled forecast over the base forecast.

The MSE_k is computed as the average error of all prediction steps of a given aggregation level k from

$$\mathrm{MSE}_{kj} = \frac{1}{h}\sum_{t=1}^{h} e_{kj,t}^2, \tag{4.27}$$

$$\mathrm{MSE}_k = \frac{1}{N_k}\sum_{j=1}^{N_k} \mathrm{MSE}_{kj}, \tag{4.28}$$

where $e_{kj,t} = y_{kj,t} - \hat{y}_{kj,t}$ is the forecast error at a starting reference time $t \in \mathbb{R}^h$ of an node kj with k being the aggregation level of the hierarchy possessing N_k elements and j the node observation. The starting reference time t points to the very first time step considered in the hierarchy and is employed to anchor the nomenclature of temporal as well as spatio-temporal hierarchies, which usually encase time frames of [t, t+m], in similar notations as spatial ones.

Measures based on scaled errors

An alternative way of removing the inherent structural-scale differences present in hierarchical structures is producing structurally-scaled errors. This can be achieved by dividing the error vector $\boldsymbol{e}_t = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_t$ by the structural vector $\boldsymbol{\kappa}_{str}$, where each element contains the number of nodes contributing to the forecasted error of that aggregation level, such that

$$\boldsymbol{\kappa}_{str} = S\mathbf{1}_m, \tag{4.29}$$

$$\boldsymbol{e}_t^{str} = \boldsymbol{e}_t \oslash \boldsymbol{\kappa}_{str}, \tag{4.30}$$

where $\oslash$ is a Hadamard division and $\boldsymbol{e}_t^{str}$ is the structurally scaled error vector at a time step t. The hierarchy illustrated in Fig. 4.2, for instance, gives $\boldsymbol{\kappa}_{str} = (6, 3, 3, 1, 1, 1, 1, 1, 1)$.

Structurally-scaled errors can then be employed in any given evaluation metric. We consequently define the Mean Structurally-Scaled Square Error (MS3E) as

$$\mathrm{MS3E}_{kj} = \frac{1}{h}\sum_{t=1}^{h} {e_{kj,t}^{str}}^2, \tag{4.31}$$

which can be averaged either per aggregation level or over the entire hierarchy.

4.3 Hierarchical learning

While traditional hierarchical forecasting approaches have treated forecasting and reconciliation phases separately, we propose to unify these steps under a singular machine-learning method. To introduce our approach in a step-wise manner, let us first provide a comprehensive overview of the diverse ways machine learning may be employed within the frame of hierarchical forecasting, supported by the illustrative schemes provided in Fig. 4.6.

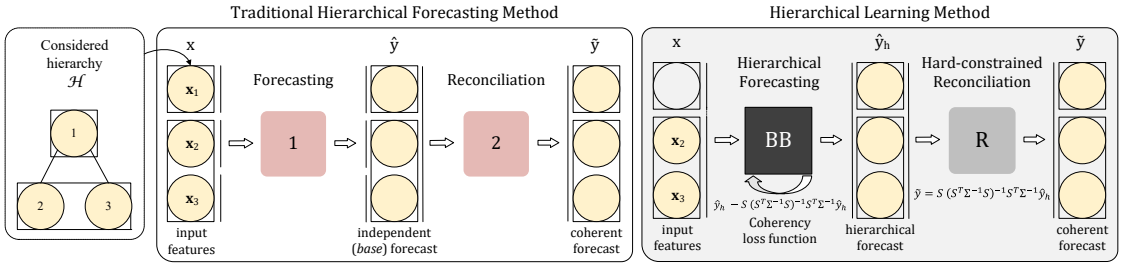


Figure 4.6: Hierarchical forecasting method with highlighted traditional approach steps (top) and proposed hierarchical learning method (bottom).

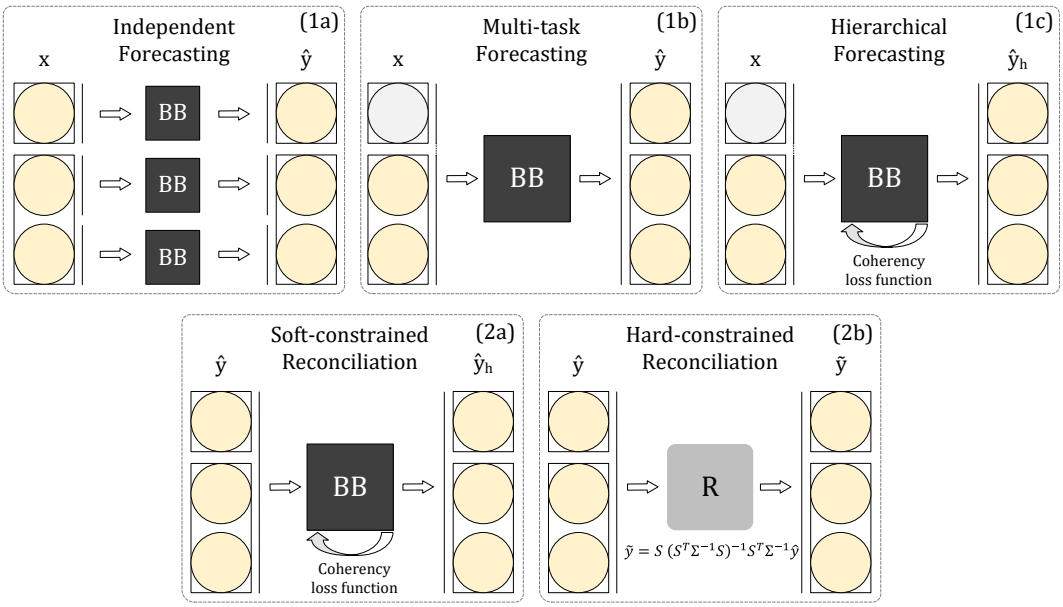


Figure 4.7: Hierarchical forecasting methods employing machine learning (here illustrated with the acronym BB, standing for Black Box) are illustrated (1a-c) along with two reconciliation approaches (2a-b). Forecasting methods encompass independent forecasting (1a), multi-task forecasting (1b), and our proposed hierarchical learning method (1c), working as a combined forecasting and reconciliation learner. Reconciliation approaches presented cover our machine learning method employed as a soft-constrained coherency enforcement over the base forecast (2a) and the traditional hard-constrained coherency enforcing method (2b).

We start by detailing the forecasting phase composing hierarchical forecasting with machine learning and continue with the description of the reconciliation step and its subsequent approaches.

4.3.1 Introducing hierarchical forecasting with machine learning

Machine learning regressors employed for hierarchical forecasting can here be employed in one of three ways.

Independent forecasting

First, with independent models each forecasting a unique node of the hierarchy, see (1a) of Fig. 4.7. The models leverage data information uniquely related to the considered node and do not exchange information with one another. They produce typical independent (*base*) forecasts of the hierarchy. Notable variations of this process involve either; (*i*) exploiting transfer learning across the models to allow exchange of information throughout the hierarchy. The work of Sagheer et al. [189] precisely employed such a scheme using a top-down approach to determine coefficients of the lower-level models as proportions of the learnt top-level one. This process secures the coherency of the forecasted tree all the while providing privacy protection of data from one site to another, as transferred model coefficients retrieve data sharing dependence. Or (*ii*) by employing a unique single-output model for the forecasting of each node of the tree, namely a multivariate learner. This allows one model to gather more information as it learns from a much larger database than independent models. However, the disadvantage of this approach originates from the generalization intention of the learnt model applied to what could be, very different forecasted processes, e.g., heterogeneous buildings. This is why this approach works best when considering processes exhibiting similar characteristics, e.g., time series range, and typical patterns, which are generally obtained through a prior clustering phase [118]. In addition, because the approach relies on the formulation of a unique model, any miss-specification could drastically impact the performance of the forecast, consequently making its design a key consideration for scientists.

The loss function of independent forecasting regressors are typically designed around a given error metric, e.g. mean squared error, describing the differences between forecasted and true values. Typically

$$\mathcal{L}^b(\mathcal{Y}, \widehat{\mathcal{Y}}|\Theta) = \frac{1}{h}\sum_{t=1}^{h}\left(y_t - \hat{y}_t\right)^2, \tag{4.32}$$

where $\mathcal{L}^b$ denotes the mean square loss function between the predicted independent *base* forecast set $\widehat{\mathcal{Y}}$ subject to a set of parameters Θ and a set of observed values $\mathcal{Y}$.

Multi-task forecasting

Second, by taking the concept of multivariate regressors even further, a multi-task regressor can be contemplated, see (1b) of Fig. 4.7. The regressor now produces a *hierarchical*, dependent, forecast of the tree as a single vector output. The model notably accepts features from the bottom layer of the tree for spatial hierarchies, as aggregate levels would provide

redundant information already present in the tree leaves. However, temporal hierarchies typically benefit from the inclusion of aggregate-level features, allowing them to exploit important information about the time series at different frequencies. Requirements for coherency are, however, not included with such a scheme.

The loss function of multi-task learners is similar to single-task ones other than considering vector rather than point errors, i.e.,

$$\mathcal{L}^h(\mathcal{Y}, \widehat{\mathcal{Y}}|\Theta) = \frac{1}{h}\sum_{t=1}^{h}\Big(\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t\Big)^2. \tag{4.33}$$

Hierarchical forecasting

This takes us to our third and last approach, crystallizing the intention and concepts behind the contributions of our work, namely, hierarchical forecasting, see (1c) of Fig. 4.7. This technique builds on the aforementioned multi-task forecasting model while extending it with the inclusion of a coherency-informed learning process thanks to a custom loss function employing established coherency taxonomy from the literature. The coherency loss function is formulated as the difference between the predicted values $\hat{\boldsymbol{y}}$ and its reconciled counterpart $\tilde{\boldsymbol{y}}$, following the reconciliation constraint of Eq. (4.12). The coherency loss function $\mathcal{L}^c$ can consequently be expressed as

$$\mathcal{L}^c(\mathcal{Y}, \widehat{\mathcal{Y}}|\Theta) = \frac{1}{h}\sum_{t=1}^{h}\Big(\hat{\boldsymbol{y}}_t - S(S^T\Sigma^{-1}S)^{-1}S^T\Sigma^{-1}\hat{\boldsymbol{y}}_t\Big)^2. \tag{4.34}$$

To combine both accuracy and coherency in the learning process of the regressor, the coherency loss is added to the hierarchical loss function defined in Eq. (4.33) forming the hierarchical-coherent loss function $\mathcal{L}^{hc}$,

$$\mathcal{L}^{hc}(\mathcal{Y}, \widehat{\mathcal{Y}}|\Theta) = \alpha\mathcal{L}_t^h + (1-\alpha)\mathcal{L}_t^c, \tag{4.35}$$

where $\alpha \in [0, 1]$ weights the hierarchical loss against the coherency loss. This avoids the over-adjustment of weights during the training of the regressor due to the addition of the coherency loss to the loss function. We typically set α to 0.75 for hierarchical forecasting to favor accuracy learning of produced predictions over coherency, yet this parameter should commonly be tuned by hyper-parameter optimization in the validation process of the model development, see Sect. 4.4.2 for implementation details.

The method regroups numerous key advantages of machine learning-based forecasts. With large and rich multi-dimensional data to learn from the regressor effectively makes use of all the information provided by the most detailed layer of the hierarchy, i.e., the tree leaves, all the while incorporating hierarchical structure information as a soft-constrained learning mechanism. Loss function augmentation via regularization and penalty methods

has grown to become the most popular way of introducing constraints in deep learning [197]–[199]. Although the approach comes at the price of sacrificing hard constraints, it has been shown that soft-constrained penalty methods perform well in practice and often exceed hard constraint methods [200], [201]. In addition, machine learning approaches are powerful at capturing non-linear relationships in the targeted predicted values. In particular, deep-learning methods are known for effective and automatic feature extraction from the data, thus reducing the need for guesswork and heuristics, which could provide a much-needed solution to the problem of non-identifiability of the covariance matrix.

Its disadvantages are similar to those of hierarchical forecasting approaches. By relying on a unique model, architecture considerations become paramount for the accurate performance of the regressor and consequently require careful, tailored tuning, e.g. with hyper-parameter grid-search.

4.3.2 Reconciliation with machine learning

While our proposed hierarchical learning approach (1c) blurs the limit between the traditionally delineated forecasting and reconciliation steps, it can also be employed as a classic reconciliation step, see (2a) of Fig. 4.7. Proposed as a soft-constrained coherency regressor, the machine learning model now takes the entire *base* forecast $\hat{\boldsymbol{y}}$ as input and outputs a coherency-informed forecast $\hat{\boldsymbol{y}}_h$. The weighting coefficient α presented in Eq. (4.34) can here be set to 0.25 to favor coherent outputs for example. The evaluation of such a scheme, lays, however, outside the scope of this work, as our contribution targets hierarchical forecasting performance evaluation on varying dimensions. This setup rather showcases the flexibility of our approach as applicable to both the forecasting and reconciliation phases of traditional hierarchical forecasting methods.

For hard-constrained reconciliation, optimal reconciliation is considered, see (2b) of Fig. 4.7. It imposes coherency to its input forecast and can be employed a posteriori to the hierarchical learning step (1c) for eventual non-coherent outputs. In addition, as an established reconciliation method, it provides a good benchmark to evaluate the performance of our proposed method to both forecasting (1c) and reconciliation (2a).

4.3.3 Structural hierarchical learning

To propose a regressor that is robust to scale differences and avoids normalization requirements we leverage the Mean Structurally-Scaled Square Error (MS3E) defined in Ref. [161] for the definition of the learning loss function. Structurally scaled errors must, indeed, be considered in hierarchical forecasting as produced predictions will possess major scale differences inherent to hierarchical structures. Designing hierarchical loss functions notably differentiates itself in this way from the scaled multi-output regressors presented earlier. To avoid a biased learning process favoring the top levels of the aggregation, where predicted values possess larger magnitudes, such structural scale differences must be adjusted. To consider this, we

employ structural scaling leveraging the aggregation-level vector

$$\boldsymbol{\kappa}_{str} = S\mathbf{1}_m. \tag{4.36}$$

For example, for the hierarchy illustrated in Fig. 4.2, $\boldsymbol{\kappa}_{str} = (6, 3, 3, 1, 1, 1, 1, 1, 1)$. The structural-hierarchical loss function $\mathcal{L}^h$ can then be expressed as

$$\mathcal{L}^{sh}(\mathcal{Y}, \widehat{\mathcal{Y}}|\Theta) = \frac{1}{T}\sum_{t=1}^{T}\Big((\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t) \oslash \boldsymbol{\kappa}_{str}\Big)^2. \tag{4.37}$$

where T is the number of time-steps, $\mathcal{L}^{sh}$ denotes the mean structurally-scaled square loss function between the predicted independent *base* forecast set $\widehat{\mathcal{Y}}$ subject to a set of parameters Θ and a set of observed values $\mathcal{Y}$. The operator $\oslash$ is a Hadamard division. It should be noted that while the proposed hierarchical forecasting loss function employs the MS3E metric, any other accuracy metric can be exploited provided they use structurally scaled error $\boldsymbol{e}_t^{str} = (\boldsymbol{y}_t - \hat{\boldsymbol{y}}_t) \oslash \boldsymbol{\kappa}_{str}$ as error reference.

The structural-coherency loss function is formulated as the structurally scaled differences between predicted values $\hat{\boldsymbol{y}}$ and their reconciled counterpart $\tilde{\boldsymbol{y}}$, following the reconciliation product of Eq. (4.12). The coherency error $\boldsymbol{e}_t^{coh}$ and subsequent scaled loss function $\mathcal{L}^{sc}$ can consequently be expressed as

$$\boldsymbol{e}_t^{coh} = \hat{\boldsymbol{y}}_t - S(S^T\Sigma^{-1}S)^{-1}S^T\Sigma^{-1}\hat{\boldsymbol{y}}_t, \tag{4.38}$$

$$\mathcal{L}^{sc}(\mathcal{Y}, \widehat{\mathcal{Y}}|\Theta) = \frac{1}{T}\sum_{t=1}^{T}\Big(\boldsymbol{e}_t^{coh} \oslash \boldsymbol{\kappa}_{str}\Big)^2. \tag{4.39}$$

Both structural accuracy and coherency losses are then combined together similarly to Eq. (4.35), forming the structural-hierarchical-coherent loss function

$$\mathcal{L}^{shc}(\mathcal{Y}, \widehat{\mathcal{Y}}|\Theta) = \alpha\mathcal{L}_t^{sh} + (1-\alpha)\mathcal{L}_t^{sc}. \tag{4.40}$$

Hierarchical learning regressors are here tailored to exploit topological structures of hierarchies, allowing the design of bespoke *structural* hierarchical models. This is accomplished by manipulating two distinct characteristics of machine learning models, namely, neuron partitions and network weights connecting partitions together. Both features are here ingeniously shaped echoing common hierarchical attributes.

Partitioning hierarchical models

We begin by defining a single-task fully connected deep neural network of 3 hidden layers and 2 inputs, as illustrated by the (a) *base* model of Fig. 4.8. To best conceptually illustrate network partitions, we propose to compare two stacked variations of said single-task regressor

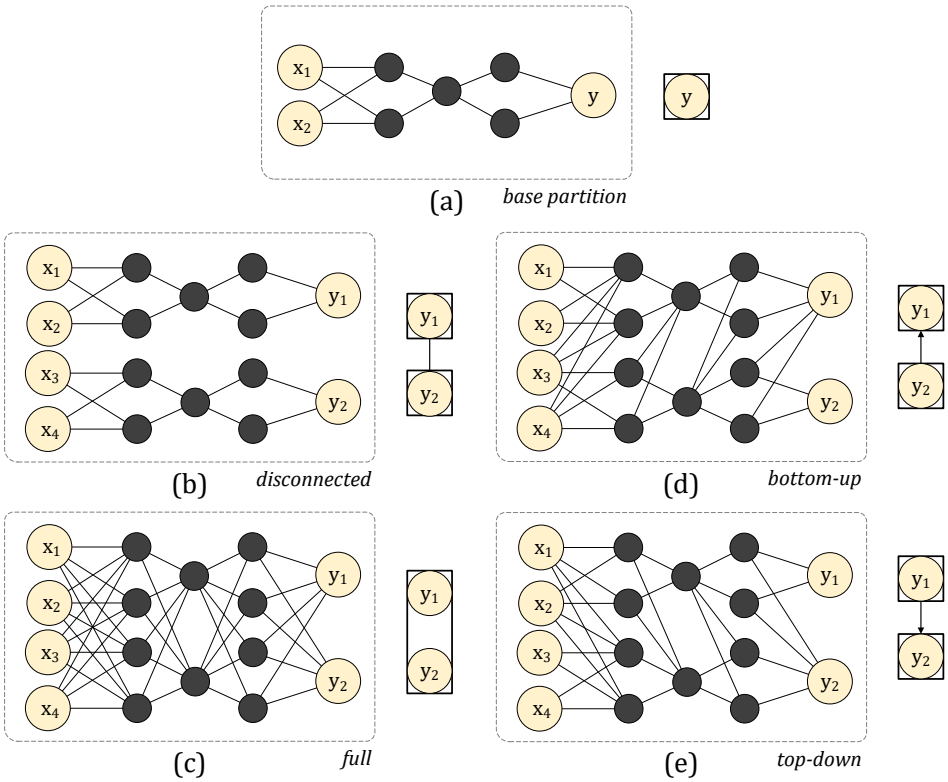


Figure 4.8: Example illustrations of (a) *base* single-task fully-connected deep neural network, composed of 3 hidden layers and 2 inputs, followed by four multi-task stacked variations of the former model where y_1 and y_2 relate to parent and child nodes of a larger hierarchical ensemble respectively. The multi-task models encompass (b) *disconnected*, (c) *full*y-connected, (d) *bottom-up*, and (e) *top-down* partition links, creating a deep neural network, composed of 3 similar hidden layers, 4 inputs, and 2 outputs.

(a), thus forming the multi-task regressors (b) and (c). While both models possess similar number of layers, neurons, inputs, and outputs, the weights connecting these elements into a common model are here notably different. In model (c) all neurons composing the hidden layers are connected to their preceding and succeeding elements, i.e., input, hidden layer, or output, while in model (b) two distinct *partitions* can be recognized formed by their parent regressor (a). Naturally, in a common prediction task, it would not be considered valuable to assemble two disconnected models together to form model (b). There would be no performance or computational gains to expect from such a setup compared to employing two independent models (a) instead for instance. In a hierarchical forecasting setting, however, it becomes on the contrary rather beneficial to gather produced outputs under the hood of one model. Indeed, in this way coherency requirements of produced predictions can be exploited by leveraging the coherency loss function. Additionally, it becomes clear that producing specific partitions over a defined neural network considerably reduces the number of weights

to update. Taking, again, the illustrated example of Fig. 4.8, model (c) possesses a total of 36 weights to update, while model (b) has only 20. This significant reduction in the number of weights to learn can serve two desirable outcomes. First, fewer weights to learn implies fewer iterations required to calibrate the model and subsequently fewer data instances, an entity that is typically expensive to gather in large and qualitative quantities. And, second, it can effectively support targeted single-task learning by helping isolate eventual conflicting multi-output predictions from one another in the model design.

We propose three partition variations, serving as alternatives to the fully-connected hierarchical model of Ref. [161], echoing established tree topologies, namely, fully-disconnected, leaf-connected, and k-level partitions, which we visually introduce in Fig. 4.9 as vertical lattices. Fully disconnected partitioning intends on mirroring independent, or base, forecasting by considering each tree node as a disjointed partition of the neural network. Similarly to the (b) model of Fig. 4.8, tree-node partitions are stacked together to form a larger, hierarchical model. We refer to this partitioning as *tree*. Leaf-connected partitioning, referred to as *cutree*, suggests linking leaf elements with identical parents together while considering the rest of the tree nodes as independent partitions. This setup is particularly interesting for hierarchies built from time-series clustering, where leaf elements typically display similar dynamics. Finally, k-level partitioning proposes to group hierarchical elements possessing alike aggregation levels into common partitions. The advantage of this approach is granting time series with similar levels of aggregation, and subsequent signal-to-noise ratios, a shared partition to capture these, possibly similar, dynamics. We refer to this partitioning as *klvl*.

Creating topological bridges

While leveraging distinct node-specific models in hierarchical forecasting is profitable, it becomes relevant to create bridges between separated layers, allowing targeted and effective learning across defined hierarchical modeling partitions. Two topological bridges have already been presented, namely disconnected partitioning, or *disc*, and fully connected ones.

To complement these we propose two well-established hierarchical connections, namely, (d) bottom-up and (e) top-down illustrated in Fig. 4.8, to bridge the disconnected partitions of y_1 and y_2 together. We here assume y_1 and y_2 to be parent and child nodes within a larger hierarchical ensemble respectively. The bottom-up setup connects the lower levels of the hierarchy with their closest higher-level elements. This implies creating neuron links, or weights, between each element of lower-level partitions, y_2, and their next respective parent partition, y_1. We refer to this partitioning as *bu*. Top-down topological bridging works in the opposite way. It creates connections between higher-level elements of the hierarchy and their associated children, i.e., here connecting y_1 partition elements to its child y_2. This partitioning is thereafter mentioned as *td*. Lastly, a subsequent topological bridge combining the above methods is put forward, i.e., bottom-up and top-down, or *butd*.

Figure 4.9 visually summarizes all of the introduced partitions and topological bridges,

exemplified on a two-level hierarchy.

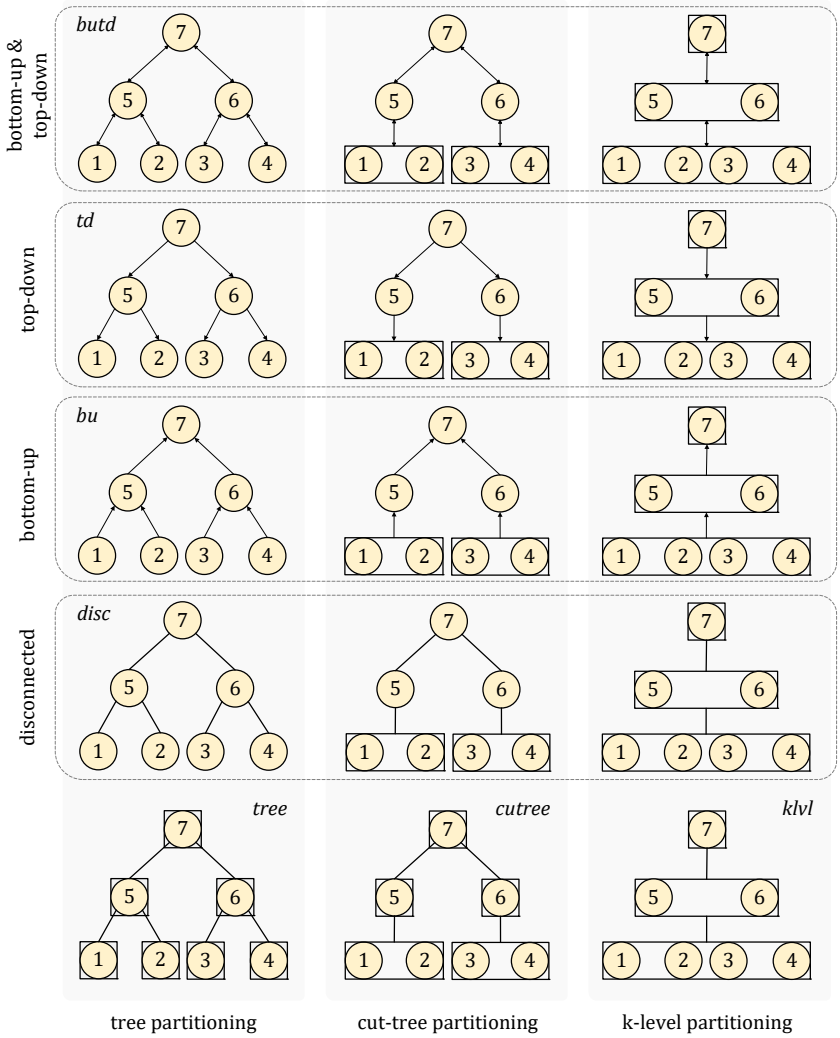


Figure 4.9: Schematic illustration of introduced model partitions (vertical) and topological bridges (horizontal) on a two-level hierarchy

4.4 Implementation

This section details the implementation-related details of our study, namely, considered case studies, hierarchical structures, and predictive-learning setup.

Our study considers two large datasets of building smart-meter measurements to demonstrate the usability and performance of our method to real-life scenarios. Case Study 1 considers the 2NECO data set. Measurements are gathered at resolutions of 10 seconds over a period of 3 years starting from January 1st 2019 to the 2nd of August 2021. Case Study 2 employs the BDG2 [20] open data set.

4.4.1 Hierarchies

Spatial hierarchies are defined by hierarchically clustering the prediction target time series, i.e., electricity demand. This step is carried out employing the Ward variance minimization algorithm [202]. The obtained hierarchy is reduced in size by *cutting the tree* using a defined distance threshold over visual inspection of the derived dendrogram. In this way, hierarchical structures located below the defined distance threshold will be clustered together, effectively reducing the number of connection nodes of the tree. Figure 4.10 illustrates the attained reduced tree of the *Fox* site of case study 2.

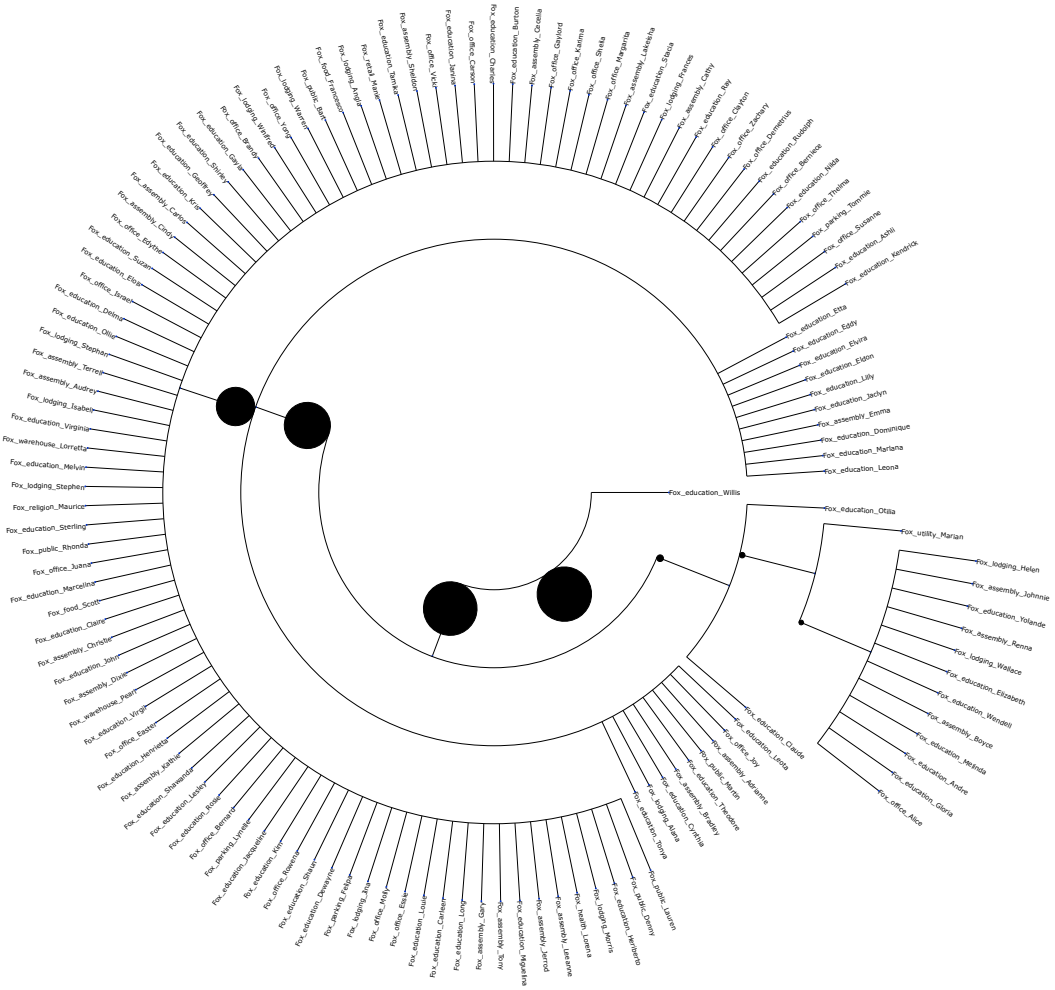


Figure 4.10: Hierarchical spatial tree structure of the *Fox* site from Case Study 2

Temporal hierarchies are considered with a horizon of one day (tree root sampling frequency) while reaching down to granularities of hourly sampling intervals (tree leaves). Aggregation levels encompass sampling frequencies every 6 and 3 hours, resulting in a tree with sampling frequencies of 1 day, 6 hours, 3 hours and 1 hour per k-level, as illustrated in Fig. 4.11. Spatio-temporal trees are then obtained as a result of the dimensional combination of spatial

Table 4.1: Characteristics of assembled hierarchy per case study

	Characteristics		Spatial	Temporal	Spatiotemporal
Case study 1	n	[#]	383	37	14,171
2NECO	m	[#]	192	24	4,608
	horizon	[hours]	1	24	24
Case study 2	n	[#]	140	37	1,998
BDG2	m	[#]	133	24	1,200
	horizon	[hours]	1	24	24

and temporal hierarchies, as detailed under Sect. 4.2.1. To limit the exponential explosion in tree size from dimensional combination, spatial trees are limited to 50 leaves in case study 2. Table 4.1 details the different characteristics of the considered hierarchies per case study.

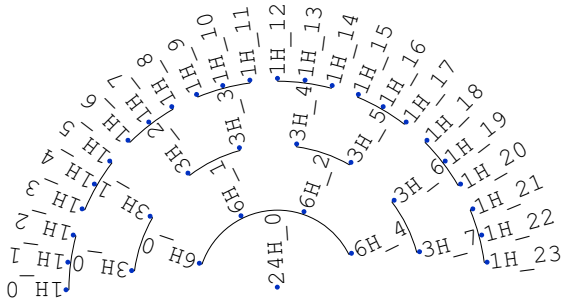


Figure 4.11: Hierarchical temporal tree structure for day-ahead forecasts

4.4.2 Model learning setup

In both case studies, we proceed to resample the time-series to hourly intervals. Time-series with no cumulative missing values larger than 2 hours are considered and smaller gaps are interpolated via a moving average using a window size of 8 hours.

Feature engineering

Data sets are then treated per dimensional batches, namely, per site, sub-site sample, or building for spatial, spatio-temporal, and temporal dimensional hierarchies respectively. Features are selected based on their Maximum Information Coefficient (MIC) [203] computed in relation to the learning target. MIC is a powerful indicator that captured a wide range of associations both functional and not while providing a score that roughly equals the coefficient of determination (R2) of the data relative to the regression function. It ranges between values of 0 and 1, where 0 implies statistical independence and 1 a completely noiseless relationship. The advantage of using MIC for feature engineering over the more commonly employed

person correlation indicator [204] is that it captures non-linear relationships present in the data, which deep-learning models are popularly capable of detecting. We retain features exhibiting MIC values higher than 0.25, as electric loads can typically become quite volatile and impede MIC values with noise.

Additionally, to feed the learner with the most relevant historical information of the predicted target, we select the 3 top auto-correlation values per temporal aggregation level above 0.25 as model input features. If no target auto-correlation value is above 0.25, we consider the most recent historical information, i.e., $t_k - 1$ where t_k is the first k-level time-step value of the predicted horizon.

Both MIC and autocorrelation selection thresholds are settings that should typically be included in the hyper-parameter optimization of the model validation phase. While evaluating the performance of hierarchical regressors over three varying dimensional considerations and two different case studies, this work considers the tuning of these thresholds to lay outside of its scope, as such computations rapidly become excessively burdensome.

Data partitioning and transformation

Training and testing sets are then defined employing TimeSeriesSplit, a times-series cross-validator of the sklearn package [205], with equal test-size in a rolling window setup.

Scaled considerations - For the hierarchical learning regressor defined in Sect. 4.3.1, we proceed to standard normalize the data per batch using batch-specific available historical information such that each batch-scaler is first fitted to the current, and past, training set. The fitted-scaler is then employed to transform batch-specific test sets, as depicted by Fig. 4.12. This process avoids data leakage situations, which refers to the inadvertent use of data

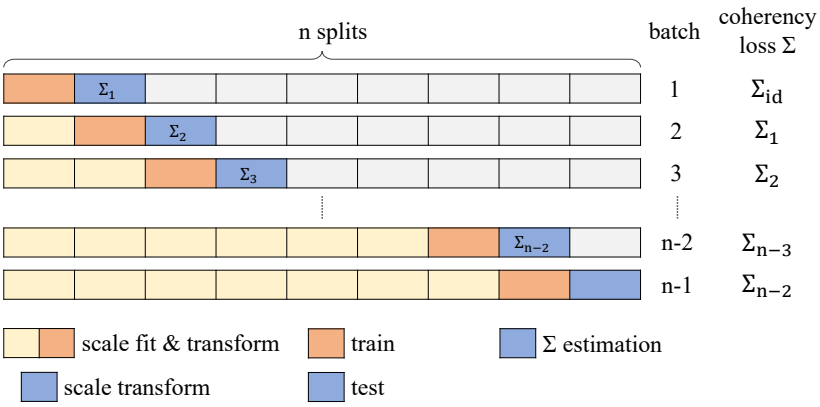


Figure 4.12: Data partitioning, transformation and covariance matrix estimation setup

from test sets, or more generally data not available during inference while training a model. This typically occurs when the data is normalized prior to partitioning for cross-validation, i.e., by performing smoothing or normalization over the whole series before partitioning

for training and testing [206]. While it benefits the performance of (deep) neural networks to normalize input features and predicted target, i.e., from unscaled $\boldsymbol{y}_x$ to scaled $\boldsymbol{y}_z$, this shatters the hierarchical relationship of the regressors' outputs; thus, affecting the soundness of the coherency loss-function. To integrate the coherency loss function in such a setting, hierarchical relationships of predicted values $\hat{\boldsymbol{y}}_z$ are restored by reverse transformation prior to coherency loss calculation. The obtained reversed-scaled prediction $\hat{\boldsymbol{y}}_x$ is reconciled to $\tilde{\boldsymbol{y}}_x$ following Eq. (4.12) and is finally re-scaled to $\tilde{\boldsymbol{y}}_z$ to calculate the coherency loss function against its original predicted self $\hat{\boldsymbol{y}}_z$.

Unscaled considerations - For the structural hierarchical learning regressor defined in Sect. 4.3.3, a priori normalization of training and testing sets is not required. We employ the heterogeneous variance approximation of the covariance matrix, which displayed good prediction performances in Ref. [161]. The heterogeneous variance includes separate variance estimates for each node. With the example hierarchy of Fig. 4.2 this gives $\Sigma_{hvar} = \text{diag}(\sigma_{11}^2, \sigma_{21}^2, \sigma_{22}^2, \sigma_{31}^2, \ldots, \sigma_{36}^2)$. The covariance matrix is recursively estimated in the test sets. For the first batch training, we employ the identity covariance estimate id as no forecasts are yet available. Each batch training i then comes with a new covariance matrix estimate Σ_i that is employed in the coherency loss function of the next training set $i+1$. This setup echoes the adaptive covariance matrix estimation proposed by [178] employed for temporal hierarchies, anchored here quite organically in the learning process of neural networks.

Coherency settings

The estimation of the covariance matrix is performed over test sets. For the first batch training, we employ the identity covariance estimate id as no forecasts are yet available. Each batch training i then comes with a new covariance matrix estimate Σ_i that is employed in the coherency loss function of the next training set $i+1$, see Fig. 4.12. This setup echoes the adaptive covariance matrix estimation proposed by [178] employed for temporal hierarchies, anchored here quite organically in the learning process of neural networks.

Designing hierarchical regressors

We select deep neural network regressors to best serve the benchmarking of hierarchical-coherent forecasts. Such machine-learning regressors possess well-developed packages supporting custom implementations that serve our approach well. The regressor is structured as a series of sequential layers designed according to three predominant features: partition widths, sequential layer depth, and topological bridges (weights). Each partition is designed as a series of sequential layers decreasing proportionally in size, from the defined input layer width w_{1p} to the desired output dimension w_D, such that

$$w_{ip} = (w_{1p} - w_{Dp}) \cdot \frac{i}{D} \ , \tag{4.41}$$

defines a partition's width in function of its design depth D. The subscripts i and p stand for the sequential layer depth and sequential layer index respectively where $i \in [1, ..., D]$ and $p \in [1, ..., P]$. Aggregating the partitions together in the regression model then produces n forecasts, ensuring that $\sum_p w_{Dp} = n$. We select the aggregated number of features per partition as the input layer width w_{1p}.

Structural regressor specifics - Between each sequential partition, we further introduce batch normalization and dropouts, both serving different purposes. Batch normalization is a technique to standardize activations in intermediate layers of deep neural networks across mini-batches. It has demonstrated improved accuracies and faster convergences due to its stabilization of the learning process [207]. Additionally, introducing batch normalization allows the in and outputs of the regression model to remain unscaled, thus retaining the hierarchical structure of the coherency-loss function. This is an essential design improvement from Ref. [161], which allows the tackling of observed faulty-coherency learning engendered from scaled trees. Dropout is a technique introduced by N. Srivastava et al. [208] designed to prevent overfitting by combining exponential numbers of combinations of neural network architectures efficiently. The term "dropout" refers to dropping out units of a neural network. Dropped-out units are removed from the network, along with all their incoming and outgoing connections, thus producing a thinned network. In essence, dropout simulates model assembling without creating multiple networks [209] while increasing convergence time. Topological bridges are then established between neurons of initially disconnected partitions following the presented connections of Sect. 4.3.3, namely disconnected (*disc*), bottom-up (*bu*), top-down (*td*), and bottom-up top-down (*butd*).

The optimal number of layers of the model is selected heuristically based on prediction performances while increasing step-wise the network's depths starting from shallow 1-layer perceptrons. This allows the selected architecture to serve an "as simple as possible yet as complex as necessary" design. Model hyper-parameters are later tuned over a concise grid encompassing loss function parameter α, activation functions, and dropout fraction, further improving the performance of the model. These tests resulted in the design of a deep neural network of 3 layers, leveraging sigmoid activation functions and dropout ratios of 0.2 on all but the last layer favoring a linear activation and no dropouts, and a retained α coefficient value of 0.75 The presented models of Sect. 4.3 were implemented in Python using the `TensorFlow` package [210].

4.5 Results and discussion

4.5.1 Hierarchical learning

We describe the outcome of the implementation of the hierarchical learning regressor here over spatial, temporal and spatio-temporal hierarchical structures per case study. In particular, we evaluate the accuracy and coherency of the forecasted building loads outlined in an

annotated heatmap and bar plot respectively, where the presented coherency loss relates solely to the output of the forecasting method, i.e., reconciliation referred to *None*, as reconciled forecasts all possess null coherency losses. Evaluated forecasting methods cover the independent (*base*), *multi-task*, and *hierarchical* forecasting methods presented under Sect. 4.3.1. Hierarchical forecasting and reconciliation methods each consider the covariance approximations presented under Fig. 4.5, i.e., ordinary least square (*id*), structural (*str*), heterogeneous variance (*hvar*), homogeneous variance (*svar*), shrunken covariance (*cov*) and shrunken k-level covariance (*kcov*). Necessary computational resources inherent to the forecasting methods are also discussed.

Case study 1 - 2NECO

Spatial - Performances of spatial hierarchical forecasts are presented under Fig. 4.13, where illustrated hierarchical losses showcase *svar* as the best hierarchical forecast performer, with and without reconciliation. The lowest hierarchical MS3E originates from *base* forecast

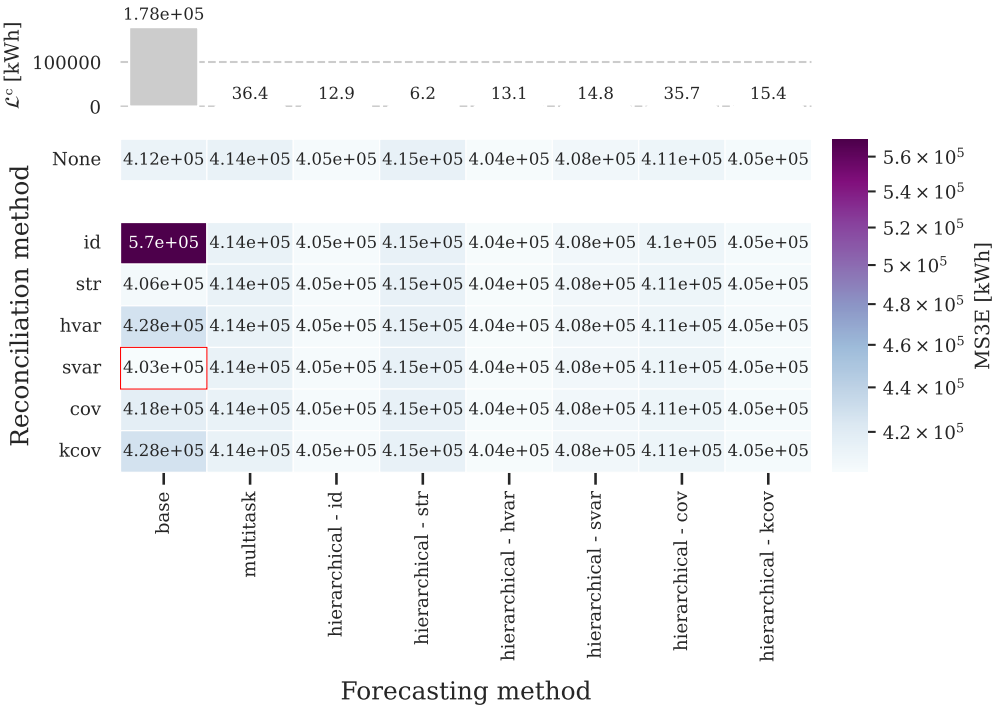


Figure 4.13: Spatial hierarchy forecasting performance of case study 1. The best accuracy performer is highlighted by a red rectangle.

reconciled with *id* covariance matrix approximation, while *hvar* and *kcov* also notably perform quite poorly for this forecasting method. Overall, the performance of the forecasts seems to rely more on the selected forecasting method rather than their reconciliation approaches. Coherency losses seem in line with expected results; *base* forecast is showcased as the most incoherent outcome, holding coherency errors ranging up to 1.783e5 kWh, while *multi-task*

and *hierarchical* regressors score MS3Es of 36 kWh and 16 kWh (on average) respectively.

Temporal - Temporal hierarchical forecasting performances, on the other hand, portray a much different behavior. As illustrated by Fig. 4.14, it is here the *base* and *multi-task* regressors that possess the lowest hierarchical losses, with 1.232e6 and 1.116e6 kWh respectively. The poorer performer without reconciliation in this setup is *svar*, with an

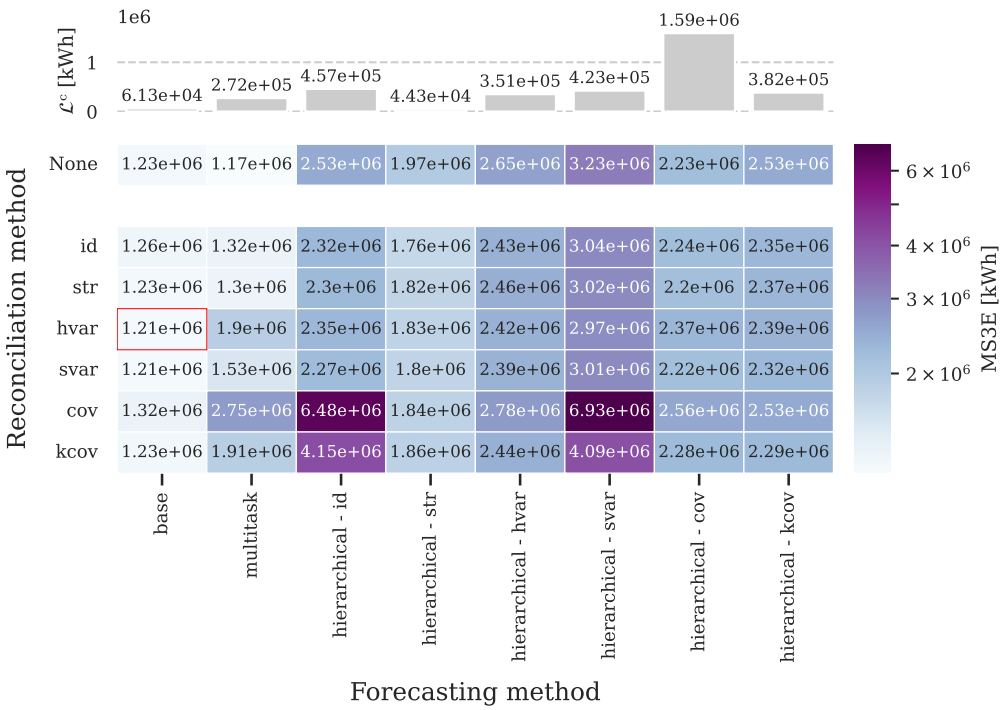


Figure 4.14: Temporal hierarchy forecasting performance of 40 buildings from case study 1. The best accuracy performer is highlighted by a red rectangle.

MS3E of up to 3.23e6 kWh. Extreme poor performances are noticeable for the *cov* and *kcov* reconciliations of *id* and *svar* forecasting methods. Overall, the performance of the forecasting methods here seems also more driven by the considered forecasting method than reconciliation. In terms of coherency, the *str* forecast exhibits the most coherent outputs next to the simpler *base* method with MS3Es of 4.43e4 and 6.13e4 kWh respectively. Other forecasting methods then follow featuring inconsistency errors ranging between 2.72e5 kWh and 1.59e6 kWh.

Spatio-temporal - Finally, spatio-temporal forecasting performances exposed in Fig. 4.15 reveal contrasting outcomes compared to previous hierarchies. First, all *cov* and *kcov* reconciliations here perform extremely poorly, irrespective of the forecasting method employed, with hierarchical losses ranging between 1.57e6 and 2.62e6 kWh. Similarly to the temporal hierarchy, *base* and *multi-task* forecasts perform overall better than *hierarchical* ones. The *multi-task* regressor without reconciliation is showcased as the best performer in this setup with an MS3E of 3.187e5 kWh. It can notably be observed here that all *hierarchical* and

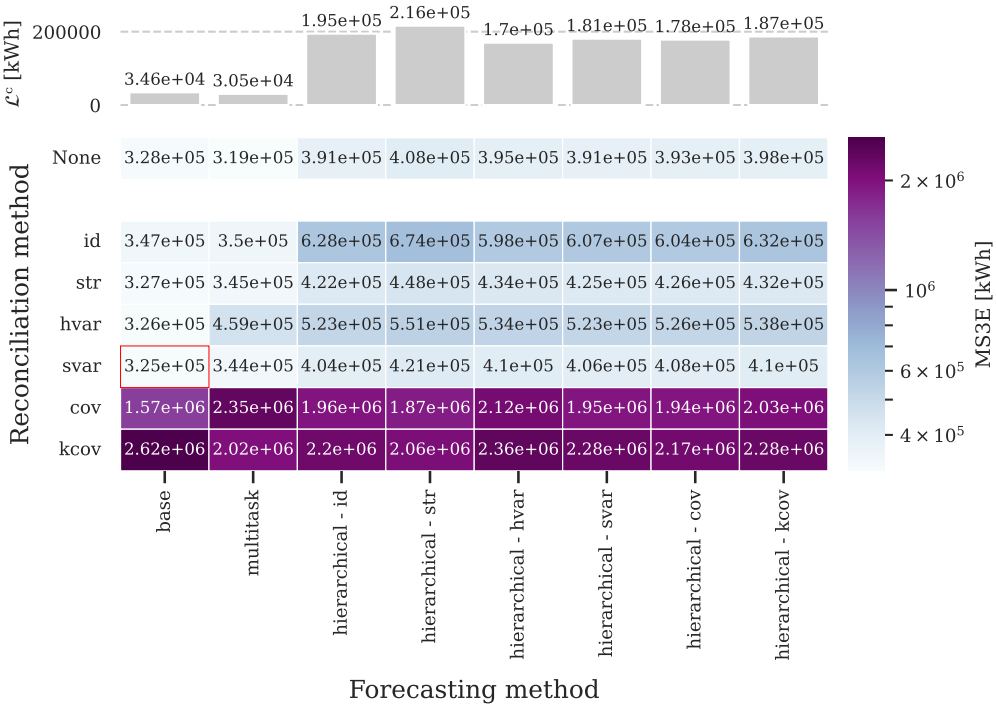


Figure 4.15: Spatio-temporal hierarchy forecasting performance of 41 buildings from case study 1. The best accuracy performer is highlighted by a red rectangle.

multi-task forecast reconciliations do not improve the accuracy of their original forecast. Additionally, exposed performances here display a much stronger dependency on the considered reconciliation approach than forecasting.

Concerning coherency losses, spatio-temporal hierarchies produce two distinct performances; where *base* and *multi-task* forecasts exhibit inconsistencies of 1 order of magnitude lower than all *hierarchical* ones, i.e., 3.255e4 kWh against 1.878e5 kWh on average.

Case study 2 - BDG2

Spatial - Concerning case study 2, the spatial hierarchical forecasting performance presented under Fig. 4.16, depicts noticeable variations from case study 1. Here, the *base* case exhibits the most accurate forecast, although at the cost of higher inconsistencies across the tree. *Multi-task* forecasts followed by structural, *str*, hierarchical ones both produce the most coherent outcomes. Surprisingly, while the *multi-task* forecast is trained without coherency information, its forecast displays the best coherency performance in this scenario. Overall, the best forecast accuracy is obtained from *base* forecasting reconciled with the *cov* approximation, while the worst performer for this scenario is the *cov* hierarchical forecasting with either *kcov* or *hvar* covariance approximations. It displays hierarchical MS3Es ranging from 611.5 to 1.727e3 kWh and coherency MS3Es varying between 0.156 and 245 kWh.

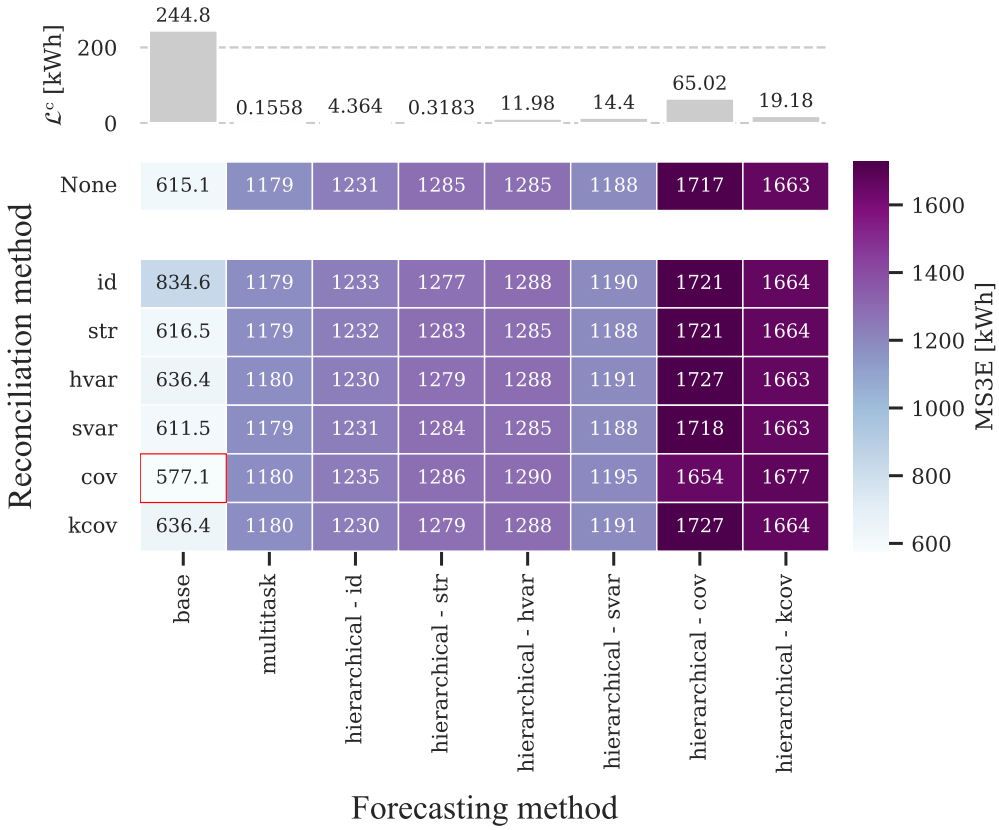


Figure 4.16: Spatial hierarchy forecasting performance of the *Fox* site of case study 2. The best accuracy performer is highlighted by a red rectangle.

Temporal - The averaged temporal hierarchical forecast performance of 66 buildings from the *Fox* site is exposed under Fig. 4.17. Forecasting performances are overall significantly worse than those of spatial-hierarchies, with hierarchical MS3Es now ranging between 1.164e3 and 6.327e4 kWh, while coherency losses fluctuate from 180 to 6.213e4 kWh; an order of magnitude about 3 times higher than temporal trees. Here, the best-performing forecast belongs to the *multi-task* forecast with no reconciliation, which also displays the highest inconsistency score. The lowest performing forecast interestingly resides with the *id* reconciliation of that same forecast. The most coherent forecast produced for temporal-trees peculiarly originate from *base* forecasts, which neither share information across the hierarchy, nor possess coherency-knowledge. Other *hierarchical* forecasts produce coherency losses ranging between 2.120e3 and 2.416e4 kWh.

Spatio-temporal - Lastly, the forecast performance of spatio-temporal structures considering 50 buildings from the *Fox* site is presented under Fig. 4.18. Similarly to the temporal-tree, hierarchical losses display much poorer performances compared to their spatial antecedent, with MS3Es ranging between 1.485e3 and extreme 9.57e9 kWh values, while coherency

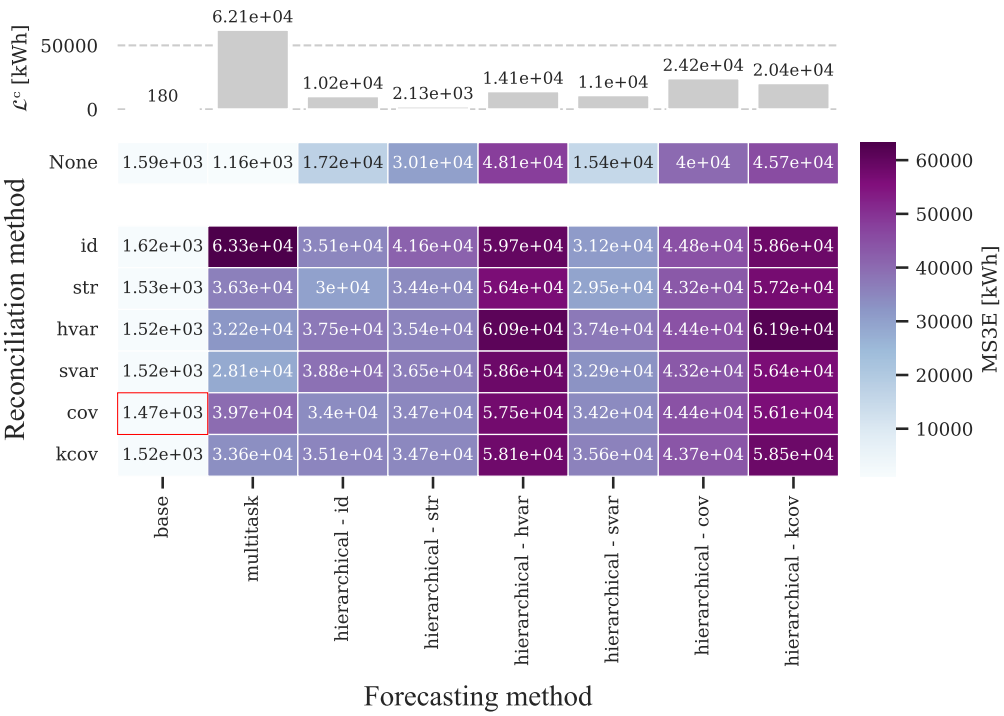


Figure 4.17: Temporal hierarchy forecasting performance of 66 buildings from the *Fox* site of case study 2. The best accuracy performer is highlighted by a red rectangle.

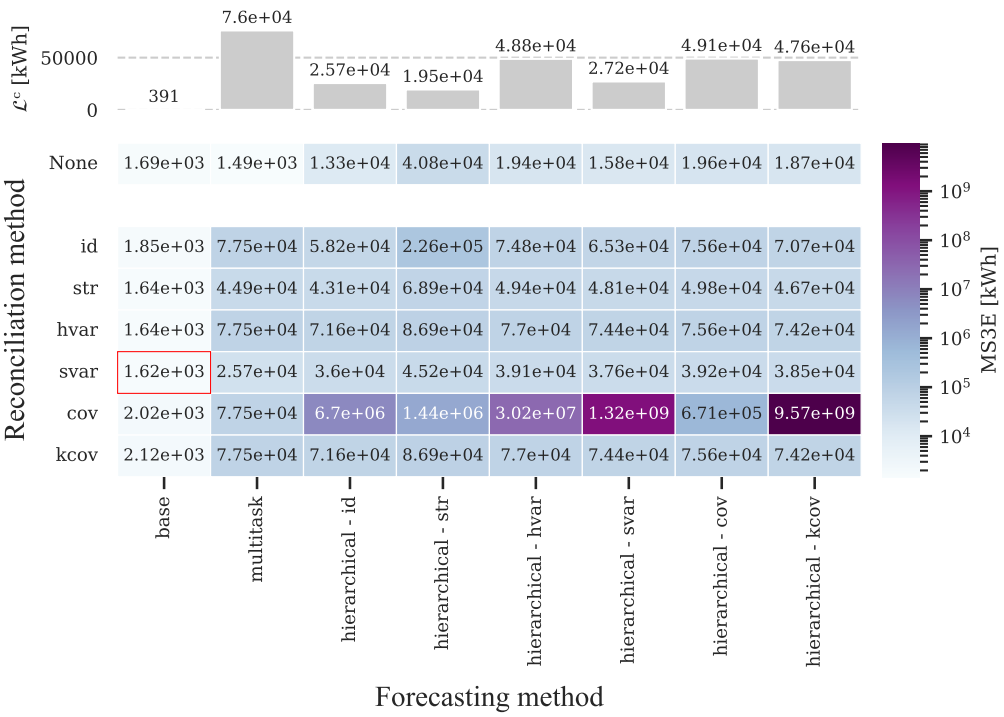


Figure 4.18: Spatio-temporal hierarchy forecasting performance of 50 buildings from the *Fox* site of case study 2. The best accuracy performer is highlighted by a red rectangle.

Table 4.2: Averaged computing times (in seconds) of evaluated forecasting methods

	tree size	forecasting method		
	n	base	multi-task	hierarchical
Case study 1	14,171	3.6	392	397.3
	383	34.9	90	96.6
	37	12.2	12	13
Case study 2	1,998	2.5	70	77
	140	22.5	70	80.6
	37	2.2	20	19.3

losses vary from 391 to 7.603e4 kWh. Mirroring the results from temporal-hierarchies, the forecasting technique withholding the lowest hierarchical loss is the *multi-task* learner without reconciliation which is also characterized by the highest coherency loss. A series of extreme poor performers are identified as a result of the *cov* reconciliation over all hierarchical-learners. Contrary to temporal-tree, reconciled forecasts performances here seem driven by the reconciliation method rather than the considered forecasting technique. Coherency scores display overall poor performances across all *hierarchical* and *multi-task* learners with losses ranging 2 orders of magnitude higher than the best case *base* regressor.

Computational prospects

Computational performances of forecasting approaches are here considered, providing a complete overview of evaluated methods. Table 4.2 presents the computation time of each forecasting method averaged over all training batches. Two anticipated findings can be noted from it.

First, the computing time is positively correlated to the size of the hierarchy. One exception seems to deviate from that rule in case study 2, between tree sizes of 1,998 and 140, which display relatively close computing times. Second, smaller regressors, i.e, *base*, train faster than larger ones, namely, *multi-task* and *hierarchical*. Both these observations can be explained by the increasing number of weights to update in the larger regressor. The more weights to update, the longer the training will take.

Although independent regressors seem attractive due to their noticeably faster computing times, it should be noted that the displayed performances depict only the average computing time of a unique independent regressor. Should such regressors not be trained and tested in a distributed computational setup, then these numbers would need to be multiplied by the hierarchy size to obtain an appropriate estimation of the required computing period.

Analysis

Although presented case studies bear varying results, these also display a number of commonalities supporting interpretation and analysis, which are here discussed.

Hierarchical-coherency value

Unifying the forecast of hierarchical structures under one regressor possesses attractive data-efficient prospects, i.e., cross-tree information exchange combined with embedded-structural learning provided from coherency loss. However, produced outcomes from hierarchical-coherent learners were only found to bring added value in one setting, namely, the spatial hierarchy of case study 1. This can be explained by the similarities in building loads of case study 1, which encompassed time series of similar patterns and dynamics, all originating from residential constructions, while case study 2 included a broader collection of construction types covering offices, college classrooms, lodging, warehouses, and parkings. Such profile diversities are challenging to learn from limited measurements, particularly for a large model involving considerable numbers of regression weights.

It can consequently be found that while the results of the spatial hierarchy of case study 1 are promising, these unveil, in fact, important challenges hierarchical forecasting must face. While some promising performances were observed, hierarchical forecasting was seen to face three important challenges put to light by our results;

(i) **A unified but arduous learning process**
Although the outcome of hierarchical learning demonstrated promising performances, identified in the spatial hierarchy of case study 1, the resulting number of weights to update and possibly conflicting forecasted outputs can become burdensome, i.e., as unveiled by the performance of the spatial hierarchy of case study 2. Indeed, with hierarchical regressors growing in size, their number of neuron connections increases by an exponential factor of 2. This renders the learning process of these models laborious as more data should support the learning of larger number of weights. Additionally, multi-output regressors are faced with the challenging task of predicting numerous outcomes which might exhibit highly different, possibly antipodal, dynamics. This also affects the learning process, which might struggle to identify these discrepancies from limited training data.

(ii) **Induced coherency over accuracy**
Overall, temporal hierarchies of the considered case studies were seen to perform significantly worse than spatial ones. This significant change can be attributed to the combination of two factors. First, the longer forecasting horizon of temporal trees compared to spatial ones, i.e., 24 hours against 1, implies that forecasts must rely on fewer data and less recent information while dealing with higher uncertainties, thus negatively affecting their performances. Secondly, building electrical loads are endowed with a periodicity that falls precisely on the forecasted horizon of 24 hours.

This consequently leads to little variations in the forecasted element of its hierarchy. And, while this characteristic is desirable for ordinary forecasting, the addition of the coherency-loss function, although weighted by the α coefficient - see Eq. (4.35), may push the regressor to produce constant predictions, tailored more to coherency than accuracy, thus resulting in unrealistic and inaccurate predictions.

(iii) **Faulty coherent-learning from normalized trees**
In some settings, hierarchical-coherent learning displayed particularly poor perfor-

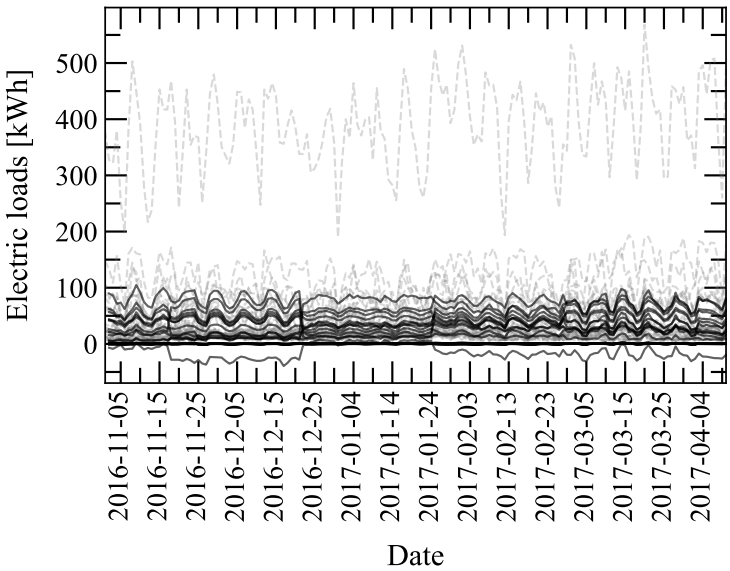


Figure 4.19: Illustration of faulty coherent-learning from normalized trees. The predicted (black) versus true (dashed grey) electric loads of the *Fox_assembly_Lakeisha* temporal hierarchical tree showcase the mirrored top-level forecast predicted in the negative domain.

mances from extreme hierarchical and coherency errors, i.e., temporal and spatio-temporal hierarchies. Following further inspection, it was noticed that these poor performers all withheld abnormal top-level forecasts which mirrored their expected true values in the negative domain, as illustrated in Fig. 4.19. These undesirable, yet peculiarly common, results can be traced back to the normalization of the target hierarchical time series. Indeed, while neural networks benefit from normalized targets, serving fair and balanced learning across the network's weights, this also shatters the coherency structure of the tree. The existing setup, detailed in Sect. 4.4.2, proceeds to tackle this issue by reverse-transforming these target values prior to the coherency constraint computation and re-scaling them for coherency loss calculation. This ensures both loss functions, namely hierarchical and coherency, see Eqs. (4.33) and (4.34) respectively, to operate on akin normalized time series. However, coherency learning can eventually produce adjustments larger than the original normalization ranges, e.g., lowering the top-level forecast $\hat{\boldsymbol{y}}_z$ fully into the negative domain such that the reverse standard transformation $\hat{\boldsymbol{y}}_x = \hat{\boldsymbol{y}}_z \cdot u + s$, where u and s refer to the mean and

standard deviation of the fitted time series respectively, also produces a fully negative reverse-scaled $\hat{\boldsymbol{y}}_x$. This evidently improper outcome consequently negatively impacts both the learning and the forecasting performance of the regressor and should be dealt with in future work.

4.5.2 Structural learning

We describe and discuss the outcome of the structural hierarchical learning regressor implementation here. In particular, we evaluate the accuracy and coherency of case study 2 (BDG2) building load forecasting outlined in varying heatmaps allowing insights into the performances of the forecast across the tree and forecasting methods. The improvement ratio brought by the coherency loss function is also highlighted both for accuracy and coherency forecast performances.

Forecast accuracy

Figure 4.20 presents the forecasted accuracy of all evaluated methods over the tree nodes, sorted by their performances across the overall hierarchy. Both extreme values of the heatmap present the tree partitioning with bottom-up (*bu*) connections and structural hierarchical-coherent loss function (*shc*) as the better performer across the forecasting methods, while the k-level partitions with top-down (*td*) topological bridges and *shc* loss function performs the worst, by an impressive 8 order of magnitude RMS3E difference. The notable better performers possess RMS3Es ranging from 42 to 100 kWh and all bear tree partitions that are either bridged in a disconnected (*disc*) or *bu* fashion. These two leading contenders each perform best with the inclusion of the coherency requirement in the loss function, i.e., *shc* versus *sh*. On the other end of the heatmap, we can regroup flawed performers ranging from 3.4e5 to 2.1e8 kWh RMS3E. The structural characteristics of these networks display k-level, cut-tree, and full partitions coupled to varying topological bridges, mostly *td* and bottom-up-top-down (*butd*).

The tendencies that can be extracted from Figure 4.20 expose that (*i*) structural models with fewer connections perform overall better than models with larger numbers of connections, and (*ii*) that within good performers, the inclusion of coherency information in the loss function improves the performance of the overall accuracy of the forecast.

Indeed, considering the number of connections per topological-design places the tree partition as the one with the least amount of connections, followed by cut-tree, k-level, and full partitions. Inter-partition connections follow the logical increasing ordering of disconnected, *bu*/*td*, and *butd*. It is consequently observed that tree partitions perform best as they result in narrower layers compared to cut-tree and k-level ones. However, the least connected model design, *tree-disc*, stands as the second best performer, thus demonstrating that some amount of information exchange between hierarchical layers, here *bu*, is valuable for the performance of the forecast. The flawed performers exhibit similar inclinations, where k-level partitions

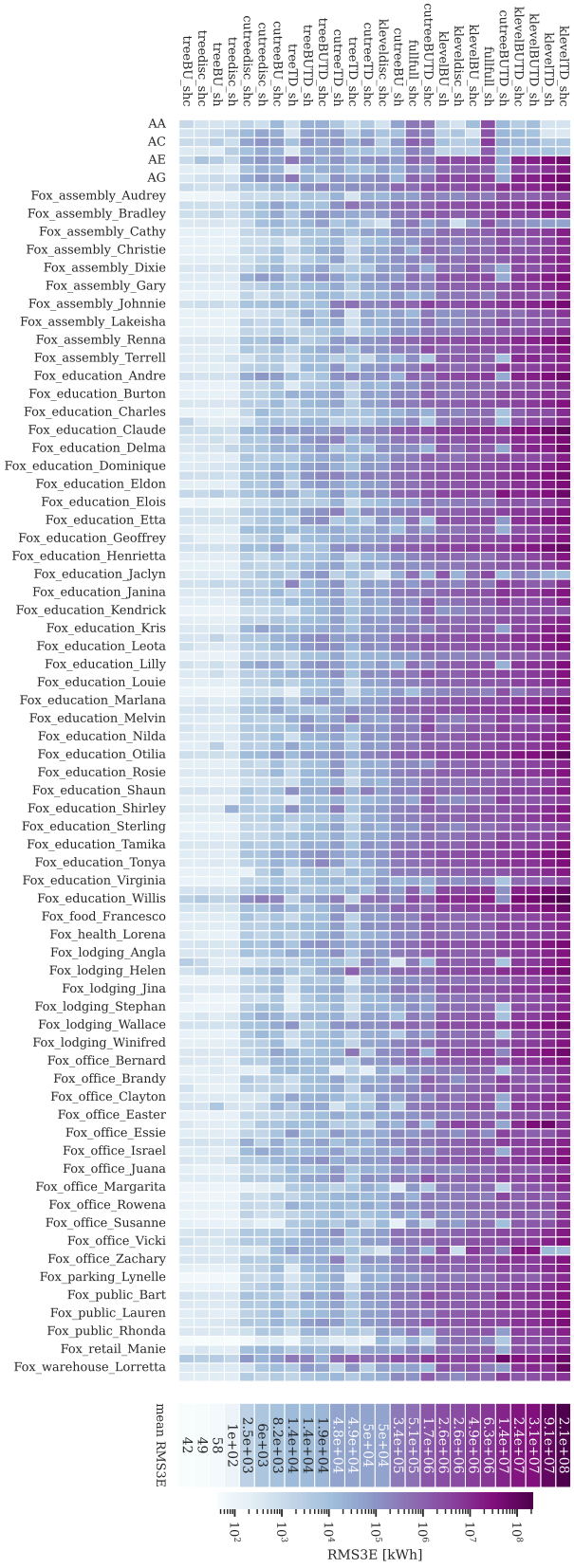


Figure 4.20: Heatmap of the accuracy performance in Root Mean Structurally-Scaled Square Error (RMS3E) across tree nodes and forecasting methods. Structural hierarchical forecasting methods described against the y-axis are designated by their respective partition, topological bridge, and considered loss function, and are here sorted according to their forecasted accuracy.

perform overall worse than cut-tree ones, which disregard the connections between leaves of dissimilar parents, thus cutting down their numbers. Then, the inclusion of coherency information in the learning mechanism of the regressor produces improved forecasts for the better half of the models, with the exception of a few cases, namely *td* and *butd* trees. This will be further discussed under Sect. 4.5.2.

It can be noticed that the *td* connections systematically perform much worse, by at least a RMS3E order of magnitude, than their *bu* counterparts, i.e., within similar layer partitionings and loss functions. The only exception that ignores this observation is the cut-tree partitioning with *sh* loss. This poorer performance of the *td* connection also seems to negatively impact the performance of its derivative *butd*. In turn, the *butd* linkage exclusively performs worse than its *bu* setup, in similar neural network designs. This topological bridge design, indeed, suffers from the influence of meager *td* performances coupled with greater numbers of weights to learn, in a data-limited setting.

Lastly, a few peculiar cases seem to produce results that deviate from observable trends. The fully connected model, although possessing a larger amount of weights by design is surprisingly not amongst the worst performers. It also displays a much more uniform forecasting performance across its hierarchy than its neighboring k-level or cut-tree models. Both observations can be explained by the fact that it possesses a number of connections in a similar order to k-level and cut-tree partitions while profiting from a more uniform design. This allows the dropout layer to reduce the network in an unconstrained manner, thus functioning under optimal conditions. Another peculiar behavior can be examined under the tree *td* with *shc* loss which displays few, but impacting, poor performances across its hierarchy, thus negatively affecting its mean accuracy performance.

Forecast coherency

While information exchange across a hierarchy in a forecasting setting has demonstrated accuracy gain potentials, the coherency improvements of the produced hierarchical time series must be evaluated. Figure 4.21 subsequently presents the coherency RMS3E, as defined by Eq. (4.39), sorted across evaluated hierarchical model designs. The coherency errors can be compared to their associated accuracy biases, thus providing a complete overview of a method's performance.

The models producing the most coherent forecast range between 5.2e2 and 9.7e2 kWh RMS3E and all benefit from tree partitions, either connected in a *bu*, *butd*, or *disc* fashion, by decreasing order of performance respectively. These top coherency performers also relate to top accuracy ones, with the exception of the tree-*butd* model. While both tree-*bu* and tree-*butd* produce slightly more coherent forecasts without the inclusion of coherency information in their loss functions, i.e., *sh*, compared to their coherent counterpart, *shc*, the coherency MS3Es are fairly similar, and theses differences can here be neglected.

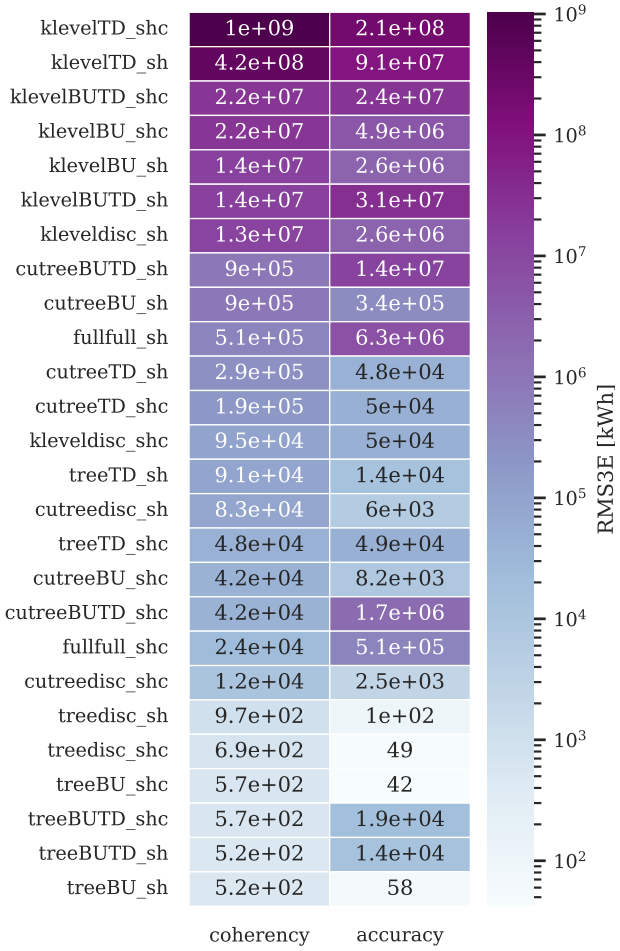


Figure 4.21: Heatmap of the accuracy and coherency Root Mean Structurally-Scaled Square Error (RMS3E) across the forecasting methods. Structural hierarchical forecasting methods described against the y-axis are designated by their respective partition, topological bridge, and considered loss function, and are here sorted according to their forecasted coherency.

The most incoherent forecasts are here produced by models with k-level partitions and *td*, *butd*, or *bu* linkages, ranging between 1e19 and 1.3e7 kWh RMS3Es respectively. Models including coherency information in their learning process here also display poorer coherency performances but are, however, associated with extremely poor accuracy performances.

A surprising observation showcases the fully connected model, *full*, and cut-tree-*butd* with coherency losses as some of the better coherency performers, in spite of their poor accuracies and large number of weights to learn. Generally, however, coherency performances display similar tendencies as their associated accuracy ones.

Coherency information value

Finally, to formally investigate the value brought by coherency information in the learning process of structural-hierarchical models, we evaluate the relative performance ratio between *sh* and *shc* loss functions of similar models. The improvement ratios for accuracy, *acc*, and coherency, *coh*, are defined as

$$r_{acc} = \frac{\mathcal{L}^{sh} - \mathcal{L}^{sh}_{coh}}{\mathcal{L}^{sh}} \ , \tag{4.42}$$

$$r_{coh} = \frac{\mathcal{L}^{sc} - \mathcal{L}^{sc}_{coh}}{\mathcal{L}^{sc}} \ , \tag{4.43}$$

where r is the improvement ratio defined by the difference between structural-hierarchical losses $\mathcal{L}^{sh}$ or structural coherent ones $\mathcal{L}^{sc}$ and their respective counterparts with the inclusion of coherency in the learning mechanism of the model, i.e., Eq. (4.40). This difference is then normalized by the reference loss, which does not consider coherency information in its loss function, i.e., Eq. (4.37). As such, positive improvement ratios relate to a performance improvement brought by coherency knowledge, whereas negative ratios point to performance regressions. This echoes the relative root mean square error (RRMSE) [176] evaluation metric typically employed to estimate the value brought by a reconciliation approach to a base forecast. The main difference in this setting is that instead of a common base forecast, we consider the structural-hierarchical forecast performance from each individual model architecture. This allows a relative performance evaluation per model architecture of the inclusion of coherency information in the learning process of the regressors.

Figure 4.22 presents the improvement ratios categorized by their network design characteristics, i.e., per partition and topological bridge arrangement. Both accuracy and coherence improvements brought by the coherency loss only display four cases of performance regression, three of which are similar: k-level-*bu*, k-level-*td*, and tree-*butd*.

Models k-level *bu* and *td* are extreme poor performers both in accuracy and coherency and can thus be disregarded in the remainder of the examination, together with k-level *butd*.

The tree-*butd* and tree-*bu* designs are the two best coherency performers, with equivalent 5.2e2 to 5.7e2 kWh RMS3Es. The accuracy of the tree-*butd* model, however, is more modest, with 1.4e4 and 1.9e4 kWh RMS3E for *sh* and *shc* corresponding losses. Due to the equivalent, top-performing coherencies of these models, their obtained negative coherency improvement ratios can thus be considered null and consequently disregarded. The remaining network characteristics all exhibit improved coherency forecasts thanks to the inclusion of the coherency loss in their learning procedure. This significant finding places structural hierarchical coherent learning as a valuable method, bringing forecasts one step closer to coherency, prior to reconciliation.

Regarding the accuracy improvements brought by the coherency loss, the tree-*td* design demonstrates an interesting behavior where coherency is improved but accuracy deteriorates.

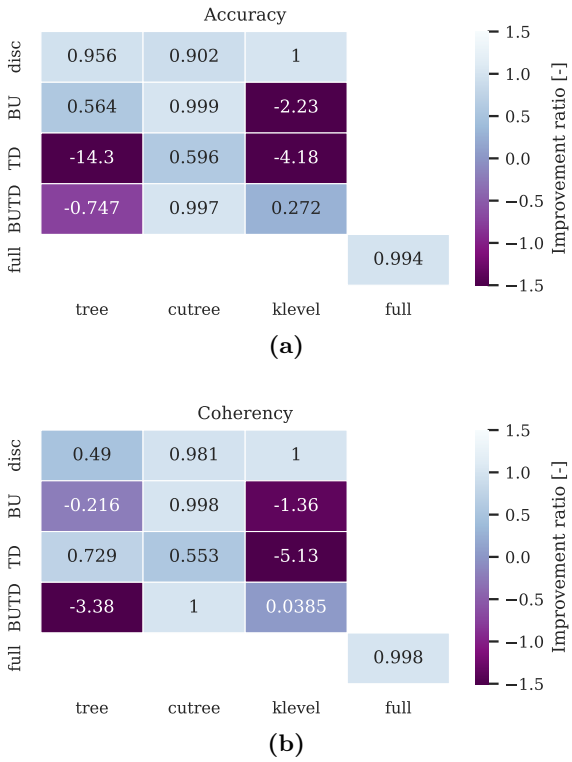


Figure 4.22: Value ratio brought by the coherency loss function for (a) the accuracy, and (b) coherency performances of the forecast.

By looking further into the accuracy performance of this method in Fig. 4.20, it was noted that the model produced overall good accuracies across its nodes with the exception of a few extreme cases, which significantly impact the overall performance of the forecast. As such, the *shc* loss function consequently pushes the forecast to a more coherent outcome than its *sh* equivalent, at the cost of a poorer accuracy across the hierarchy. A similar, but less pronounced, outcome can be observed for the tree-*butd* network, which maintains a similar coherency score but tapers its accuracy by adjusting fewer excessive forecasts. The coherency value investigation consequently allows us to claim that coherency knowledge improves the accuracy of produced hierarchical forecasts provided individual forecasts are generated within reasonable accuracy limits.

4.6 Summary

Ensuring coherent previsions of the future is crucial to support better informed and aligned decision-making processes across hierarchical structures. And while previous works have attempted to exploit spatio-temporal hierarchical reconciliation using disparate steps [164], [171], [193], [194], no common formulation of multi-dimensional hierarchical structures had, to this date, been proposed. Furthermore, traditional hierarchical forecasts use disjointed

forecasting and reconciliation processes that inherently deprive forecasting algorithms of (*i*) the benefits of information transfer across (hierarchical) models, as well as (*ii*) capitalizing on the coherency requirements of the produced forecast. This work proposes a solution to these shortcomings.

First, by formally defining multi-dimensional hierarchical structures, it extends conventional hierarchical forecasting methods, allowing the exploitation of spatio-temporal structures unified under a common frame, i.e., a unique summation and covariance matrix resulting from spatio-temporal function composition.

Second, rather than considering reconciliation a posteriori to forecasting, this work brings together independent forecasting models into a unique machine-learning regressor embedded with coherency information. This provides the regressor with (*i*) a global overview of information across its hierarchy, permitting a cross-dimensional and data-rich learning process, while (*ii*) learning coherency-requirements as a soft constraint thanks to a custom hierarchical-coherent loss function. The approach can notably be tuned thanks to an adjustable α coefficient to either consider multi-task, hierarchical or only reconciliation in its learning process. Coherency of the produced hierarchical forecasts can then be enforced as a hard constraint using established reconciliation technics. The outcome is a unified and coherent forecast across all examined dimensions, granting a common view of the future serving aligned and better decision-making. The approach provides a data-driven solution to assemble diverging parts of an organization and blend information from varying sources, hierarchy levels, or scales [164].

Third, we evaluated our hierarchical learning approach on two different case studies, across all hierarchical dimensions, considering established state-of-the-art reconciliation approaches. Results revealed spatial hierarchies to perform best while temporal and spatiotemporal structures suffered from coinciding forecasted horizon with the periodicity of electric loads from buildings. Although the value potential of hierarchical-coherent learning was observed in case study 1, the performances of the approach were quite disparate in other settings. In this regard, a comprehensive analysis was reported revealing important challenges the approach faces. In particular, dealing with predicted outputs of conflicting trends while fitting an exponentially large number of weights to the model is a recurring fragility of the approach. Additionally, correcting faulty coherency training from normalized tree structures is another frailty future work may tackle.

Finally, to undertake the above shortcomings, we investigate custom neural network designs echoing the structural topologies of hierarchies. The approach notably exploits layer partitions producing distinct model components tailored to node-specific elements, while sharing specific information across the model from varying topological bridges resulting in 13 different model architectures. Batch normalization is notably included between layers of the model, providing structural-scale robustness to the learning process, while exempting input hierarchical time

series from prior normalization and its identified subsequent biased coherency learning. We investigate all designs under two novel structurally-scaled learning functions, i.e., structural-hierarchical loss and structural hierarchical-coherent loss, leveraging the mean structurally-scaled square error (MS3E) [161], and subsequently entitle our approach *structural hierarchical learning*. The varying neural network designs are evaluated over the accuracy and coherency performances of their produced forecasts from the BDG2 [20]. Models with tree partitionings notably performed best, particularly coupled to bottom-up and disconnected topological bridges, for both structural-hierarchical and structural hierarchical-coherent losses. Links between the performance of a model and its network topology specifically revealed that (*i*) structural models with fewer connections performed overall better than models with larger numbers of connections, and (*ii*) that the inclusion of coherency information in the loss function improved both the accuracy and coherency performances of forecasts, provided individual forecasts were generated within reasonable accuracy limits.

4.6.1 Outlooks and future research

This study proposes a novel hierarchical learning method yielding important implications for forecasting theory. Indeed, by directly forecasting hierarchies this work opens the door to leveraging multi-scale and multi-frequency measurement information driving improved forecast accuracies. It notably expands and unites traditionally disjointed methods together providing a path toward a novel generation of forecasting regressors. Our work confirms the value potential brought by coherency information in structural hierarchical regressors and places structural hierarchical learning as a successful hierarchical-forecasting method, bringing forecasts one step closer to coherency, prior to reconciliation. By putting forward tailored, ingenious architectures of neural networks we effectively reduced hierarchical model complexities while serving advanced and coherency-aware hierarchical forecasts. The approach could notably support domains such as retail, stock management, and distribution networks, thanks to improved and more consistent predictions across all levels of considered hierarchies.

Meanwhile, numerous directions for future work can already be distinguished, as the approach opens the door to a variety of interesting investigations. For instance, one could imagine evaluating the effect of existing patterns on the performance of the forecast across the hierarchy. To this end, the relationship between cluster validity scores of energy patterns per node could be compared to their associated forecasting performances across varying hierarchical structures and sizes. Also, exploring the robustness of hierarchical learners to varying levels of disturbance could provide compelling insights to energy network operators seeking stable estimators. Presumably, the regressor could give more importance to the less volatile predictions of the hierarchy and compensate input disturbances by leveraging its coherency constraint. Further, formally examining how the approach scales against larger numbers of model weights, hierarchy nodes, or leaves, would provide a complete overview

of the computing performance of the method in a real-world deployment setting. Such investigations, while laying outside the scope of this work, comprise interesting pathways for future studies. Finally, comparing hierarchical learning performances against established models, i.e., grey- or white-box, that benefit from the inclusion of domain expertise to tackle targeted behaviors, such as seasonality, advances another interesting endeavor for future work.

4.6.2 Linking coherent forecasts with decision-making applications

This chapter has repeatedly stated that providing coherent forecasts at multiple levels of aggregation supports better decision-making processes. Let us showcase how, in a smart-grid setting, this is put into practice. Hierarchical and distributed optimization approaches are particularly adapted to leverage information at varying levels of the network (spatial hierarchy) and/or varying horizons (temporal hierarchy) to profit from decomposed, smaller, problems coordinated towards a consistent strategy across its considered structure.

For instance, Saad et al. [211] employed hierarchical distributed model predictive control to enable the optimization over both long (upper-layer) and short (lower-layer) time-horizons of smart grids. The upper layer worked to produce operational strategies for the grid operator and gave guidelines to the lower layer. Instead, the lower layer focused on high-power variability periods and had the responsibility to coordinate centralized optimization objectives and physical power system constraints. In another study, Jiang et al. [212] adopted a hierarchical optimization method to separate the coordination of both demand response and distributed energy resource management in a smart-building-to-grid setting. Load demands were first requested and scheduled by users, then distributed generation and storage utilities met the required demand with power.

In both settings, coherent forecasts can support the decision-making process by either or both (*i*) providing more accurate level-specific forecasts and (*ii*) producing coherent information dispatched across the energy system considered. These particularly serve multi-layer hierarchical optimization systems which efficiently deal with multi-vision control objectives while incorporating multi-level information. The next chapter presents such a prescriptive analytical application where centralized optimization problems are distributed across disconnected spatial scales of the urban energy system.

Chapter 5

From building occupants to urban energy planning

Chapter overview

- Can occupant behaviors affect urban energy planning?
- Optimal stochastic energy community design
- Distributed optimization framework for scalable urban system design
- Case study: 41 residential buildings, 2NECO
- Occupant behavior impacts energy community planning the most compared to climate and economic uncertainties
- GitHub repository: /energycommunityplanning

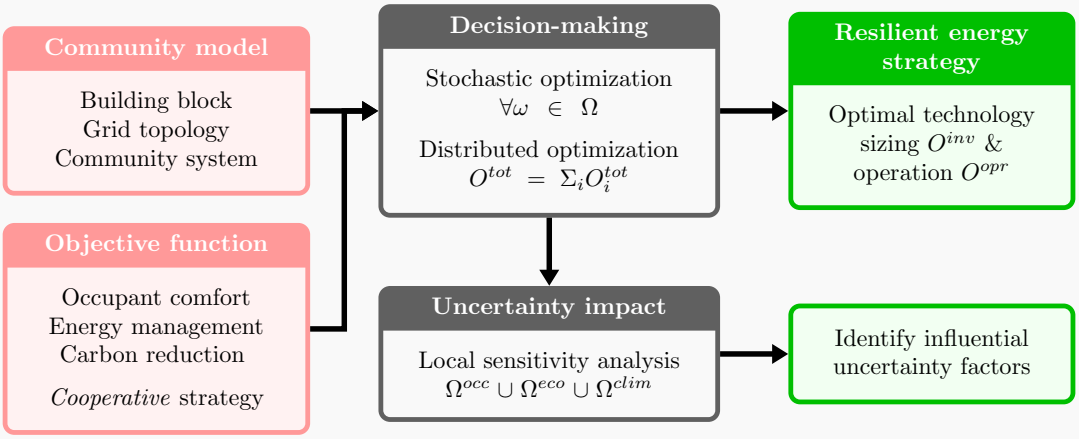


Figure 5.1: In this chapter, the decision-making process first produces a resilient energy strategy, then the impact of uncertainty factors on the community design is assessed by a local sensitivity analysis.

This chapter has been published as Leprince et al. [213].

"Remember that the happiest people are not those getting more, but those giving more."
H. Jackson Brown Jr.

5.1 Preface

Shifting our energy systems to resilient, and sustainable processes has never been more important than today. To tackle the global climate crisis and meet net-zero targets set by the European Green Deal [3], in line with the Paris agreement [4], countries around the world urgently need to decarbonize their economies by 2050. This requires them to simultaneously reduce their current energy demand while significantly increasing the penetration of renewable energy sources in decentralized energy systems [5]. Recent statistics reveal the building sector as the largest global energy-related CO_2 emission contributor [214], consequently placing it as the primary policy target of multiple regions of the globe [215]–[218]. A reliable integration of decentralized energy generation systems into the grid, such as photovoltaics, wind energy converters, geothermal heat pumps, or biomass-driven combined heat and power [5], is, however, challenging due to the variability of weather-dependent sources [219]. Couplings to energy storage utilities with robust and flexible control strategies are subsequently required to ensure energy demand and supply meet. To increase the reliability of renewable and sustainable energy systems, smart grid technologies and demand-side management approaches have been exploited over the last decades to profit from available energies more efficiently. Thereby, peaks in electricity demand can be shifted to periods where energy from intermittent renewable sources is available [220].

The concept of energy hubs and communities emerged from these ideas, to create autonomous areas optimally supplied with multiple energy sources. Energy communities are defined by the European Commission as a "legal entity which is effectively controlled by local shareholders or members, generally value rather than profit-driven, involved in distributed generation and in performing activities of a distribution system operator, supplier or aggregator at a local level" [221]. They form a combination of distribution, conversion, and storage technologies controlled to supply communal consumers of energy. Such consumers represent individual households or apartments but also large building complexes or district facilities. Typical energy communities extend over the urban energy system as districts. They integrate renewables such as photovoltaics, wind turbines, solar thermal collectors, or hybrid collectors with buildings and are connected to local and regional scale distribution technologies such as smart (micro-)grids and district heating & cooling networks [222]–[224]. The design of such communities is not a trivial task and necessitates computational methods gathering multiple energy sources and technologies while optimizing urban to user-level energy flows [225]. If done correctly, however, the pooling of communal resources into energy planning has demonstrated significant energetic and economical gains [226]. For instance, Orehouning et al. [225] showed that combining energy supply and local energy storage systems together lowered energy demand peaks on the electrical grid and reduced the overall consumption of the neighborhood. Maroufmashat et al. [227] demonstrated that developing synergies between up to three energy hubs resulted in significant economic and carbon emission reduction gains, i.e. 11% to 29%, as well as a 13% reduction in natural gas consumption.

The inherent challenge in planning and controlling such systems stems from the stochastic processes driving its three founding pillars, i.e., (*i*) investment strategies, (*ii*) renewable productions, and (*iii*) energy demands. The sources of these uncertainties can be attributed to either of these distinct phenomenons, i.e., (*i*) economic circumstances, (*ii*) climate conditions, and (*iii*) building occupant behaviors. Economic and climate-related uncertainties are important factors commonly considered in the design of urban energy systems [228]. These provide a uniform setting for the planning of energy districts and have been amply investigated in recent years [229]–[231]. Occupant behavior, on the other hand, is a notoriously heterogeneous constituent of building energy systems. Driven by multiple contextual, sociological, or psychological factors, they are exceedingly tedious to characterize [49] and have consequently become the leading source of uncertainty in predicting building energy use [50], [51] contributing to the so-called building performance gap [52]. These behaviors commonly include interactions with thermostats, plug-in appliances, operable lights, windows, or blinds. The control of window blinds by occupants may be motivated by factors such as the desire to either secure privacy or maintain view or a sense of connection to the outdoors for instance [51].

Under these circumstances, urban energy planners typically leverage energy demand measurements induced from occupant behavior to exploit samples of identified behaviors in the design phase [232]. It becomes, however, increasingly precarious to develop systems resilient to behavioral variations that are likely to come from either demographic or behavioral transformations [233]. Subsequently, there exists, to this date, no study examining the impact of varying behavioral groups on strategic urban energy planning.

This shortcoming is typically due to the scale and difference in modeled details between building to room-level energy management problems and urban energy planning ones [54]. Energy planning problems at the neighborhood, city, or country scale typically need to reduce the encompassed dimensionality through spatial and temporal aggregations to render resulting optimization problems computationally tractable. For example, the planning of a residential neighborhood would consider both a typical, representative, year of operation, to reduce the temporal dimension (horizon) of the problem, as well as aggregated energy demands from clusters of buildings or apartments to simultaneously downscale its spatial granularity [234], [235]. Yet, these necessary simplifications deprive planners from exploiting the full extent of available synergies between prosumers of energy communities. Activating untapped energy flexibility potentials such as demand-side management in the planning phase could significantly improve system efficiency and reduce planning costs. The question of relevant scale identification in urban energy planning is in fact, not a new one. Cajot et al. [55] stated that it should be regarded rather as an open question, for future research to provide planners and decision-makers with rigorous and systematic tools necessary to quantify the gains and losses of different boundaries.

5.1.1 Motivation

In this context, it becomes clear that unifying occupant and building-level information to the urban energy infrastructure can uncover additional reductions in energy demands and carbon emissions while allowing the design of energy communities resilient to the intrinsic uncertainty induced by occupants.

First, to bring these, so far, disconnected scales together, we incorporate detailed spatial information from buildings and their associated occupants into the scope of an urban energy planning problem. In particular, calibrated building models coupled to electricity base loads and set point temperature time series are exploited to include granular occupant-level information in the energy system. Second, a stochastic programming formulation is employed to account for the uncertainty present in the system and identify, in consequence, a resilient energy strategy leading to less costly and energy-demanding planning solutions. For this, climate, economic, and occupant-related uncertainties are incorporated as varying stochastic scenarios into the problem. Then, energy community design variations brought by encompassed uncertainties are subsequently examined to evaluate their respective impact and associated relevance in the context of urban planning problems. This particularly elucidates the entitled interrogation "Can occupant behaviors affect urban energy planning?". The proposed sensitivity analysis allows a relative, and holistic, appreciation of the respective impact of the varying uncertainties affecting the system, consequently placing the particular study of occupant behavior in the context of other uncertainties. Finally, to ensure the problem remains computationally tractable, we present an uncomplicated distributed optimization scheme allowing our approach to scale across the urban landscape. The distributed sub-problem architecture echoes that of typical decentralized energy management systems, thus anchoring the problem design in a real-world operational control setting, suited for field deployment.

In short, the contributions of this work can be summarized as four-fold:

1. We propose to bridge occupant behavior and strategic urban energy planning. By means of an optimal energy community design leveraging identified clusters of occupant behaviors along with sub-hourly calibrated building heat dynamics models, we effectively connect granular, detailed spatiotemporal scales of building energy systems to the coarser resolutions commonly employed for urban infrastructure planning.

2. We identify the optimal design and operation of an energy community under varying system uncertainties, i.e., occupant, climate, and economic, built upon stochastic programming.

3. We evaluate the impact of occupant behavior on the identified optimal design of the system against other uncertain factors, thus answering the questions: can occupant behavior affect urban energy planning? And, is this effect significant in the context of other system uncertainties?

4. Lastly, we propose an instinctive distributed optimization formulation, both securing the computational tractability of the problem and setting the stage for the decentralized control of the community in a real-world setting.

5.2 Energy community model

The operating limits of the energy community considered in this study are represented by the energy systems composing the urban energy infrastructure. It comprises groups of individual residential buildings and residents sharing communal resources for the optimal operation and design of the overall community. To model the investigated district, we consider three principal, and connected, modeling blocks; namely, building, grid topology, and community-level system.

The building block encapsulates residential building utilities providing the electric and heat loads induced by occupant behavior. The models are coupled to detailed thermal characterizations of the building heat dynamics from calibrated stochastic differential equations founded on heat transfer physical laws. This allows the optimization to leverage the thermal inertia of buildings in the operation planning, thus activating their energy flexibility potential. The grid topology gathers information about the low-voltage electric distribution system connecting the residential buildings together as a community along with power-line and transformer-level power constraints. Finally, the community system encompasses shared utilities operating on a medium-voltage level while ensuring the overall system connection to the high-voltage distribution grid. Figure 5.2 illustrates the energy community modeling blocks schematic.

The behavior of devices and system constraints are modeled using mixed-integer linear programming (MILP), more specifically two-stage stochastic programming. The technique has been widely employed in research to formulate optimization problems as well as perform building services energy optimizations [236]–[238]. An advantage of employing MILP stems from the general-purpose solver packages that can be exploited.

5.2.1 Formulating conventions

The following naming convention is used in the rest of the chapter:

- An italic letter stands for a scalar variable, while a bold roman letter represents a vector, commonly indexed over time steps. As an example, the symbols E and $\boldsymbol{E}$ symbolize the electric energy in scalar and vector format, respectively. These annotations are mainly employed to differentiate design variables of devices, i.e., a single value over the entire optimization period, with optimal control equipment variables, i.e., a vector with one value per sample time t_s over the optimization horizon H.
- We differentiate power from energy variables using the time derivative notation $\dot{\boldsymbol{E}}$,

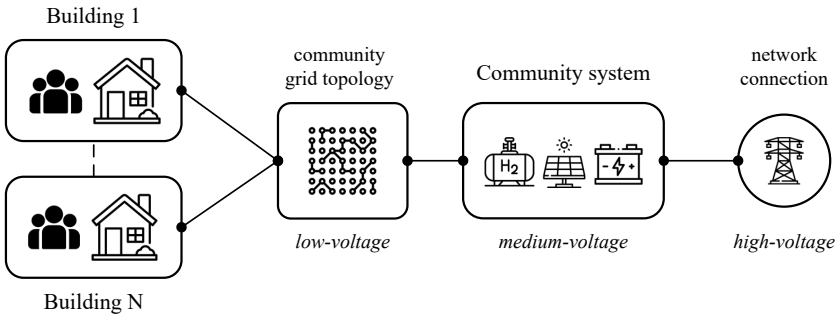


Figure 5.2: Energy community system schematic divided into building, community, and grid topology blocks.

here expressing the electric power vector. The relationship between power and energy can be derived using $\boldsymbol{E} = \dot{\boldsymbol{E}} \cdot t_s$, where t_s is the sampling time.

- A superscript d is employed to symbolize independent decision variables. As a result, $\boldsymbol{Q}^d$ indicates a controlled thermal energy manipulated by the optimization.
- Parameters of the model which depend on uncontrolled variables or external inputs are pre-calculated before the optimization starts. As an example, the COP (coefficient of performance) of the air-source heat pump is a function of ambient temperature, hence it is pre-calculated leveraging weather measurements over a typical meteorological year.
- The nature of utility investment compels us to employ binary decision variables representing the consideration or disregard of a particular device in the energy community design. For example, the binary decision variable χ_{U}^d takes a value of '1' if the unit U is included in the community design and '0' otherwise.
- All decision variables are declared as non-negative real numbers $\mathbb{R}_{\geq 0}$ such that $\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} \mid x \geq 0\}$.

The following sections will describe the structure of the three main model blocks, namely the building, the grid topology, and the community block.

5.2.2 Building system

The building block considers a series of residential storage and conversion technologies commonly employed in Dutch residential homes. Utilities considered englobe solar thermal collectors, photovoltaic panels, storage technologies with a battery and hot water tank, as well as thermal energy converters, i.e., a heat pump, and gas boiler. The building thermal dynamics are implemented built upon calibrated lumped resistance capacity models, allowing

Table 5.1: Nomenclature employed in Chapter 5: symbols

Symbol	*Description*	*Unit*
α	constant parameter	-
$\dot{\boldsymbol{E}}$	electric power	W
$\dot{\boldsymbol{Q}}$	thermal heat power	W
$\dot{\boldsymbol{E}}$	electric energy vector	kWh
$\boldsymbol{p}$	price or cost vector	€/kWh
$\boldsymbol{Q}$	thermal heat energy vector	kWh
$\boldsymbol{s}$	slack variable vector	-
$\boldsymbol{T}$	temperature vector	K
χ	existence (boolean variable)	-
η	efficiency	-
γ	power coefficient	1/h
λ	random variable	-
$\mathbb{B}$	building set	-
$\mathbb{C}$	community system set	-
Ω	set of scenarios	-
ω	scenario index	-
π	scenario realization probability	-
σ	self discharge rate	-
τ	lifetime	years
C	capacity or thermal capacity	J or W/K
E	electric energy	kWh
O	objective term	€
p	price or cost	€/kWh
Q	thermal heat energy	kWh
R	thermal resistance	K/W
r	interest rate	-
s	slack variable	-
T	temperature	K
t_s	time step	hour
U	thermal transmittance	W/K or W/m^{2}K
V	volume	L
H	time horizon	hour

Table 5.2: Nomenclature employed in Chapter 5: subscripts and superscripts

Subscript	*Description*	U	unit
b	building	w	window
BAT	battery		
BOL	boiler	*Superscript*	*Description*
ch	charge	$clim$	climate conditions
COM	community	d	decision variable
dch	discharge	eco	economic circumstances
EL	electrolyzer	inv	investment
FC	fuel cell	occ	occupant behavior
gas	gas	opr	operation
HP	heat pump	slk	slack
HV	high-voltage	T	vector transpose
HYD	hydrogen storage	tot	total
LV	low-voltage	amb	ambient
lvl	levelized costs	base	baseline
MV	medium-voltage	dist	distribution
nom	nominal	e	envelope
PV	photovoltaic panel	h	heater
SP	space heating	i	inside
STC	solar thermal collector	m	medium
T	temperature	s	sensor
TES	thermal energy storage	sol	solar

the controller to leverage the full thermal energy flexibility potential of the communal building stock. Modeled utilities ensure occupant-driven electric loads and building thermal conditions are met, all the while serving smart community energy management thanks to their connection to the low-voltage distribution network. Figure 5.3 illustrates the building model block.

Heat dynamics model

Formulating models that support the inclusion of occupant comfort needs while leveraging energy flexibility potential requires an a priori characterization of a building's thermal dynamics. This allows the thermal mass of the dwelling to be exploited as a dynamic storage asset. Conventional building control strategies do not typically consider the thermal mass of the building in their control scheme. Typically, the thermostat is set back to a lower temperature when the building is not occupied such that heating equipments are generally off during these periods. However, exploiting the building mass as a thermal storage asset

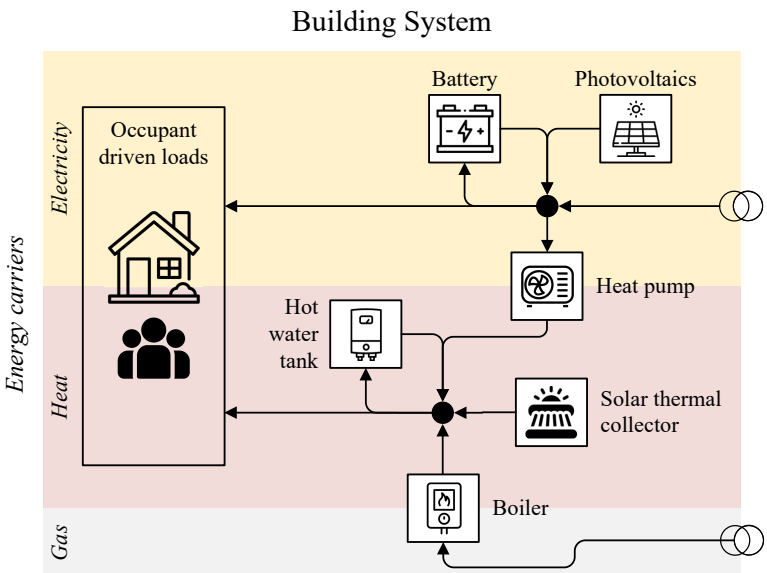


Figure 5.3: Building model block with highlighted energy carriers and connections to the community grid.

has been shown to significantly reduce operational costs in a context of varying energy prices thanks to load-shifting. These smart control strategies exploit the use of low-cost off-peak electrical energy with improved mechanical heating efficiencies at times where more favorable part-load and ambient conditions occur [239]. Additionally, the aggregation of load shifting and load curtailment demand-side management coordinated on a neighborhood scale delivers significant cost reductions [240], [241]. Recent study results showed that the implementation of the demand response program significantly reduced the demand for power during peak hours, thereby reducing the installed capacity of the combined heat and power unit [242]. This highlights the added value brought by considering energy-flexible buildings in the planning phase of building energy systems.

The building thermal models considered in this work are based on established lumped resistance capacity (RC) models [243] ranging from 1st to 5th order. Model parameters were calibrated employing the automated selection and evaluation procedure proposed and open-sourced in Chap. 3 providing 225 calibrated models of Dutch residential building heat dynamics.

The building inside temperature state variable $\boldsymbol{T}_b^d$, with associated building index b and time step t, is determined with

$$\boldsymbol{T}_b^d(t) = \boldsymbol{T}_b^d(t-1) + \Delta \boldsymbol{T}_b^d(\dot{\boldsymbol{Q}}_{SP}^d, \boldsymbol{T}^{amb}, \dot{\boldsymbol{Q}}^{sol}) \ , \tag{5.1}$$

where $\Delta \boldsymbol{T}_b^d$ is the incremental heat exchange between the building and ambient conditions defined by the RC model. It is a function of the input space heating decision variable $\dot{\boldsymbol{Q}}_{SP}^d$ and weather conditions, with ambient temperature $\boldsymbol{T}^{amb}$ and solar irradiance $\dot{\boldsymbol{Q}}^{sol}$. The complete 5th order model is formulated as sets of stochastic differential equations describing the building heat flows [122], here described in discrete time by

$$\Delta \boldsymbol{T}_b^d \equiv \Delta \boldsymbol{T}^i \ , \tag{5.2}$$

$$\begin{aligned}\text{Interior: } \Delta \boldsymbol{T}^i = & \frac{1}{R^{is}C^i}(\boldsymbol{T}^s - \boldsymbol{T}^i)t_s + \frac{1}{R^{im}C^i}(\boldsymbol{T}^m - \boldsymbol{T}^i)t_s \\ & + \frac{1}{R^{ih}C^i}(\boldsymbol{T}^h - \boldsymbol{T}^i)t_s + \frac{1}{R^{ie}C^i}(\boldsymbol{T}^e - \boldsymbol{T}^i)t_s \\ & + \frac{1}{R^{ia}C^i}(\boldsymbol{T}^{amb} - \boldsymbol{T}^i)t_s + \frac{1}{C^i}A_w\dot{\boldsymbol{Q}}^{sol}t_s \ ,\end{aligned} \tag{5.3}$$

$$\text{Sensor: } \Delta \boldsymbol{T}^s = \frac{1}{R^{is}C^s}(\boldsymbol{T}^i - \boldsymbol{T}^s)t_s \ , \tag{5.4}$$

$$\text{Medium: } \Delta \boldsymbol{T}^m = \frac{1}{R^{im}C^m}(\boldsymbol{T}^i - \boldsymbol{T}^m)t_s \ , \tag{5.5}$$

$$\text{Heater: } \Delta \boldsymbol{T}^h = \frac{1}{R^{ih}C^h}(\boldsymbol{T}^i - \boldsymbol{T}^h)t_s + \frac{1}{C^h}\dot{\boldsymbol{Q}}_{SP}^d t_s \ , \tag{5.6}$$

$$\begin{aligned}\text{Envelope: } \Delta \boldsymbol{T}^e = & \frac{1}{R^{ie}C^e}(\boldsymbol{T}^i - \boldsymbol{T}^e)t_s \\ & + \frac{1}{R^{ea}C^e}(\boldsymbol{T}^{amb} - \boldsymbol{T}^e)t_s + \frac{1}{C^e}A^e\dot{\boldsymbol{Q}}^{sol}t_s \ ,\end{aligned} \tag{5.7}$$

where the subscripts i, s, m, h, and e point to inside, sensor, medium, heater, and envelope state components respectively. For a detailed description of the models, the reader is suggested to refer to the work of Bacher and Madsen [243]. The temperature is initiated at the set point (Eq.(5.8)) and is kept within acceptable boundaries to maintain occupant comfort using Eq. (5.9),

$$\boldsymbol{T}_b^d(0) = \boldsymbol{T}_b^{set}(0) \ , \tag{5.8}$$

$$\boldsymbol{T}_b^{set} - b \leq \boldsymbol{T}_b^d \ , \tag{5.9}$$

where b is a buffer parameter commonly set to 0.5 °C.

Battery storage

Storage devices are modeled employing straightforward state variable update relationships, commonly used in control-oriented frameworks [236]. Although more complex and accurate models exist, their consequent additional computational cost should be avoided for large two-stage optimization problems such as urban energy planning. For instance, the battery BAT model employed here ignores degradation from charge and discharge cycles as well as synchronous charging and discharging behaviors.

$$\chi^d_{\mathrm{BAT}} \cdot \underline{C}_{\mathrm{BAT}} \leq C^d_{\mathrm{BAT}} \leq \chi^d_{\mathrm{BAT}} \cdot \overline{C}_{\mathrm{BAT}} \tag{5.10}$$

$$\underline{C}_{\mathrm{BAT}} \leq \boldsymbol{E}_{\mathrm{BAT}} \leq C^d_{\mathrm{BAT}} \tag{5.11}$$

$$\begin{aligned}\boldsymbol{E}_{\mathrm{BAT}}(t) = \boldsymbol{E}_{\mathrm{BAT}}(t-1) \cdot \sigma_{\mathrm{BAT}} + \dot{\boldsymbol{E}}^d_{BAT,ch}(t) \cdot \eta_{BAT,ch} \\ -\dot{\boldsymbol{E}}^d_{BAT,dch}(t) \cdot \frac{1}{\eta_{BAT,dch}}\end{aligned} \tag{5.12}$$

$$0 \leq \dot{\boldsymbol{E}}^d_{BAT,ch} \leq \gamma_{BAT,ch} \cdot C^d_{\mathrm{BAT}} \tag{5.13}$$

$$0 \leq \dot{\boldsymbol{E}}^d_{BAT,dch} \leq \gamma_{BAT,dch} \cdot C^d_{\mathrm{BAT}} \tag{5.14}$$

$$\boldsymbol{E}_{\mathrm{BAT}}(0) \leq \boldsymbol{E}_{\mathrm{BAT}}(H) \tag{5.15}$$

The main design variable is the battery capacity C^d_{BAT} which sets the limit for the amount of energy stored at any given time in Eq. (5.11), which is also lower bounded by a minimum energy state-of-charge parameter $\underline{C}_{\mathrm{BAT}}$. The existence (binary) variable χ^d_{BAT} forces the design variable either to zero or within the allowed limits through Eq. (5.10). The battery state-of-charge or stored energy is calculated using Eq. (5.12), where σ_{BAT} is the self-discharge rate and $\eta_{BAT,ch}$ and $\eta_{BAT,dch}$ stand for the battery unit charging and discharging efficiencies respectively. Equations (5.13) and (5.14) restrict the maximum charging and discharging powers using the coefficients $\gamma_{BAT,ch}$ and $\gamma_{BAT,dch}$ respectively, while Eq. (5.15) proposes a relaxed cyclic constraint for the storage system over the problem horizon H.

Thermal energy storage

The thermal energy storage (TES), i.e., a hot water tank, is modeled analogously to the battery unit. The design variable here is the energy storage capacity C^d_{TES}.

$$\chi^d_{\text{TES}} \cdot \underline{C}_{\text{TES}} \leq C^d_{\text{TES}} \leq \chi^d_{\text{TES}} \cdot \overline{C}_{\text{TES}} \tag{5.16}$$

$$0 \leq \boldsymbol{Q}_{\text{TES}} \leq C^d_{\text{TES}} \tag{5.17}$$

$$\begin{aligned} \boldsymbol{Q}_{\text{TES}}(t) = \boldsymbol{Q}_{\text{TES}}(t-1) \cdot \sigma_{\text{TES}} + \dot{\boldsymbol{Q}}^d_{TES,ch}(t) \cdot \eta_{TES,ch} \\ -\dot{\boldsymbol{Q}}^d_{TES,dch}(t) \cdot \frac{1}{\eta_{TES,dch}} \end{aligned} \tag{5.18}$$

$$0 \leq \dot{\boldsymbol{Q}}^d_{TES,ch} \leq \gamma_{TES,ch} \cdot C^d_{\text{TES}} \tag{5.19}$$

$$0 \leq \dot{\boldsymbol{Q}}^d_{TES,dch} \leq \gamma_{TES,dch} \cdot C^d_{\text{TES}} \tag{5.20}$$

$$\boldsymbol{Q}_{\text{TES}}(0) \leq \boldsymbol{Q}_{\text{TES}}(H) \tag{5.21}$$

Boiler

Gas boilers (BOL) are typically employed in Dutch residential heat systems and provide the necessary heat for space heating and hot water demand. Its main design variable is the outlet heating power capacity C^d_{BOL}.

$$\chi^d_{\text{BOL}} \cdot \underline{C}_{\text{BOL}} \leq C^d_{\text{BOL}} \leq \chi^d_{\text{BOL}} \cdot \overline{C}_{\text{BOL}} \tag{5.22}$$

$$0 \leq \dot{\boldsymbol{Q}}^d_{\text{BOL}} \leq C^d_{\text{BOL}} \tag{5.23}$$

$$\dot{\boldsymbol{Q}}^d_{\text{BOL}} = \dot{\boldsymbol{V}}^d_{gas} \cdot \eta_{\text{BOL}} \tag{5.24}$$

The output heating power of the boiler $\dot{\boldsymbol{Q}}^d_{\text{BOL}}$ is obtained by converting input gas $\dot{\boldsymbol{V}}^d_{gas}$ to heat given a fixed unit efficiency η_{BOL}.

Air source heat pump

Heat pump technologies have become a popular heating solution for buildings given their high efficiencies and environmental performances [244]. They serve as a sustainable alternative to the gas boiler thanks to reduced operational carbon emissions. The air source heat pump (HP) is implemented such that

$$\chi^d_{\mathrm{HP}} \cdot \underline{C}_{\mathrm{HP}} \leq C^d_{\mathrm{HP}} \leq \chi^d_{\mathrm{HP}} \cdot \overline{C}_{\mathrm{HP}} \ , \tag{5.25}$$

$$0 \leq \dot{\boldsymbol{Q}}^d_{\mathrm{HP}} \leq C^d_{\mathrm{HP}} \ , \tag{5.26}$$

$$\dot{\boldsymbol{Q}}^d_{\mathrm{HP}} = \dot{\boldsymbol{E}}^d_{\mathrm{HP}} \cdot \mathbf{COP}_{\mathrm{HP}} \ , \tag{5.27}$$

where C^d_{HP} is the design variable and $\dot{\boldsymbol{Q}}^d_{\mathrm{HP}}$ the output heat power. The parameter $\mathbf{COP}_{\mathrm{HP}}$ is pre-calculated using an exponential function of the ambient temperature $\boldsymbol{T}^{amb}$ and the, fixed, distribution temperature T^{dist}, as defined in Ref. [236]:

$$\begin{aligned}\mathbf{COP}_{\mathrm{HP}} = \alpha_{HP,1} \cdot \exp(\alpha_{HP,2} \cdot (T^{dist} - \boldsymbol{T}^{amb})) \\ +\alpha_{HP,3} \cdot \exp(\alpha_{HP,4} \cdot (T^{dist} - \boldsymbol{T}^{amb})) \ ,\end{aligned}$$

The parameters $\alpha_{HP,*}$ depend on the type of heat pump considered and are provided by the manufacturer.

Photovoltaic

Buildings are emerging as growing electricity prosumers who not only produce energy from distributed energy resources but also consume generated energy locally [245]. With the European Union mandating PV on all commercial, public, and new buildings by 2027 [246], photovoltaic systems will soon become an irreplaceable element of our built environment. We model PV via:

$$\chi^d_{\mathrm{PV}} \cdot \underline{A}_b \leq A^d_{\mathrm{PV}} \leq \chi^d_{\mathrm{PV}} \cdot \overline{A}_b \ , \tag{5.28}$$

$$\dot{\boldsymbol{E}}_{\mathrm{PV}} = A^d_{\mathrm{PV}} \cdot \boldsymbol{I}^{sol} \cdot \eta_{\mathrm{PV}} \ , \tag{5.29}$$

where A^d_{PV} is the upper bounded design variable by available building roof surface $\overline{A}_b$. A theoretical limitation for $\overline{A}_b$ would be the area of the roof. However, roof obstacles typically result in a few locations becoming unusable for installing PV. Additionally, PV modules are commonly mounted with an inclination, hence its calculations hinge on the geometries of the roof. The energy conversion equation is straightforwardly implemented in Eq. (5.29) employing the nominal efficiency η_{PV}.

Solar thermal collector

The solar thermal collector (STC) absorbs sunlight and converts it to heat. The amount of absorbed solar power depends on the collector's surface area A^d_{STC}, the total solar incident on the STC surface $\boldsymbol{I}^{sol}$, and the ambient temperature $\boldsymbol{T}^{amb}$, as defined in Ref. [236].

$$\chi^d_{\text{STC}} \cdot \underline{A}_b \leq A^d_{\text{STC}} \leq \chi^d_{\text{STC}} \cdot \overline{A}_b \tag{5.30}$$

$$\dot{\boldsymbol{Q}}_{\text{STC}} = A^d_{\text{STC}} \cdot \eta_{\text{STC}} \cdot (\boldsymbol{I}^{sol} - U_{\text{STC}} \cdot (T_{\text{STC}} - \boldsymbol{T}^{amb})) \tag{5.31}$$

Thermal losses of the collector are modeled in Eq. (5.31) by the term $U_{\text{STC}} \cdot (T_{\text{STC}} - \boldsymbol{T}^{amb})$, with U_{STC} being the thermal transmittance to the surroundings and T_{STC} denoting temperature of the water entering the STC.

Lastly, to consider limited roof area for both PV and STC modules, an upper bound linking both design variables is imposed such that

$$A^d_{PV} + A^d_{\text{STC}} \leq \overline{A}_b \ . \tag{5.32}$$

Energy balance

To connect considered devices of the building model block with occupant-driven energy needs a heat (Eq. (5.33)) and electricity (Eq. (5.34)) energy balance are modeled as

$$\dot{\boldsymbol{Q}}^d_{SP} + \dot{\boldsymbol{Q}}^d_{TES,ch} = \dot{\boldsymbol{Q}}^d_{\text{HP}} + \dot{\boldsymbol{Q}}^d_{\text{BOL}} + \dot{\boldsymbol{Q}}^d_{TES,dch} \ , \tag{5.33}$$

$$\dot{\boldsymbol{E}}^{\text{base}}_b + \dot{\boldsymbol{E}}^d_{BAT,ch} + \dot{\boldsymbol{E}}^d_{\text{HP}} + \dot{\boldsymbol{E}}^d_{b,out} = \dot{\boldsymbol{E}}^d_{BAT,dch} + \dot{\boldsymbol{E}}_{PV} + \dot{\boldsymbol{E}}^d_{b,in} \ , \tag{5.34}$$

where $\dot{\boldsymbol{E}}^d_{b,in/out}$ stands for the input and output power flows connecting the building model block to the low-voltage grid. The left-hand side elements of both equations denote the energy demands of the building and its utilities while the right-hand side elements provide the required energy to meet the demands. It can here be noted that while the building's space heat load $\dot{\boldsymbol{Q}}^d_{SP}$ is optimally controlled by the optimization, as a result of ensuring suitable thermal condition (Eq. (5.9)), the baseline electricity load $\dot{\boldsymbol{E}}^{\text{base}}_b$ associated with occupant-behavior is a fixed, non-shiftable, load measurements.

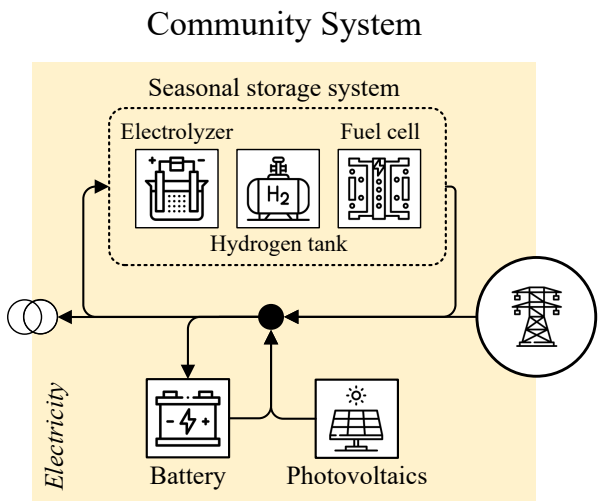


Figure 5.4: Community model block connected to the low-voltage distribution network towards buildings (left-hand side) and to the high-voltage distribution grid (right-hand side).

5.2.3 Community system

The community system exemplifies the concept of the energy hub, operating at a medium voltage network scale, linking the building community to a shared set of utilities with the high-voltage energy grid (Fig. 5.4). In this setting, the community system considers utilities that might benefit from increased performances due to their larger capacities, namely photovoltaics coupled with short and/or seasonal storage systems. The seasonal storage system is composed of three devices set up in series, i.e., an electrolyzer converting electricity to hydrogen, a hydrogen tank for long-term energy storage, and a fuel tank converting hydrogen back to electricity [247].

Models of the photovoltaic and battery community system devices are analogous to the ones presented in the building block. The main differentiation between them stems from their techno-economic parameters. We detail the particularities of the seasonal storage device to explicitly illustrate the sizing of three separate entities under a unified storage utility.

Seasonal storage system

The important value brought by the consideration of seasonal storage devices originates from offsetting seasonal mismatches between renewable energy generation and energy demands. With hydrogen storage tanks featuring negligible energy losses, they are popularly considered a promising solution for long, inter-seasonal, storage systems [248]. The hydrogen tank (HYD) is coupled to the electrolyzer (EL) and fuel cell (FC) to produce, store, and use hydrogen

respectively. A compressor device is connected to the storage tank to store hydrogen at a high pressure of 200 bars, and while hydrogen storage possesses limited energy losses, the round-trip efficiency of the seasonal storage system is much lower than that of the battery, i.e., about 35% against 95% respectively. For these reasons, hydrogen storage has been investigated as an efficient alternative to store energy for long periods of time [248], [249]. It is finally worth mentioning that due to the differences in usage between batteries and seasonal storage systems, this typically translates into larger installed capacities for seasonal storage devices.

The seasonal storage system is modeled analogously to other storage devices with the addition of three distinct design variables C^d_{HYD}, C^d_{EL}, and C^d_{FC} standing for the hydrogen tank, electrolyzer and fuel cell respectively, all linked by a unique existence variable χ^d_{HYD} in Eqs. (5.35), (5.36), and (5.37).

$$\chi^d_{\text{HYD}} \cdot \underline{C}_{\text{HYD}} \leq C^d_{\text{HYD}} \leq \chi^d_{\text{HYD}} \cdot \overline{C}_{\text{HYD}} \tag{5.35}$$

$$\chi^d_{\text{HYD}} \cdot \underline{C}_{\text{EL}} \leq C^d_{\text{EL}} \leq \chi^d_{\text{HYD}} \cdot \overline{C}_{\text{EL}} \tag{5.36}$$

$$\chi^d_{\text{HYD}} \cdot \underline{C}_{\text{FC}} \leq C^d_{\text{FC}} \leq \chi^d_{\text{HYD}} \cdot \overline{C}_{\text{FC}} \tag{5.37}$$

$$\underline{C}_{\text{HYD}} \leq \boldsymbol{E}_{\text{HYD}} \leq C^d_{\text{HYD}} \tag{5.38}$$

$$\begin{aligned} \boldsymbol{E}_{\text{HYD}}(t) = \boldsymbol{E}_{\text{HYD}}(t-1) \cdot \sigma_{\text{HYD}} + \dot{\boldsymbol{E}}^d_{EL,ch}(t) \cdot \eta_{EL,ch} \\ - \dot{\boldsymbol{E}}^d_{FC,dch}(t) \cdot \frac{1}{\eta_{FC,dch}} \end{aligned} \tag{5.39}$$

$$0 \leq \dot{\boldsymbol{E}}^d_{EL,ch} \leq \gamma_{EL,ch} \cdot C^d_{\text{EL}} \tag{5.40}$$

$$0 \leq \dot{\boldsymbol{E}}^d_{FC,dch} \leq \gamma_{FC,dch} \cdot C^d_{\text{FC}} \tag{5.41}$$

$$\boldsymbol{E}_{\text{HYD}}(0) \leq \boldsymbol{E}_{\text{HYD}}(H) \tag{5.42}$$

The electrolizer and fuel cell fix the charging $\eta_{EL,ch}$ and discharging $\eta_{FC,dch}$ efficiencies of the storage tank, and limit its inlet $\dot{\boldsymbol{E}}^d_{EL,ch}$ and outlet $\dot{\boldsymbol{E}}^d_{FC,dch}$ powers through Eqs. (5.40) and (5.41) respectively.

Power balance

The power balance equation linking the community system devices together with the low- and high-voltage distribution energy grids is modeled as

$$\dot{\boldsymbol{E}}^{d}_{MV\rightarrow LV} + \dot{\boldsymbol{E}}^{d}_{BAT,ch} + \dot{\boldsymbol{E}}^{d}_{EL,ch} = \dot{\boldsymbol{E}}^{d}_{LV\rightarrow MV} + \dot{\boldsymbol{E}}_{PV,in} + \dot{\boldsymbol{E}}^{d}_{BAT,dch} + \dot{\boldsymbol{E}}^{d}_{FC,dch} + \dot{\boldsymbol{E}}^{d}_{HV,in} \ , \tag{5.43}$$

where HV represents the input high-voltage power flow and $MV \rightarrow LV$ and $LV \rightarrow MV$ stand for the medium-to-low and low-to-medium voltage network connections respectively. BAT, PV, EL, and FC refer to battery storage, photovoltaics, electrolyzer, and fuel cell utilities on a community level, respectively.

5.2.4 Grid topology

The topology of the low-voltage distribution network, connecting the buildings forming the energy community together is here presented.

$$\dot{\boldsymbol{E}}^{d}_{MV\rightarrow LV/LV\rightarrow MV} \leq \overline{\dot{E}}_{\mathrm{MV}} + s^{d}_{MV} \tag{5.44}$$

$$\dot{\boldsymbol{E}}^{d}_{b,in/out} \leq \overline{\dot{E}}_{\mathrm{LV}} + s^{d}_{LV,b} \qquad \forall b \in \mathbb{B} \tag{5.45}$$

$$\dot{\boldsymbol{E}}^{d}_{MV\rightarrow LV} + \sum_{\mathrm{b}\in\mathbb{B}} \dot{\boldsymbol{E}}^{d}_{\mathrm{b},out} = \sum_{\mathrm{b}\in\mathbb{B}} \dot{\boldsymbol{E}}^{d}_{\mathrm{b},in} + \dot{\boldsymbol{E}}^{d}_{LV\rightarrow MV} \tag{5.46}$$

In a typical distribution network, power flows are limited by one of two factors: the maximum capacity of the power lines (Eq. (5.44)) or the maximum capacity of the micro-grid transformer (Eq. (5.45)), here represented by $\overline{\dot{E}}_{\mathrm{MV}}$ and $\overline{\dot{E}}_{\mathrm{LV}}$ respectively. Penalized slack variables s^d are additionally included to relax both LV and HV line maximum capacities in order to secure problem feasibility. Each individual building of the energy community belongs to the set $\mathbb{B}$ such that Eq. (5.45) holds for all $b \in \mathbb{B}$ and the power balance of Eq. (5.46) sums the in and output power of all the buildings belonging to the community.

5.3 Methodology

The principal objective of the energy community optimization problem is to identify the optimal selection, sizing, and operation of available building and community-level components while accounting for the uncertainty affecting the system in order to minimize the total communal costs. Optimal system designs are, however, inherently co-dependent on their associated operational strategy. As such, simultaneous optimization of control and design approaches [250] perform well for one realization of the uncertainty affecting the system, but may not for others. These approaches imply that if the suggested design is used, no other

control strategy yields better results, i.e., a lower cost function, and vice versa. Designing an energy system that is resilient to these uncertainties is an important goal of this work.

This section consequently presents the objective function of the optimization problem and details how the uncertainty affecting the energy community is captured into representative scenarios. Then, uncertainty is incorporated into the optimization problem as a two-stage stochastic programming model, and a sensitivity analysis is proposed to evaluate the specific contribution of varying uncertainty factors on the optimal energy community system design, in particular the occupant behavior. Lastly, a distributed formulation of the problem is put forth dealing with computational tractability issues endowed from large and granular optimization problems.

5.3.1 Objective function

Considering the optimal energy planning goal of the considered community, the optimization is performed for a full year encompassing all seasonal variations, while the objective is extended to varying equipment lifetimes ranging from 10 to 25 years. The total objective function O^{tot} to be minimized consists of three terms associated with levelized investment, operation, carbon emission reduction objectives, and slack penalties:

$$\min \qquad O^{tot} = O^{inv}_{lvl} + O^{opr} + O^{co2} + O^{slk} \ . \tag{5.47}$$

The levelized investment objective O^{inv}_{lvl} is calculated in Eq. (5.48) as the sum of all levelized price variables p_U, which indicates the overall cost of purchase, installation, maintenance and replacement of an arbitrary unit U belonging to the building and community system sets $\mathbb{B}$ and $\mathbb{C}$ respectively. The prices are levelized over the technology lifetime τ_U at a discount rate r such that their operation horizons serve as weights for their investment costs in the optimization.

$$O^{inv}_{lvl} = \sum_{U \in \mathbb{B} \cup \mathbb{C}} p_U \cdot \frac{r}{1-(1+r)^{-\tau_U}} \tag{5.48}$$

$$p_U = a_U \cdot D^d_U + b_U \cdot \chi^d_U \tag{5.49}$$

The device price p_U is affected by the existence variable χ^d_U and the main design variable D^d_U of the device, which refers either to the capacity C^d_U or area A^d_U of the unit. Equation (5.49) shows the calculation of the price for including a device U in the energy community. The parameter b_U defines the price for the existence of the device, while the parameter a_U represents the relative sizing price of the unit.

The operational cost O^{opr} accounts for the total amount of electricity and gas consumed by

the energy community.

$$O^{opr} = \dot{\boldsymbol{E}}^{d}_{\mathrm{HV}} \cdot \boldsymbol{p}^{T}_{el} + \sum_{b \in \mathbb{B}} \dot{\boldsymbol{V}}^{d}_{gas,b} \cdot \boldsymbol{p}^{T}_{gas} \tag{5.50}$$

The carbon emission reduction objective O^{co2} is modeled as a carbon emission penalty associated with the natural gas consumption of buildings. The term $\boldsymbol{p}_{co2}$ indicates the carbon pricing set by the European Union.

$$O^{co2} = \sum_{b \in \mathbb{B}} \dot{\boldsymbol{V}}^{d}_{gas,b} \cdot \boldsymbol{p}^{T}_{co2} \tag{5.51}$$

Lastly, the slack penalty costs O^{slk} associate a predefined slack penalty p_{slk} to all declared slack variables, such that

$$O^{slk} = p_{slk} \cdot s^{d}_{MV} + p_{slk} \cdot \sum_{b \in \mathbb{B}} s^{d}_{LV,b} \ . \tag{5.52}$$

The penalty value is set high such that the optimization would only consider relaxing the constraints in cases of problem infeasibility.

It should be noted that the energy community optimization problem shares a unique, global objective function O^{tot}. Defining a global objective function ensures a cooperative behavior between all elements of the community, i.e., building and energy community blocks, working toward the reduction of the aggregated costs of the system, rather than sub-optimal individualistic objectives. Leveraging such cooperative behaviors between individual agents of an energy system is recognized to substantially improve economic and energetic performances [251].

5.3.2 Representative scenario identification

To perform urban energy planning in a computationally tractable manner, it becomes necessary to trim encompassed spatiotemporal dimensions to a reduced, but representative, number of spatiotemporal frames. Indeed, urban distributed energy resources design procedures commonly cluster encompassed input data to a typical reference year, assumed constant over the lifetime of the energy system, e.g., 25 years [252]. Downscaling the spatial resolution of the energy community by clustering buildings to fewer representative ones would, however, deprive the optimization of the diversity of information-rich occupant behaviors and varying building thermal dynamics. The purpose of the present work is to consider the complete building community stock in the energy planning process, allowing building energy flexibility activations to be exploited on an aggregated urban scale while bridging the pluralities of occupants to urban infrastructure planning. The determination of characteristic years of measurements across buildings, weather, and economic conditions provides scenarios serving

both two-stage stochastic programming (Sect. 5.3.3) and the latter sensitivity analysis (Sect. 5.3.4). In order for these scenarios to approximate the underlying uncertainty as closely as possible, many scenarios are initially bootstrapped (Sect. 5.3.2), then reduced to few representative ones by clustering (Sect. 5.3.2).

Scenario generation using seasonal bootstrapping

In this study, three categories of parameters are selected as uncertain, namely building electrical load demands and set-point temperatures for occupant behavior, energy prices for economic conditions, and ambient temperature and solar irradiance for weather conditions.

To artificially increase the number of years of collected data while retaining the auto-correlations of energy consumption profiles and day-ahead pricing, we apply a seasonal block bootstrapping technique to generate 1000 synthetic years of data. Block bootstrapping for seasonal time series has been found suitable for periodic time series with fixed-length periodicities of arbitrary block and sample size [253]. Given the diurnal patterns of building energy consumption and day-head electric forecasts, we consider block samples of 24 hours that are sampled across the entire data set to secure the correlations between building energy needs and weather and economic conditions. The blocks are bootstrapped over a seasonal-dependent sub-space to retain the periodic behaviors present in the original data. Thereby, weekday and weekend variations are preserved while the sampling space is restricted to a region of 8 weeks surrounding the sampling block [254].

Scenario reduction using clustering

Gathered scenarios are then reduced to a more manageable number employing k-medoids clustering [255] to obtain identifiable cluster centers (*medoids*) and associated probabilities [256]. The advantage of uncovering medoids, which are superimposed on existing input data, is that it preserves the volatility of the original input data as opposed to k-means clustering which produces *centroids* that are averages of their cluster members, thus resulting in the curtailment of their individual stochastic properties.

Identifying a suitable number of clusters is commonly performed from cluster intra-class homogeneity and inter-class separation indexes that assess the validity of obtained clusters. Such metrics focus on the grouping validity of input scenarios but cannot deliver information on the grouping validity of the resulting optimal policy of their associated optimization problems. Presumably, one might obtain 25 clear representative scenarios characterized by high intra-class homogeneity and inter-class separation, yet produced policies might be very dissimilar within scenarios of a cluster just as inter-cluster ones might produce very similar policies. With this in mind, we consider 10 scenarios to provide a sufficient number of samples and a representative description of the diversity of existing system uncertainty, with no need for cluster validity assessment. We consequently reduce the 1000 bootstrapped scenarios to $N_\Omega = 10$ distinct clusters, where N_Ω is the number of scenarios considered and

Ω is the set of scenarios. Their associated probabilities $\pi(w)$ is subsequently obtained from

$$\pi(\omega) = P(\omega|\lambda = \lambda(\omega)), \quad \text{where} \quad \sum_{\omega\in\Omega} \pi(\omega) = 1 \ , \tag{5.53}$$

and $\lambda(\omega)$ is a random variable associated with a scenario index ω, while the scenario realization probabilities are represented by $\pi(\omega)$. The probabilities $\pi(\omega)$ correspond to the relative cluster sizes obtained via k-medoids clustering.

5.3.3 Stochastic programming formulation

Introducing uncertainty in the design of energy communities involves a decision-making problem structure well suited to a two-stage stochastic programming model [257]. Indeed, the design problem features the concurrent determination of both design and operation variables, which are commonly decided in different stages. This means that decisions on the design variables must be adequate to adapt to varying realizations of energy demand and supply profiles over the year. Consequently, design variables D_d and their associated existence variables χ^d are categorized as first-stage variables, to be decided prior to the resolution of uncertainty, and all other operational (decision) variables are considered second-stage variables, which can later be adapted in function of the uncertainty scenario unfolding. A two-stage stochastic programming approach thus undertakes the simultaneous determination of the optimal configuration of a distributed energy system given varying (optimal) operating conditions [258].

By incorporating the uncertainty into the mixed-integer optimization problem, a two-stage stochastic programming problem is formulated as follows [259]:

$$\begin{aligned} \min \quad & \overbrace{O^{inv}_{lvl}}^{1^{\text{st}}\text{-stage costs}} + \overbrace{\sum_{\omega\in\Omega} \pi(\omega)\cdot(O^{opr}(\omega) + O^{co2}(\omega) + O^{slk}(\omega))}^{\text{expected } 2^{\text{nd}}\text{-stage costs}} \ , \\ s.t. \quad & \boldsymbol{A}\boldsymbol{x}^d = \boldsymbol{b} \ , \\ & \boldsymbol{T}(\omega)\boldsymbol{x}^d + \boldsymbol{W}(\omega)\boldsymbol{y}^d(\omega) = \boldsymbol{h}(\omega) \ \ \forall\omega\in\Omega \ , \end{aligned} \tag{5.54}$$

where $\boldsymbol{x}^d$ gathers the 1st-stage decision variables and $\boldsymbol{y}^d(\omega)$ the 2nd-stage decisions. The matrices and vectors $\boldsymbol{A}$, $\boldsymbol{T}(\omega)$, $\boldsymbol{W}(\omega)$, $\boldsymbol{b}$, $\boldsymbol{c}$, $\boldsymbol{q}(\omega)$, and $\boldsymbol{h}(\omega)$ are known parameters of the system, that can be gathered from Eqs. (5.1)-(5.46) and (5.48)-(5.51).

The objective thus becomes to determine the first-stage design and existence variables by taking the sum of the deterministic first-stage costs, namely, the levelized investments costs, defined in Eq. (5.48), and the expected second-stage operational costs corresponding to the sum of the operation and carbon emission costs, see Eqs. (5.50) and (5.51) respectively, weighted by their respective realization probabilities $\pi(\omega)$.

5.3.4 Uncertainty impact on energy community design

While the consideration of the uncertainty in the form of a stochastic program allows the identification of the optimal policy given the probability of varying scenarios unfolding, it, however, does not inform energy planners on the relative *impact* of its considered uncertainty factors. To evaluate the relative influence of the different categories of uncertain parameters on the design of energy communities, in particular occupant behavior, it becomes necessary to undertake a sensitivity analysis. There are two principal methods for sensitivity analysis, local and global ones. Local sensitivity analysis methods typically analyze how the uncertainty in each input parameter affects an output of interest [260]. Uncertain parameters are commonly altered one at a time with other parameters fixed at their nominal values, or through the definition of scenarios, i.e., combinations of uncertain parameter values [228]. Global sensitivity analysis on the other hand considers all of the input parameters simultaneously. The impact of each input parameter on the performance indicator of interest are commonly evaluated by variance-based methods. These are, however, computationally expensive for urban energy planning due to their large number of inputs [261], and require the characterization of uncertainties a priori, else would result in false rankings when employing generic uncertainty ranges [262].

To keep the scope of this work within manageable limits, we consider a local sensitivity analysis method performed over all-encompassed uncertainty factors.This supports the assessment of the impact of occupant behavior on urban energy planning while providing a relative evaluation in the context of other uncertainties.

Local sensitivity analysis

To undertake the sensitivity analysis, uncertainty factor-dependent scenario sub-sets and their nominal values must first be identified. The scenario ensemble is thus divided into three distinct subsets

$$\Omega = \Omega^{occ} \cup \Omega^{eco} \cup \Omega^{clim} \ , \tag{5.55}$$

corresponding to occupant, economic, and climate conditions respectively. The nominal scenario ω_{nom} for each uncertainty factor is identified from k-medoid clustering with $k = 1$ subsets of Ω and extracting its medoid scenario. This is performed over the subsets $\Omega^{eco} \cup \Omega^{clim}$, $\Omega^{occ} \cup \Omega^{clim}$, and $\Omega^{occ} \cup \Omega^{eco}$ for occupant, economic, and climate conditions uncertainty factors respectively.

Then, the influence of varying uncertainty factor-dependent scenarios is assessed in a one-at-a-time fashion via the optimal design variables retained by the energy community planning

problem. The problem (5.54) is then iteratively solved for either of the following variations

$$\Omega_{ijl} = \Omega_i^{occ} \cup \Omega_j^{eco} \cup \Omega_l^{clim} \begin{cases} \forall i \in \Omega^{occ}, j = \omega_{nom}^{eco}, l = \omega_{nom}^{clim} \ , \\ \forall j \in \Omega^{eco}, i = \omega_{nom}^{occ}, l = \omega_{nom}^{clim} \ , \\ \forall l \in \Omega^{clim}, i = \omega_{nom}^{occ}, j = \omega_{nom}^{eco} \ . \end{cases} \tag{5.56}$$

Note that the evaluated set Ω_{ijl} becomes singular, thus the stochastic program (5.54) becomes a deterministic problem as the 1st and 2nd stage costs are evaluated over a unique scenario.

It should be acknowledged, however, that such a setup disregards the existing inter-correlations between the considered uncertain parameters. For instance, the energy flexibility leveraged in demand-side management applications from occupant-established comfort buffer regions is known to possess a strongly correlated relationship to the energy price levels [263]. Similarly, weather conditions typically affect occupant thermal preferences. This implies that separating these uncertainties in factor-dependent scenarios is intrinsically flawed, making it tedious to differentiate their independent contributions to the problem policy. We consider this approximation, however, to be a necessary simplification for the evaluation of the distinct impact of these uncertainties.

5.3.5 Distributed optimization

Stochastic optimization problems are notoriously known for their associated computational burden [264]–[266]. As problems get larger, the state space is proportionally multiplied by the number of considered scenarios, and computing intractability problems quickly arise. To alleviate this charge, we propose to partition the problem into smaller sub-problems. This allows scaling of the considered energy community system as sub-problems are easier to solve. simultaneously, this increases the overall system resilience in a control setting should one of its components (sub-problems) fail or become obsolete. The computational load of the problem is subsequently eased by dividing the initially larger problem into multiple smaller ones [267], resulting in a *distributed* optimization problem.

The subsequent partitioned problems, however, require careful coordination not to result in conflicting local actions and threaten the global system stability. To this end, we consider an uncomplicated sequential solving approach [267], allowing information exchange between the different sub-systems while keeping the complexity inherent to each model undisclosed [268]. During this phase, each problem partition is solved sequentially, first gathering the previously predicted operating plans of other sub-systems, then computing the local strategy [269]. The process is executed in an iterative manner until the stopping criteria is met, i.e., either by a predefined number of iterations or a cost variation threshold. This results in an information-optimized communication system requiring the sole transfer of aggregated local energy flows between sub-systems in the coupling constraint. This setup notably preserves the privacy related to individual building energy demands, while preparing the stage for the real-world deployment of the energy management system of the community.

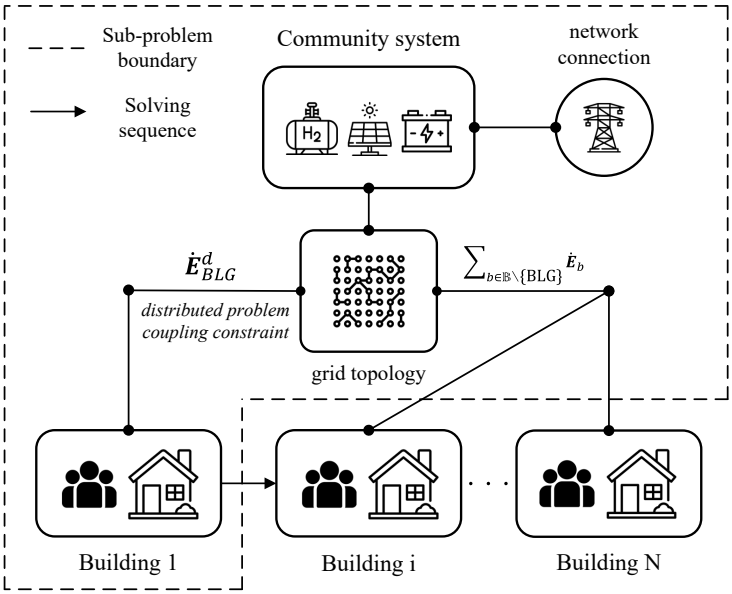


Figure 5.5: Sequential solving scheme of the distributed stochastic optimization problem.

The stochastic optimization problem is thus divided into sub-systems to form a distributed stochastic optimization problem. We consider individual building systems coupled with community-level utilities as sub-problems in order to provide the solver with available information from all spatial scales of the system. The grid topology energy balance, Eq. (5.46), here serves as an evident coupling constraint, and becomes

$$\dot{\boldsymbol{E}}^d_{MV,out} + \sum_{\mathrm{b}\in\mathbb{B}\setminus\{\mathrm{BLG}\}} \dot{\boldsymbol{E}}_{\mathrm{b},out} + \dot{\boldsymbol{E}}^d_{\mathrm{BLG},out} = \sum_{\mathrm{b}\in\mathbb{B}\setminus\{\mathrm{BLG}\}} \dot{\boldsymbol{E}}_{\mathrm{b},in} + \dot{\boldsymbol{E}}^d_{MV,in} + \dot{\boldsymbol{E}}^d_{\mathrm{BLG},in} \ , \tag{5.57}$$

where BLG is the considered building sub-system being optimized, and b indicates all other buildings belonging to the building set $\mathbb{B}$. Notice how the aggregated energy demands of other building systems $\sum_{\mathrm{b}\in\mathbb{B}\setminus\{\mathrm{BLG}\}} \dot{\boldsymbol{E}}_{\mathrm{b}}$ is now a parameter of the optimization problem, rather than a decision variable.

The distributed optimization setup is illustrated in Fig. 5.5. The problem is iteratively solved until variations of the global objective function are below a predefined threshold ε such that $\Delta O^{tot} \leq \varepsilon$ defines the stopping criterion of the distributed optimization.

Although the proposed distributed setup lacks a formal mathematical decomposition that would secure the convergence of the problem to the global optimal, we instead undertake a proof of concept, which compares a reduced problem of the proposed distributed stochastic

Table 5.3: Distributed problem proof of concept

Problem type	Parameters			Objective value
	buildings [#]	scenarios [#]	iterations [#]	[EUR]
centralized	5	10	1	15'188.173
distributed	5	10	19	15'188.173

optimization with its centralized counterpart. In particular, 5 out of the 41 available buildings are considered to reduce the problem size. The number of representative scenarios is kept to 10, and the number of iterations a priori defined for the distributed optimization to converge is set to 19. Table 5.3 summarizes the proof of concept result and system parameters. The proof of concept demonstrates that the distributed setup converges to the global optimal solution ensuing the first iteration as a result of the simple, individualistic optimal strategy identified. Typically, energy exchanges between the varying sub-systems of the distributed problem would iteratively converge to the global optimal within a 1% margin. We consider such optimal-close solutions satisfactory, and in fact valuable to the scientific and research community as these provide a simple and intuitive problem distribution arrangement supporting scalable strategic urban energy planning, thus facilitating the accessibility of our approach. Additionally, the structure of the distributed problem may be subsequently employed for the decentralized control of the energy community, by simply disregarding the investment-related variables.

5.4 Implementation

The energy community problem is implemented in Python using the PuLP package [270] as an interface to the Gurobi solver [271].

For the energy community system considered, we employ the 2NECO data set, thus anchoring our approach on data-driven techniques to induce realistic results. A total of 225 homes are originally treated, over a period of 3 years starting from January 1st 2019 to the 2nd of December 2022. Their associated building heat dynamics models are extracted from the open data set Grey-Brick Buildings [123] established from the same case study. We filter out models exhibiting nCPBES (normalized cumulated periodogram boundary excess sum) higher than 0.01 to retain models of good fit quality exclusively, resulting in 41 remaining buildings. These buildings are then assembled as a synthetic neighborhood, under a common atmospheric condition. It should here be noted that detaching building measurements from their original climate condition neglects the existing correlation between building lighting and heating demands and their ambient environment. However, due to the homogeneity of the Dutch geographical climate, we consider this approximation to be acceptable.

Weather data is assembled from publicly available Royal Netherlands Meteorological Institute

(KNMI) weather station measurements [142], employing a typical inland location in the center of the Netherlands. Economic data encompass forecasted day-ahead electricity prices coupled with residential natural gas prices. The former are collected and published by ENTSO-E (European Network of Transmission System Operators for Electricity) [272] and the latter from Eurostat [273] at granularities of 1 hour and 6 months respectively. Both prices are coupled with environmental taxes set by the Dutch government [274], and electricity day-ahead prices additionally include fixed distribution and transmission tariffs from corresponding time periods to approximate best the end-user total price. These are reported by the Netherlands authority for consumers and markets, in Annex 2 [275].

Load curves of identified representative scenarios are illustrated for weather, economic, and occupant behavior in Figures 5.6, 5.7, 5.8 respectively. These showcase sorted values of each time series allowing the inspection of their separate range and distribution. In the background of each figure is additionally plotted one unsorted scenario, showcasing eventual seasonal behaviors.

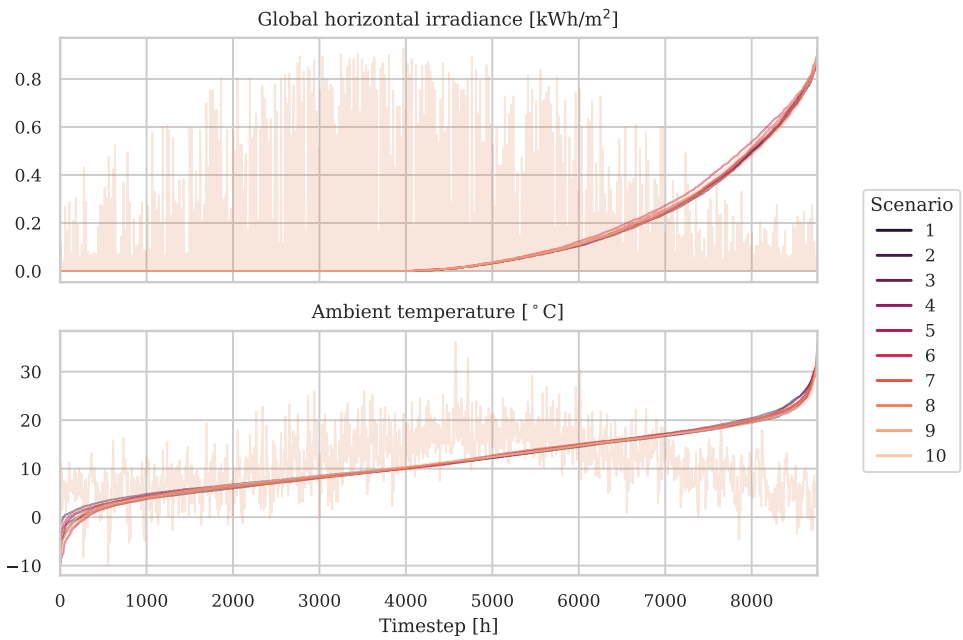


Figure 5.6: Load curves of weather uncertainty per representative scenarios

The techno-economic model parameters are gathered from the Danish Energy Agency (DEA) technology data catalogue [276]. Their referencing is summarized in Tab. 5.4.

5.5 Results and Discussion

The identified optimal energy community design is here presented along with its in situ building control operational strategy. Then we unveil which uncertainty factor impacts the

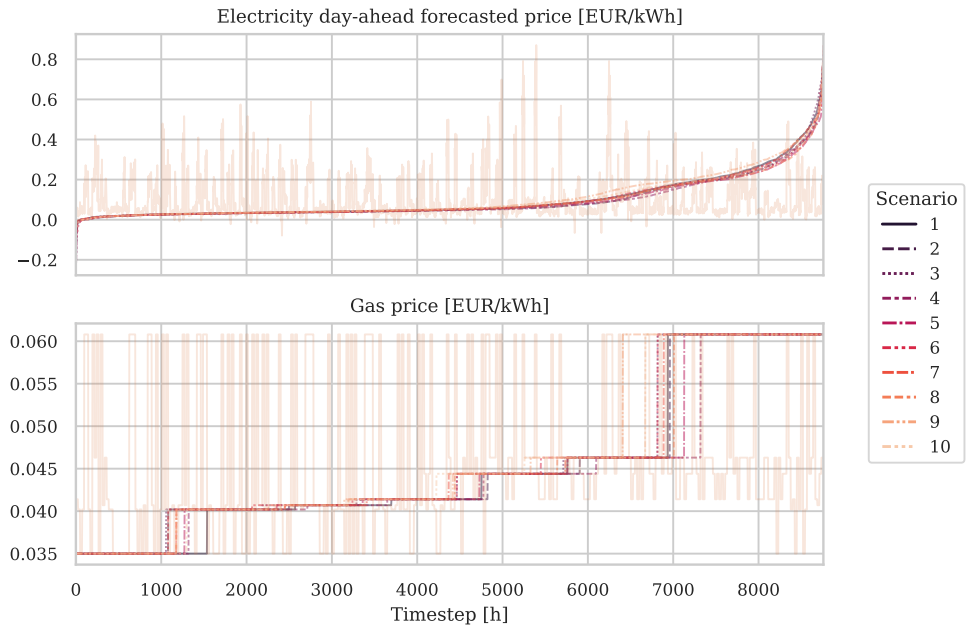


Figure 5.7: Load curves of economic uncertainty per representative scenarios

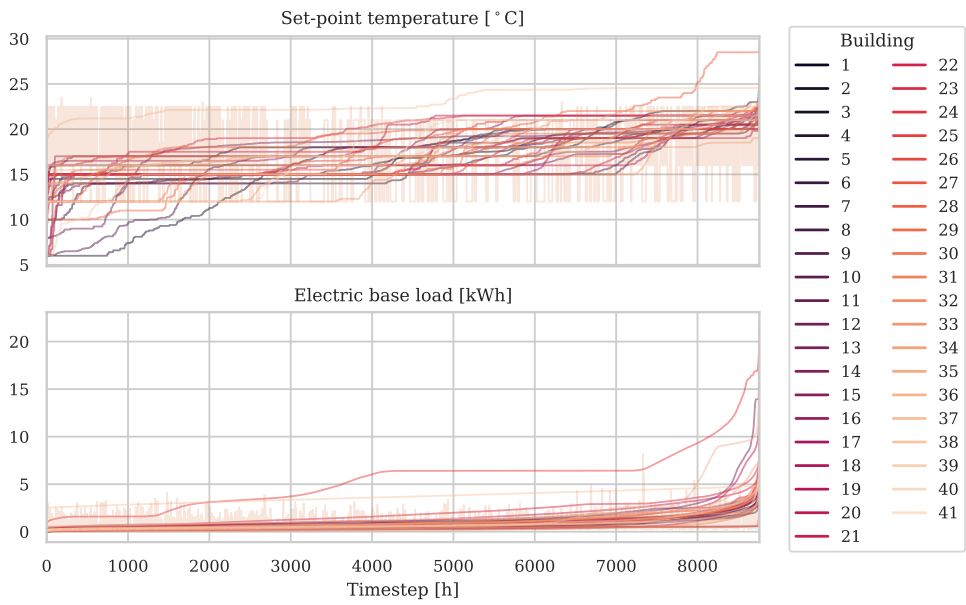


Figure 5.8: Load curves of occupant behavior uncertainty averaged over representative scenarios

Table 5.4: Techno-economical parameter DEA catalogue referencing. The reported year provides either historical or projected values of the utility parameters

Utility	Technology data catalogue	Index	Date
EL	Energy Carrier Generation and Conversion June 2017	86 Hydrogen production via alkaline electrolysis (AEC) for 1MW plant	2030
FC	power and heat production plants	12 Low temp PEM fuel cell - back pressure - hydrogen - small	2020
PV building	power and heat production plants	22 rooftop PV residential	2020
PV community	power and heat production plants	22 rooftop PV comm.&industrial	2020
BOL	heating installations	202 Gas boiler, ex single	2020
HP	heating installations	207 Heat pump, Air-to-water - apartment complex - existing building	2020
STC	heating installations	215 Solar heating system - single-family house - existing building	2020
HWT	Energy storage	142 Small-Scale Hot Water Tanks	2020
HYD	Energy storage	151a Pressurized hydrogen gas storage system (Compressor & Type I tanks 200bar)	2020
BAT building	Energy storage	181 Lithium-ion NMC battery (Utility-scale, Samsung SDI E3-R135)	2020
BAT community	Energy storage	182 NaS battery	2020

community design most through a local sensitivity analysis.

5.5.1 Resilient energy community design

The distributed stochastic optimization problem was found to converge immediately following its first iteration as no improvements to the master problem objective function were made in the following iterations. This direct convergence is a result of the identified optimal design of the energy community, which solely considers the necessary utilities to provide heating to the buildings, see Figure 5.9. The optimal design variables selected display boilers as the most energy and cost-efficient utility to supply space heating. Building #15 is the only system considering a heat pump to supplement its heating needs, due to the maximum capacity reached by the boiler. Another building, i.e., building #4, also reaches maximum boiler capacity, however, does not consider the additional input of a heat pump to provide its space heating needs.

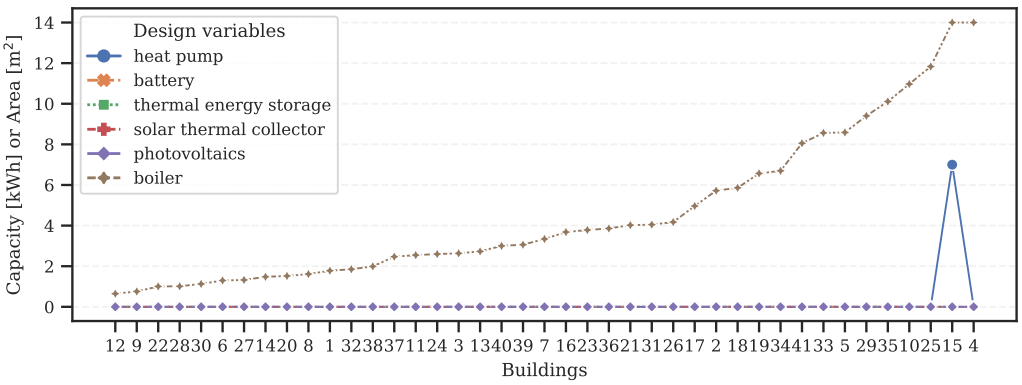


Figure 5.9: Optimal design of the energy community by taking stochasticity into account.

This optimal energy community design renders energy exchanges between the different building systems of the community disadvantageous thus resulting in a situation where cooperation between energy community members is not exploited. The efficiency and investment costs of boilers coupled with the lower energy prices of gas, make heat pumps an

unprofitable alternative for space heating and here thus only considered as a complementary utility. Higher gas prices coupled with carbon reduction incentives are consequently still needed for buildings to consider heat pumps as an interesting alternative to boilers and begin the decarbonization of the sector.

As gas prices are commonly fixed within fixed tranches of months, demand-side management control strategies in response to varying energy prices are seldom observed in the building in situ control strategies. Figure 5.10 illustrates the inside temperature control of building #1 in function of occupant-driven set-point preferences, ambient and economic conditions. The temperature set-point defined by the occupants can, however, be observed to be set at lower plateaus in periods of higher gas prices, see days 01/23 and 01/28 where it is set at a higher plateau of 17°C instead of the more common 19°C.

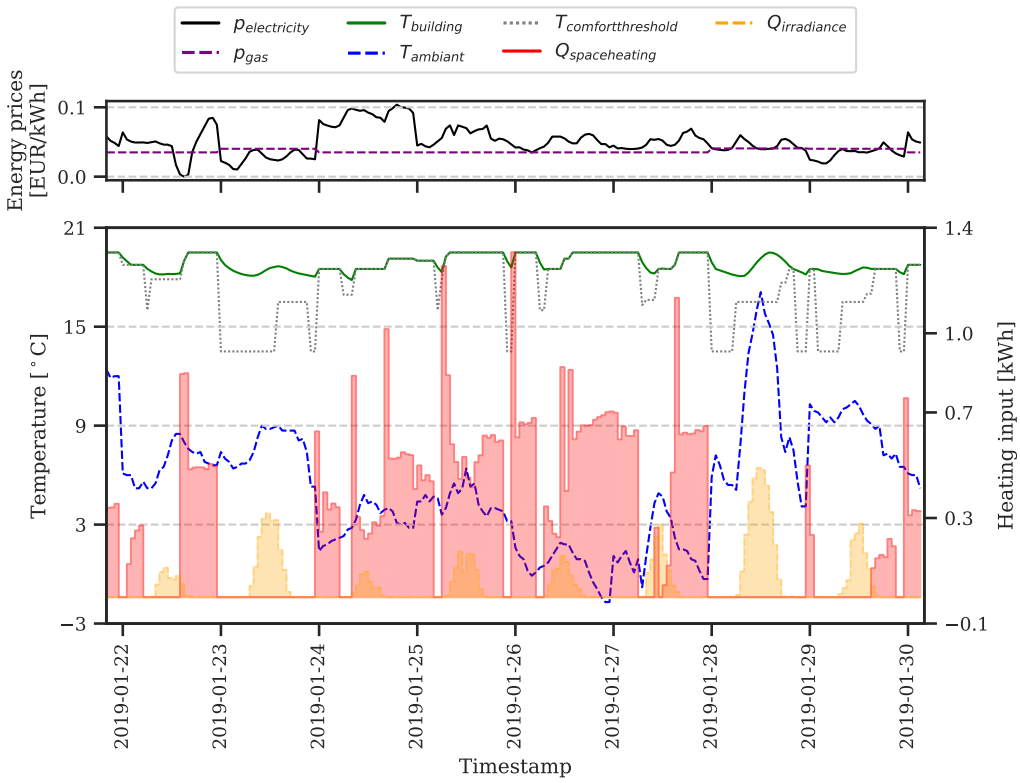


Figure 5.10: Building #1 inside temperature control.

5.5.2 Uncertainty factor impact assessment

To quantify the impact of occupant behavior on energy community planning within the context of other uncertainty factors we evaluate the optimal design of considered utilities within varying uncertain parameters. Figure 5.11 presents the spread and mean values of the considered design variable over the building stock, i.e., full circle and diamond points respectively, grouped by uncertainty parameter. Full lines represent the standard deviation of design values per uncertainty factor. The optimal design values accounting for all uncertain

parameters, i.e., presented in Fig. 5.9, are here highlighted as *optimal* to showcase the optimal design relative to the uncertainty affecting the energy system. The spread of values exposes occupant behavior as the factor with the largest impact on selected design variables, followed by climate and economic conditions. Although climate conditions often produce design variable spreads with higher central tendencies, i.e., engendering larger design variables on average, the dispersion of values is highest for occupant behavior. Buildings #4 and #10 are two examples where heat pumps are considered only when sensitive to particular uncertainty factors, namely, occupant behavior and climate conditions respectively. In both cases, the central tendency of boiler capacities is highest for climate-related uncertainties but their spread is highest for occupant-driven ones. The main difference being that identified boiler capacities of building #4 range from 1 to 14 kWh and require the additional heat pump investment in one setting while for building #10 occupant-driven uncertainties result in much lower boiler capacities than climate-related ones subsequently resulting in climate-driven heat pump investment in 4 scenarios. Building #15 on the other hand necessitates maximum design capacities both for boiler and heat pumps for all uncertainty-related factors resulting in its need for thermal storage investment in one occupant-driven setting. This is likely due to large set-point temperature shifts set by the occupant requiring additional heat inputs in specific time windows. This important finding demonstrates the significance of occupant behavior in strategic urban energy design and the value of bridging these two disconnected spatial scales.

Furthermore, optimal design variables identified by the stochastic problem formulation can often be found with values higher than its highest uncertainty analysis factor, e.g., see buildings #5, 7, 25, or 35. This is a result of the separation of correlated uncertainty factors in the sensitivity analysis. Indeed, in the local sensitivity analysis, evaluated uncertainty factors will be paired to nominal scenarios of other factors, e.g., occupant behavior scenarios ω_i^{occ} will be paired to ω_{nom}^{eco} and ω_{nom}^{clim}, whereas in the stochastic problem formulation, the combinations are different and each scenario i regroups all uncertainty factors. As nominal scenarios represent the most likely scenario per uncertainty factor, it is likely other scenarios might impact the design with more unlikely, and possibly extreme conditions, thus resulting in larger design variables. This highlights the importance of considering varying uncertainty parameters in the design phase of energy systems and the value brought by stochastic approaches, which provide robust solutions towards the more extreme conditions.

5.6 Summary

This chapter attempts to bridge two typically disconnected scales of the built environment for improved energy and carbon emission performances: occupants and the urban energy system. Strategic energy planning is undertaken by exploiting energy community concepts such as peer-to-peer cooperative energy exchanges and shared neighborhood-level infrastructure. Particularly, uncertainty factors affecting urban energy planning are embedded to the problem

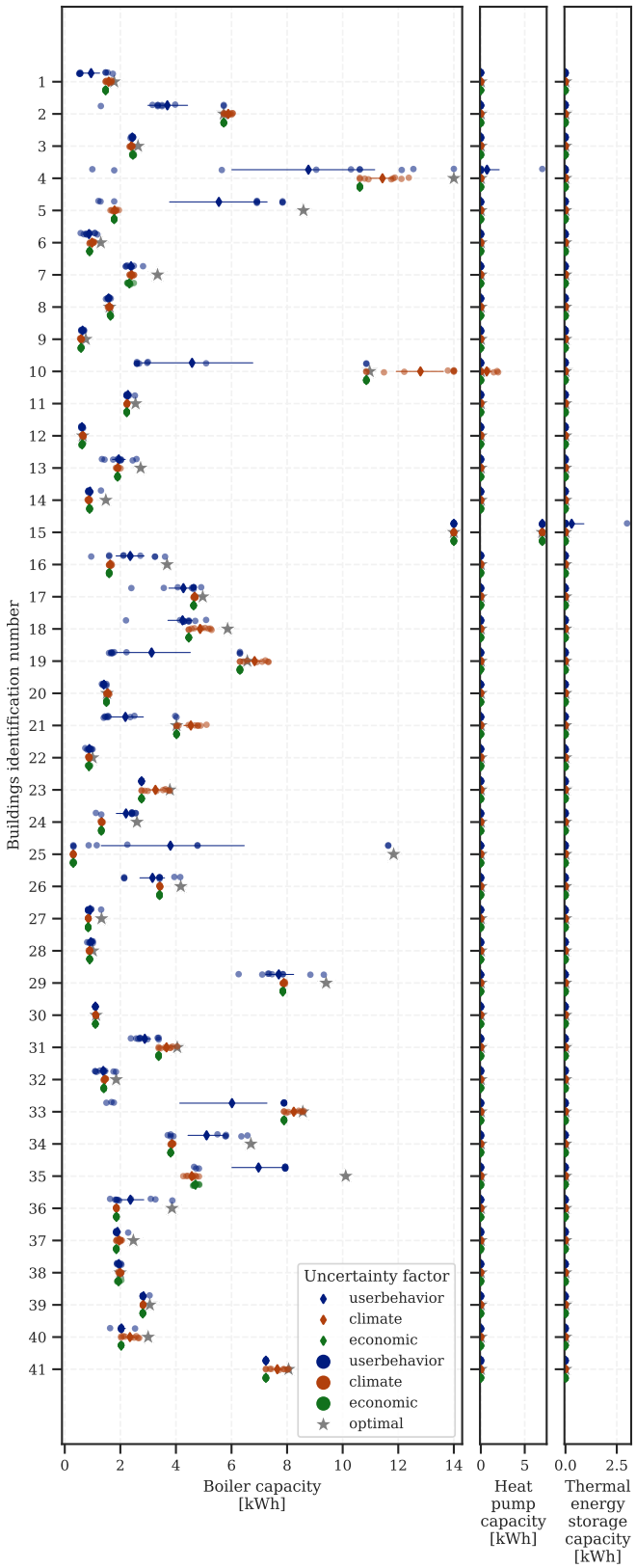


Figure 5.11: Uncertainty factor impact on energy community design.

and investigated by employing a stochastic problem formulation supplemented by a local sensitivity analysis. Computational tractability concerns are addressed, founded on an organic spatial problem distribution, which we validate by a proof of concept. The setup notably echoes that of decentralized energy management systems, thus implanting our approach in a real-world operational control setting, suitable for field deployment.

From historical measurements and accurate techno-economical parameter settings, a typical Dutch energy community composed of 41 residential buildings is designed. Results present a fast-converging distributed stochastic problem, where boilers are showcased as the winning utility provider for space heating. These expose current Dutch energy prices along with carbon emission taxes as not profitable enough for generalized heat pump adoption in typical residential buildings. It is postulated that increased electricity prices might also push energy communities to further adopt distributed energy renewables such as photovoltaics and solar thermal collectors. In such as setting, energy storage utilities will become compelling to align mismatches between renewable production and occupant-driven energy loads, as well as peer-to-peer energy exchanges.

Lastly, the impact of occupant behavior, encompassing set-point temperature and smart-meter base loads, on strategic energy planning is specifically investigated relative to other uncertainty factors, i.e., economic, electricity and gas prices, and weather, ambient temperature and solar irradiance, conditions. The analysis reveals occupants to be the leading factor affecting energy community design, thus confirming the relevance of our approach in connecting occupants to urban energy planning.

5.6.1 Limitations and future research

While our findings portray occupant behavior, i.e., building set-point temperature and electricity base loads, to be the leading uncertainty factor affecting the system design, it should be noted that representative scenarios were only sampled from historical measurements over the years 2019 to 2022. Employing older historical measurements or considerations with regard to the long-term evolution of weather and economic data might produce differing results. Thus, forecasts and uncertainty analysis related to these developments remain a goal for future research.

Additionally, varying community sizes and heterogeneity, i.e., number of buildings and representative occupant behaviors respectively, in the context of optimal stochastic urban energy planning offers an interesting analysis for urban planning decision-makers. Answering questions such as "*How large must a community be for shared utilities, such as seasonal storage, to become profitable?*" or "*How does occupant heterogeneity affect energy saving potentials?*" provide appealing research interrogations to guide subsequent studies.

Chapter 6

Conclusion

Overview

- Summary of the main results
- Recommendations and guidance for multi-scale(able) analytics for buildings
- Future perspectives

"Tell me and I forget, teach me and I may remember, involve me and I learn."
Benjamin Franklin

Data science and machine learning provide powerful solutions to support the decarbonization of the building sector. However, the diversity of buildings and disparity in their collected information hinders the broader adoption of data science practices in the building sector. This dissertation was undertaken with objectives related to the leveraging of generalizable insights gained from data science to boost energy-saving potentials for smart, connected buildings in a smart-grid context. The central goal of the effort is to develop automated approaches that may scale across the heterogeneous building stock and bridge disconnected scales of the built energy system. This led to the formulation of the main research question of this thesis in Chapter 1: *How can data science facilitate the scaling of approaches and bridge disconnected spatiotemporal scales of the built environment to deliver enhanced energy-saving strategies?* The answer is declined in four main contributions, covering methodological, descriptive, predictive, and prescriptive analytical approaches, thus providing a complete overview of multi-scale(able) data science practices for buildings. Overall, the thesis offers a broad data-driven analytical approach combining disconnected layers of the built environment together. This provides holistic insights and methods, which, together with open-sourced implementations and case studies, can effectively support, and *involve*, decision-makers in designing effective energy-saving strategies for buildings.

While methods, contributions, discussions, and results are detailed in their dedicated chapters, these concluding remarks summarize the main findings, provide recommendations and guidance for multi-scale(able) approaches for buildings, and envision future perspectives.

Main results summary

The main findings of this work are here outlined per analytical research contribution followed by their associated real-world innovation implications.

■ *Analytics*

Establishing necessary standards for multi-dimensional building data analytics incarnates the focus of Chap. 2. In this setting, we design a generic multi-dimensional data mining framework from data integration to end-use application. Leveraging data cube structures, encompassed dimensionalities can be formally mapped and examined to uncover insights from dimensional frames of interest. The method illustrates that **data cube mapping effectively breaks down high-dimensional analytical complexities**. Our method anchors popular analytical approaches in a three-dimensional data cube thus adequately linking building data dimensions to insight-driven analytics, i.e., bottom-up diagnostics {attribute, time}, top-down building stock benchmarking {site, time}, temporal drill-in analysis {site, attribute}. The method is applied to an automated building pattern identification case study, i.e., descriptive analytics, and connects two-dimensional lattice insights together exemplified by a three-dimensional data cube visualization serving practical cross-dimensional knowledge transfer.

In essence, the proposed analytical framework will support building data analyst professionals and researchers in the early design stages of their analysis. The (cube) mapping and reduction of encompassed dimensionality proposed by the method assist analysts with a visual framing of available dimensional state-space, and the generic data mining steps exposed secure a systematic approach to descriptive, predictive, and prescriptive analytics.

■ *Descriptive*

To identify building thermal dynamics in an interpretable and scalable manner, we investigate symbolic regression and lumped resistance capacity (RC) models in Chap. 3 and Annex A respectively as two data-driven modeling approaches, i.e., black- and grey-box, that can effectively scale across the heterogeneous building stock. An automated extension of RC model identification methods is first formulated concurrently with a novel model evaluation metric, namely the normalized cumulated periodogram boundary excess sum (nCPBES). Both approaches are then evaluated on a set of 225 occupied Dutch residential buildings in a non-intrusive way, which is seldom considered for building thermal identification due to the unmeasured heat gain disturbances caused by occupants. The analysis reveals that while symbolic regression allows the direct determination of analytical models in an automated manner, the interpretability of its coefficients cannot be linked back to the thermal characterization of buildings. On the other hand, the physical foundation of RC models grants an immediate thermal envelope performance overview, serving building stock performance

analysis and building-to-grid energy management applications. The method produced 144 good model fits (64%), 31 close (14%), and 50 poor quality model fits (22%). Our approach consequently showcases that **grey-box is a suitable and effective interpretable approach for scalable building heat dynamics identification across the building stock**. To support the establishment of standards and benchmarks in building models, obtained calibrated RC models were open-sourced as `Grey-brick buildings`. Guidelines for further work in the utilization of the open set for practical applications were additionally discussed.

Practically, the proposed open-source method allows industry professionals to identify the thermal performances of buildings at scale, with minimal implementation effort. These insights can support policy-makers with tangible building stock energy demand and carbon emission scenarios. Additionally, calibrated thermal models of buildings are at the foundation of most building service applications and can notably be exploited in prescriptive analytics such as strategic urban planning, see Chap. 5.

■ ***Predictive***

Predicting building loads deals with multiple spatiotemporal scales which have been unified in Chap. 4 to produce coherent forecasts supporting aligned decision-making across energy networks. First, a multi-dimensional extension of hierarchical structures is put forward. Then, a hierarchical machine learning regressor is designed, combining coherency requirement information of produced forecasts with tailored network architectures for targeted, data-efficient learning. This is a novelty compared to the literature in the field, where forecasts are typically produced disjointedly, or reconciliations undertaken a posteriori to the forecasting process. Results of the application over the BDG2 open set reveal improved accuracy and coherency performances for networks with fewer connections, in particular, tree partitionings with simple bottom-up connections. Additionally, the inclusion of coherency information in the machine learning loss function demonstrated improved accuracy and coherency performances for forecasts produced within reasonable accuracy limits. These discoveries demonstrate the **value brought by unifying disconnected scales together for building load prediction** and unveil a novel generation of forecasting regressors applicable in other fields.

Such hierarchical forecasts could rapidly change the way predictions are currently built and processed in industry and academia. For technical experts, this implies exploiting multi-level spatiotemporal characteristics of time series concurrently for improved forecast accuracies; for example, smooth, aggregated-level patterns, against more volatile, and information-rich, disaggregated elements. These multi-aggregation level data structures typically echo that of the data cube defined in Chap. 1. Thus, in the future, instead of developing isolated forecast models, one could imagine developing multi-level spatiotemporal ones as blocks, connected together by coherency constraints to further enhance the performance of a level-specific

prediction.

■ ***Prescriptive***

At last, the relationship between occupants and city energy infrastructures is investigated through the prism of uncertainty in Chap. 5. While the diversity of occupant behavior uncertainties might significantly affect strategic urban energy investment, these are seldom considered in urban planning problems due to their produced problem complexities. To examine how these commonly separate scales might correlate, an energy community planning problem is designed, incorporating resiliency and scalability targets. The impact of occupant behavior (temperature-set point and smart-meter loads) is particularly assessed in the context of other uncertainties affecting the system, i.e., climate (ambient temperature and solar irradiance) and economic (electricity and gas prices) conditions. A distributed stochastic problem formulation is outlined, following which representative scenarios and their associated probabilities are identified on a use case of 42 residential buildings. Results disclose a fast-converging distributed optimization problem, where boilers are showcased as the preferred heating utility, and distributed renewable energy and storage systems are identified as unprofitable for the community. A local sensitivity analysis particularly revealed high variabilities in selected design variables sensitive to occupant behavior. This discovery reveals that **occupants significantly impact strategic energy planning decisions**. This finding demonstrates the relevance and value of connecting occupants to cities for improved and more resilient urban energy planning strategies.

In light of this, urban planners need to exploit more detailed spatial information when designing optimization problems. Particularly, occupant-behavior uncertainty must be considered to produce resilient, more cost-effective, urban planning strategies. Capitalizing on distributed optimization technics, the subsequent computational burden endowed from large, granular problems can be practically cut down. Fully exploiting interconnected subsystems of the energy network is presently rendered possible thanks to such decentralized and coordinated approaches. These developments present signs of an energy revolution that is already underway, shifting uni-scale energy system planning and control schemes to multi-level and multi-vision ones, impacting energy network operators, consumers, markets, and policymakers, alike.

Recommendations and guidance

Founded on the obtained results and the time dedicated to the research work behind this thesis, some recommendations and guidance are provided to data analysts aiming at leveraging multi-scale building analytics. These are summarized in the following points:

- **Value against complexity** - Multi-dimensional analytics have repeatedly demonstrated value throughout this thesis, whether it produces expanded insights into system

performance, improved forecasts, or more resilient energy planning strategies. These, however, can come at the cost of greater complexities. While the mapping of encompassed dimensions leveraging data cube structures presented in Chap. 2 supports this deconstruction, it is important to first evaluate the complexity/value trade-off resulting from multi-dimensional approaches not to commit to potentially unnecessary labor-intensive tasks. In light of this, it is recommended to

- **Start simple**. Tackling scalability approaches in a sector endowed with high diversity can quickly result in complicated and hefty processes. Before undertaking multi-scale methods, it is advised to start simple and iteratively build up toward larger dimensions. As an example, the 2-dimensional lattice exploration of 3-dimensional data cubes in Chap. 2 demonstrated varying analytical pathways for the undertaking of 3-dimensional cuboid mining. Additionally, the insights gained from exploring these lattices first can be built onto in an expanded dimensional setting. This shows that multi-scale analytics calls for a priori uni-scale examination and knowledge.
- **Distribution for scalability and simplicity** - Building towards bridging disconnected scales often produces systems combining high levels of details with large spatiotemporal scopes. This inevitably engenders computational and performance complications due to the resulting problem size. The results in this thesis suggest that distributing such problems into smaller sub-systems, connected together in a coordinated fashion, offers a suitable solution tackling both simplification and scalability concerns.

Future perspectives

Future improvements of the presented work are envisioned in the context of activating buildings as energy management assets of smart-grid energy systems. Firstly, with the recent democratization of heat pumps providing residential building space heating demands, there is a growing need to anticipate, regulate, and optimize the ensuing additional load on the energy grid. This requires the identification *at scale* of building thermal characteristics, which uncovers two main challenges: (*i*) the need for scalable, non-intrusive, interpretable, and robust methods for building heat dynamics identification and (*ii*) generalizing findings to data-poor buildings, where little to no in-situ measurements, e.g., smart thermostats, are available.

While this research engages in providing solutions to the former concern, other emerging technologies should concurrently be explored. In particular, the emergence of physics-informed machine learning models has unveiled low computational and data requirements while providing physical interpretability of its parameter from eigenvalues [200]. One could further imagine developing constrained symbolic regression approaches, where analytical formulations are recursively built from an ensemble of predetermined domain-informed

equations. For example, typical heat exchange model extensions could be assembled serving automated, unsupervised, physics-informed, building heat dynamics identification from symbolic regression as an extension to Annex A. These techniques, working at the frontiers of black- and grey-box paradigms, typify the future of data-driven modeling approaches and provide possible pathways toward more efficient and robust approaches for building thermal characteristics identification.

Then, exploring data-driven technics to generalize findings across case studies is essential to unlock the full energy-saving potential of the building stock. This can be undertaken using popular machine learning technics such as classification or transfer learning, yet have seldom been explored for building thermal characteristics identification applications due to the lack of available benchmarks. Indeed, to generalize such findings to data-poor buildings a sufficient pool of thermal characteristics first need to be assembled. The published *Grey-brick buildings* open data set provides an initial step towards this goal. Yet, to establish a complete benchmark, numerous additional buildings portraying varying types and climate zone locations need to be investigated. Such a set would support the adoption of approaches to generalize the thermal characterization of entire building stocks. Overall, the broader adoption of data science best practices requires established benchmarks to back significant findings. This process is still emerging in the building sector, yet has begun gaining momentum over the last decade with open-source competitions such as the ASHRAE Great Energy Predictor [277] or the CityLearn challenge [278]. Fostering open-source data science practices supported by open sets has consequently been a major endeavor of this thesis.

Lastly, urban energy planning approaches have, to this date, relied on data-driven optimization techniques, popularly employed in control settings, e.g., model predictive control (MPC). The emergence of purely data-driven approaches applied to large multi-agent control problems has also gained traction over the last decade from research and industry both. Reinforcement learning stands as a particularly attractive technique for the intelligent control of buildings; where varying uncertainties such as occupant behaviors and unreported physical changes to the construction can be autonomously learned and handled by the agent. Its capacity to capture the long-term impact of short-term decisions by approximating the Bellman Value function possesses an appealing potential that could be applicable to the long-term design of urban energy systems comparable to optimization approaches employing mixed-integer linear programming (MILP) such as ours (Chap. 5). These investigations are on the brink of becoming a reality as frameworks bridging branch and bound methods to reinforcement learning are starting to appear in the literature [279]. This domain presents an exciting area of research for urban energy planners boosted by the development of fully data-driven, scalable, and autonomous planning agents.

Concluding remarks

This Ph.D. dissertation has demonstrated through a set of developed methods, models, and analytical frameworks, that connecting occupants, buildings and smart-energy networks together uncovers significant added value; these contribute to building stock characterization (descriptive), load forecasting (predictive), and resilient energy strategies (prescriptive analytics). The work particularly focuses on providing scalable and interpretable approaches founded on data science principles that can be replicated or expanded to other case studies. To this end, all developed implementations were open-sourced (https://github.com/JulienLeprince) and relied as much as possible on the exploitation of open data sets to secure research reproducibility and cultivate best data science practices in the building sector supported by benchmarks. Such practices further encourage research dissemination serving industry, education, and policymakers towards the decarbonization of the building sector.

Appendix A

Uncovering physical models from symbolic regressions for scalable building heat dynamics identification

Overview

This appendix is complementary to Chapter 3 and details:

- Symbolic regression as a black-box approach to produce interpretable, analytical expressions for heat dynamics identification
- Case study: 241 residential buildings, 2NECO
- Produced linear analytical expressions are analyzed and compared to typical grey-box RC models

This appendix has been published as Leprince et al. [144].

A.1 Preface

Modeling building thermal dynamics is an important challenge in characterizing performance towards various objectives. With applications in building retrofitting [13], demand-side management [16], energy forecasting [15], and model predictive control [17], it has been at the center of many research publications within the past decades. Despite its momentum, the approach is still faced with the fundamental challenge of scaling across the heterogeneous building stock. Thermal dynamics modeling fits in one of three well-established categories: physics-based methods (white-box), purely data-driven (black-box) and hybrid approaches (grey-box) [125]. Physics and knowledge-based methods (white-box) are known to be time-consuming and difficult to scale up. With many parameters to fix and human expertise required, they are better fitted to detailed and isolated case-study building models. Grey-box models, on the other hand, work as a hybrid approach bridging the gap between physical and statistical modeling. By exploiting physical knowledge in their models, grey-box models profit from interpretability, while exploiting the particularities of case-study data information for

parameter fitting which makes them good at generalization [131]. Finally, data driven models (black-box), encompass machine learning algorithms and statistical regressions, commonly fitted from input and output time-series data of the system. They are notoriously powerful at generalizing yet struggle to produce interpretable models. And while efforts in the domain have allowed the *opening of the box* through feature importance metrics, e.g., SHAP (SHapley Additive exPlanations) values [280], physical interpretation of captured models remains an existing gap in the field.

Symbolic regression was recently put to light as a powerful black-box approach for extracting analytical equations out of data. However, when dealing with high-dimensionalities, the exponential explosion of combinations make it poor at scaling. Established data-driven building heat dynamics model identification processes (grey-box) typically require as little as 4-dimensional data measurements; inside temperature, heat signal input, and outside conditions with ambient temperature and solar global irradiance [132]. Although these approaches benefit from physical knowledge included within the developed models, it becomes interesting to explore how symbolic regressions could uncover new forms of building heat dynamics.

A.1.1 Opening the (black) box

This study consequently proposes to group, categorize and analyze the analytical expression outputs of symbolic regression for building heat dynamics model identification. From a case study of 241 monitored Dutch residential buildings and exploiting the paved path provided by grey-box approaches, we propose to uncover the relationships driving inside temperature states through interpretable black-box-produced models. Prediction accuracies of the identified models are benchmarked against a commonly employed black-box regressor within the building sector, i.e., XGBoost, as well as against a naive predictor to confirm its relative performance.

A.2 Symbolic Regression

Symbolic regression is a machine learning algorithm based on genetic programming which uses a simple tree-like representation structures to build an analytical expression from given input data and mathematical operators [281]. By iteratively mutating, performing crossovers or replications of the tree branches, multiple analytical expressions are explored to determine the best fit to the given data. The procedure produces increasingly complex analytical expressions from the given input features to predict the target output. The equation with the largest fractional drop in error metric is selected as the best model [282].

Ultimately, symbolic regression derives explicit physical relations between components of a system in an automated way. Additionally by building the symbolic expression from a tree structure, increasing orders of complexities are explored as the model develops. Thus, the

algorithm only incrementally incorporates features into the model, which allows inputs with no significant impact on the target output not to be considered upfront.

A.3 Implementation

This study considers the 2NECO data set, leveraging inside temperature measurements, heat signal control inputs, ambient temperature, and solar irradiance. Electric and gas-meter data are also available at resolutions of 10 seconds and 1 hour respectively.

We consider minimum measurement periods of two months and limit the maximum times-series length to 10'000 points, which corresponds to a period of approximately 3,5 months, which is the recommended, amply sufficient, maximum input data length for the Symbolic Regression. The measurement period ranges from February 1st to the end of May 2021, which comprises the end of winter season as well as a notably cold start of spring season at the beginning of April. Electric and gas-meter data are re-sampled by average to 15 minutes intervals to match the smart-thermostat information. Available data are then filtered to obtain the most recent continuous measurement period for each building. Cumulative missing values larger than two hours are imputed and smaller gaps are filled via moving average using an eight hours window size.

We employ the open source python library *PySR* developed by Miles Cranmer et al. [282] for the symbolic learning algorithm of this study. Mathematical operators considered encompass multiplication, addition, division, cosine and sine functions. The number of iterations, or generations the regression runs for, is set to 10 and no weights are assigned to input data or operators not to influence the knowledge discovery process. Prediction accuracy of the regressor is benchmarked against a naive predictor as well as a gradient boosting regressor, i.e., XGBoost from the scikit-learn package [205], a commonly employed black-box regressor within the building sector. The naive regressor simply predicts the step-ahead inside temperature of the considered building to be the same as the last, providing a classic reference value to compare a regressor's performance to. The XGBoost regressor is trained from a classic 20-fold *TimeSeriesSplit* function of the *sklearn* python package using the same input data as fed to the symbolic regressor.

A.4 Results

The performance of the symbolic regression is evaluated using the distribution of the Mean Squared Error (MSE) of the fitted models. Figure A.1 presents the boxplot distribution of the symbolic regressor (*SR*) compared to a naive (*naive*) and gradient boosted (*XGB*) regressor. The obtained symbolic expressions present lower MSE central tendencies and spread compared to both benchmarks. Interestingly enough, it can also be noted that the simple naive regressor seems to produce lower MSEs overall than the gradient boosted

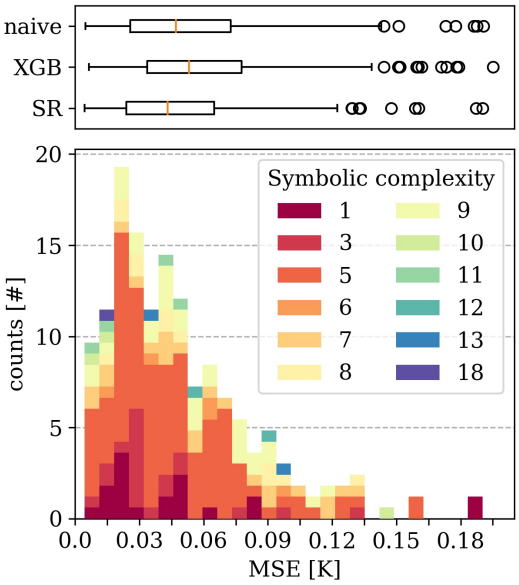


Figure A.1: Mean Squared Error (MSE) distribution of the symbolic regression (SR) output analytical functions with their respective complexities compared to naive (naive) and XGBoost (XGB) regressors.

method. This result might be due to the lack of parameter tuning for this method. A number of MSE outliers for both *naive* and *XGB* regressors are not represented given the upper limit of MSE axis that reaches as high as 0.85 Kelvins for XGB. The distribution of obtained symbolic expression complexities from the *SR* exposes a predominant presence of complexities of order five, which corresponds to a typical affine expression $a \cdot T_i + b$, where coefficients a and b, variable T_i and operators $\cdot$ and $+$ each add an order 1 of complexity to the overall expression.

The symbolic expressions derived from the black-box *SR* are post-processed to uncover 50 unique analytical expressions. These equations are then grouped into similar analytical expression families which are presented along with their group size and intra-family available attributes, i.e., building characteristics in Figure A.2 and symbolic coefficient values in Figure A.3. While examining Figure A.2, it can be noticed that all but one family constitute linear polynomial expressions of order one and two. The exception here being the $T_i + \Phi_h \cdot sin(f \cdot T_i)$ expression comprising either a *sine* or *cosine* function, where f is a coefficient and variables T_i and Φ_h represent inside temperature and space heating input signal respectively. The two preeminent analytical families, i.e., $T_i \cdot a + \Phi_h \cdot b + c$ and the simpler $T_i \cdot a + c$, where a, b and c represent affine coefficients, evoke simple first order regressions of the inside temperature considering, or not, the space heating input signal. The building characteristics distribution per analytical group seem to suggest smaller homes to be more frequent in $T_i \cdot a + V_g \cdot d + c$, where V_g represents gas-meter measurements, along with a larger proportion of family sizes of 3. The two largest family groups appear to be mainly composed of family sizes of 2 and 4,

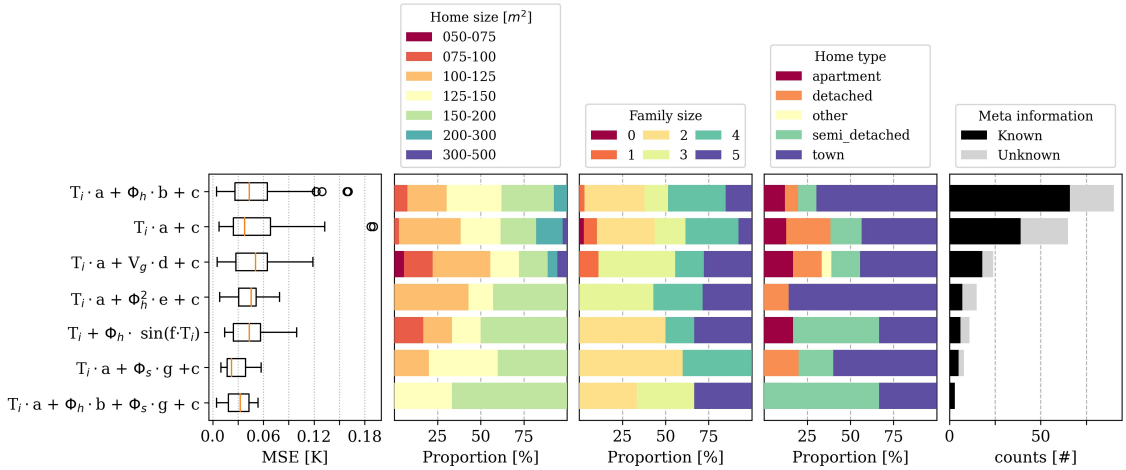


Figure A.2: Identified symbolic expression groups and their building characteristics meta-data distribution

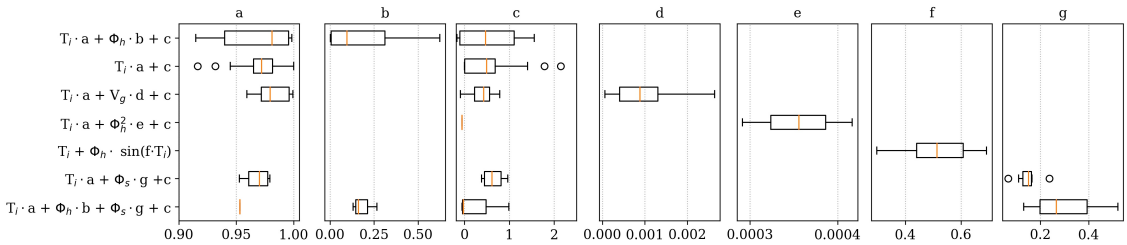


Figure A.3: Identified symbolic expressions and their coefficients boxplot distribution

while home type distributions cover a crushing majority of town types. Given the number of buildings grouped per analytical family along with their available meta-data, only the top 3 groups present results that can be considered significant.

Figure A.3 allows us to dive into the fitted coefficient distributions per family group. Coefficients of value 0 or 1 have here been removed not to bias the appreciation of distributions. Inside temperature-related coefficients a are commonly centered around 0.97, with a negatively skewed distribution, while coefficients c are spread between values of -0.1 and 2 with positively skewed distributions centered around 0.6. Overall, larger-sized groups tend to show larger coefficient distributions.

A.5 Discussion

We here discuss how the discovered findings might bring value to the building sector by (i) creating paradigm links between discovered black-box models and established grey-box ones and (ii) uncovering physical knowledge from identified models.

A.5.1 Link to grey-box paradigm

Thanks to the formulation of symbolic expressions describing dynamical systems from measurements, it naturally follows that parallels can be drawn to well-established grey-box models.

While grey-box necessitates the definition of multiple models for appropriate model selection, symbolic regression inherently iteratively builds the model, thus making it far better at generalization and automation. Commonly, multiple state point estimates are included in grey-box models, corresponding to up to fifth-order models. The complexities captured by these models echo quite naturally with the different thermal inertiae interacting in buildings. While these models increase in complexity, their inherent interpretability related to building physics allows a direct evaluation of estimated parameters. In opposition, analytical functions discovered by *SR*, while being interpretable, necessitate physical analysis by domain experts which can scale poorly given the variety of identified functions. It follows that natural links between the thermal properties of a building and identified *SR* coefficients can be drawn, building on the knowledge of grey-box models. Quite concretely, a typical grey-box building model can be represented by lumped resistance-capacity models. The below differential equation represents a 1st order model,

$$dT_i = \frac{1}{R_{ia}C_i}(T_a - T_i)dt + \frac{1}{C_i}\eta_h\Phi_h dt + \frac{1}{C_i}A_w\Phi_s dt + \epsilon$$

where the state variables T and Φ represent temperature and heat flux, estimated parameters R, C, η and A serve as heat resistance, heat capacity, appliance efficiency, and area respectively and the subscripts i, a, h, s and w relate to inside, ambient, heat, solar and window components respectively. ϵ encapsulates the measurement error, model approximations, and non-recognized or modeled phenomenons [136].

Linking identified *SR* functions to this formulation uncovers physical components such as building heat capacity C_i, space heating appliance efficiency η_h and thermal gains ϵ here linked to coefficients a, b and c respectively. Coefficient g may also be associated with solar window area gains A_w. The below equation explicitly links the above grey-box model to identified polynomial relationships.

$$C_i \cdot T_i(t+1) = C_i \cdot T_i(t) + \eta_h \cdot \Phi_h(t) + A_w \cdot \Phi_s(t) + \epsilon(t)$$
$$C_i \cdot T_i(t+1) = a \cdot T_i(t) + b \cdot \Phi_h(t) + g \cdot \Phi_s(t) + c$$

Confirming these physical links would however require knowledge of ground truth - an interesting area of research for future studies that could help uncover direct links between measurements and building characteristics.

A.5.2 Model discovery

Leveraging knowledge discovery can be used as a powerful tool to build new models, enhancing white- or grey-box model identification approaches. While all discovered polynomial models possess a linear simplicity that makes their interpretation accessible, some of the more complex or non-intuitive uncovered models might just be the starting point of a new generation of models. The cosine and sine function of the inside temperature identified in $T_i + E_h \cdot sin(f \cdot T_i)$ may here correspond to particular cyclical control strategies of the thermostat. Additionally, while no significant meta-data factor seems to separate variations of the identified polynomial functions, inside temperature patterns emerging from thermostat control and occupant behavior heat gains might unveil these model structures. This requires deeper inspections outside the scope of this work.

A.6 Summary

With this work, we bring to light an automated model identification of building heat dynamics approach from data. With 241 monitored buildings, fifty unique models were uncovered and grouped into seven main families of symbolic expressions, six of which are polynomials. These results support established differential models developed with grey-box approaches while favoring simplified symbolic complexities. It brings important perspectives to model identification in practice, e.g., for forecasting and control applications. Discovered models and coefficients may be exploited in a variety of building service applications including automated and scalable model identification and calibration for building Model Predictive Control (MPC). Building performance analytics may also leverage such findings for building characteristics benchmarking or thermostat control strategy characterization.

References

[1] P. Taylor, *Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025*, Sep. 2022. [Online]. Available: `https://www.statista.com/statistics/871513/worldwide-data-created/`.

[2] J. Han, M. Kamber, and J. Pei, "Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)," Tech. Rep., 2011.

[3] E. Commission and D.-G. for Communication, *The European Green Deal*. Publications Office, 2020. DOI: `doi/10.2775/19001`.

[4] P. Agreement, "Paris agreement," in *Report of the Conference of the Parties to the United Nations Framework Convention on Climate Change (21st Session, 2015: Paris). Retrived December*, HeinOnline, vol. 4, 2015, p. 2017.

[5] S. Perry, J. Klemeš, and I. Bulatov, "Integrating waste and renewable energy to reduce the carbon footprint of locally integrated energy sectors," *Energy*, vol. 33, no. 10, pp. 1489–1497, 2008.

[6] K. Kotilainen, "Energy prosumers' role in the sustainable energy system," in *Affordable and Clean Energy*, Springer, 2019, pp. 1–14.

[7] I. E. A. D. of Sustainable Energy Policy, *Transition to sustainable buildings: strategies and opportunities to 2050*. Organization for Economic, 2013.

[8] OECD, *Decarbonising Buildings in Cities and Regions*. 2022, p. 95. DOI: `https://doi.org/https://doi.org/10.1787/a48ce566-en`. [Online]. Available: `https://www.oecd-ilibrary.org/content/publication/a48ce566-en`.

[9] T. M. Lawrence, M.-C. Boudreau, L. Helsen, *et al.*, "Ten questions concerning integrating smart buildings into the smart grid," *Building and Environment*, vol. 108, pp. 273–283, 2016.

[10] A. H. Buckman, M. Mayfield, and S. B. Beck, "What is a smart building?" *Smart and Sustainable Built Environment*, vol. 3, no. 2, pp. 92–109, 2014.

[11] P. Domingues, P. Carreira, R. Vieira, and W. Kastner, "Building automation systems: Concepts and technology review," *Computer Standards & Interfaces*, vol. 45, pp. 1–12, 2016.

[12] A. Capozzoli, F. Lauro, and I. Khan, "Fault detection analysis using data mining techniques for a cluster of smart office buildings," *Expert Systems with Applications*, vol. 42, no. 9, pp. 4324–4338, 2015.

[13] M. Rabani, H. B. Madessa, and N. Nord, "A state-of-art review of retrofit interventions in buildings towards nearly zero energy level," *Energy Procedia*, vol. 134, pp. 317–

326, 2017, Sustainability in Energy and Buildings 2017: Proceedings of the Ninth KES International Conference, Chania, Greece, 5-7 July 2017, ISSN: 1876-6102. DOI: https://doi.org/10.1016/j.egypro.2017.09.534. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1876610217346635.

[14] Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali, and A. Amira, "Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives," *Applied Energy*, vol. 287, p. 116 601, 2021.

[15] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Transactions on Smart Grid*, vol. 10, pp. 3125–3148, 2019.

[16] E. Guelpa and V. Verda, "Demand response and other demand side management techniques for district heating: A review," *Energy*, vol. 219, p. 119 440, 2021, ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2020.119440. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544220325470.

[17] J. Drgoňa, J. Arroyo, I. Cupeiro Figueroa, *et al.*, "All you need to know about model predictive control for buildings," *Annual Reviews in Control*, vol. 50, pp. 190–232, 2020, ISSN: 1367-5788. DOI: https://doi.org/10.1016/j.arcontrol.2020.09.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1367578820300584.

[18] M. W. Ahmad, M. Mourshed, D. Mundow, M. Sisinni, and Y. Rezgui, "Building energy metering and environmental monitoring–a state-of-the-art review and directions for future research," *Energy and Buildings*, vol. 120, pp. 85–102, 2016.

[19] B. Balaji, A. Bhattacharya, G. Fierro, *et al.*, "Brick: Towards a unified metadata schema for buildings," in *Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments*, 2016, pp. 41–50.

[20] C. Miller, A. Kathirgamanathan, B. Picchetti, *et al.*, "The building data genome project 2, energy meter data from the ashrae great energy predictor iii competition," *Scientific data*, vol. 7, no. 1, pp. 1–13, 2020.

[21] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big data*, vol. 1, no. 1, pp. 51–59, 2013.

[22] T. M. Mitchell and T. M. Mitchell, *Machine learning*. McGraw-hill New York, 1997, vol. 1.

[23] M. Aghbashlo, W. Peng, M. Tabatabaei, *et al.*, "Machine learning technology in biodiesel research: A review," *Progress in Energy and Combustion Science*, vol. 85, p. 100 904, 2021.

[24] C. Fan, F. Xiao, and C. Yan, "A framework for knowledge discovery in massive building automation data and its application in building diagnostics," *Automation in Construction*, vol. 50, pp. 81–90, 2015.

[25] C. Fan, F. Xiao, H. Madsen, and D. Wang, "Temporal knowledge discovery in big bas data for building energy management," *Energy and Buildings*, vol. 109, pp. 75–89, 2015.

[26] P. C. Sen, M. Hajra, and M. Ghosh, "Supervised classification algorithms in machine learning: A survey and review," in *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*, Springer, 2020, pp. 99–111.

[27] B. Yildiz, J. I. Bilbao, and A. B. Sproul, "A review and analysis of regression and machine learning models on commercial building electricity load forecasting," *Renewable and Sustainable Energy Reviews*, vol. 73, pp. 1104–1122, 2017.

[28] A. Basu, "Five pillars of prescriptive analytics success," *Analytics magazine*, vol. 8, p. 12, 2013.

[29] R. Soltanpoor and T. Sellis, "Prescriptive analytics for big data," in *Databases Theory and Applications: 27th Australasian Database Conference, ADC 2016, Sydney, NSW, September 28-29, 2016, Proceedings 27*, Springer, 2016, pp. 245–256.

[30] C. Apte, "The role of machine learning in business optimization," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, Citeseer, 2010, pp. 1–2.

[31] D. Bertsimas and N. Kallus, "From predictive to prescriptive analytics," *Management Science*, vol. 66, no. 3, pp. 1025–1044, 2020.

[32] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[33] M. Molina-Solana, M. Ros, M. D. Ruiz, J. Gómez-Romero, and M. J. Martín-Bautista, "Data science for building energy management: A review," *Renewable and Sustainable Energy Reviews*, vol. 70, pp. 598–609, 2017.

[34] M. Manic, D. Wijayasekara, K. Amarasinghe, and J. J. Rodriguez-Andina, "Building energy management systems: The age of intelligent and adaptive buildings," *IEEE Industrial Electronics Magazine*, vol. 10, no. 1, pp. 25–39, 2016.

[35] Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, and J. Li, "A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis," *Energy and Built Environment*, vol. 1, no. 2, pp. 149–164, 2020.

[36] C. Fan, D. Yan, F. Xiao, A. Li, J. An, and X. Kang, "Advanced data analytics for enhancing building performances: From data-driven to big data-driven approaches," in *Building Simulation*, Springer, vol. 14, 2021, pp. 3–24.

[37] H. B. Gunay, W. Shen, and G. Newsham, "Data analytics to improve building performance: A critical review," *Automation in Construction*, vol. 97, pp. 96–109, 2019.

[38] C. Fan, F. Xiao, Z. Li, and J. Wang, "Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review," *Energy and Buildings*, vol. 159, pp. 296–308, 2018.

[39] S. Singh and A. Yassine, "Big data mining of energy time series for behavioral analytics and energy consumption forecasting," *Energies*, vol. 11, no. 2, p. 452, 2018.

[40] F. Xiao and C. Fan, "Data mining in building automation system for improving building operational performance," *Energy and buildings*, vol. 75, pp. 109–118, 2014.

[41] E. Keogh and S. Kasetty, "On the need for time series data mining benchmarks: A survey and empirical demonstration," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 102–111.

[42] B. Haskell, "Pecan street, inc.: Smart grid, smart meter, big data,"

[43] J. R. Schofield, R. Carmichael, S. Tindemans, M. Bilton, M. Woolf, G. Strbac, *et al.*, "Low carbon london project: Data from the dynamic time-of-use electricity pricing trial, 2013," *uK Data Service, SN*, vol. 7857, no. 2015, pp. 7857–7851, 2015.

[44] IEA, *Renovation of near 20% of existing building stock to zero-carbon-ready by 2030 is ambitious but necessary*, 2022. [Online]. Available: `https://www.iea.org/reports/renovation-of-near-20-of-existing-building-stock-to-zero-carbon-ready-by-2030-is-ambitious-but-necessary`.

[45] B. Boardman, S. Darby, G. Killip, *et al.*, "40% house," 2005.

[46] Z. Tian, X. Zhang, S. Wei, S. Du, and X. Shi, "A review of data-driven building performance analysis and design on big on-site building performance data," *Journal of Building Engineering*, vol. 41, p. 102 706, 2021.

[47] P. Torcellini, S. Pless, B. Griffith, and R. Judkoff, "Evaluation of the energy performance and design process of the thermal test facility at the national renewable energy laboratory," National Renewable Energy Lab., Golden, CO (US), Tech. Rep., 2005.

[48] B. A. Thornton, M. I. Rosenberg, E. E. Richman, *et al.*, "Achieving the 30% goal: Energy and cost savings analysis of ashrae standard 90.1-2010," May 2011. DOI: `10.2172/1015277`. [Online]. Available: `https://www.osti.gov/biblio/1015277`.

[49] S. Hu, D. Yan, E. Azar, and F. Guo, "A systematic review of occupant behavior in building energy policy," *Building and Environment*, vol. 175, p. 106 807, 2020.

[50] P. Hoes, J. L. Hensen, M. G. Loomans, B. de Vries, and D. Bourgeois, "User behavior in whole building simulation," *Energy and buildings*, vol. 41, no. 3, pp. 295–302, 2009.

[51] D. Yan, W. O'Brien, T. Hong, *et al.*, "Occupant behavior modeling for building performance simulation: Current state and future challenges," *Energy and buildings*, vol. 107, pp. 264–278, 2015.

[52] A. Paone and J.-P. Bacher, "The impact of building occupant behavior on energy efficiency and methods to influence it: A review of the state of the art," *Energies*, vol. 11, no. 4, p. 953, 2018.

[53] Y. Zhang, X. Bai, F. P. Mills, and J. C. Pezzey, "Rethinking the role of occupant behavior in building energy performance: A review," *Energy and Buildings*, vol. 172, pp. 279–294, 2018.

[54] F. E. Aliabadi, K. Agbossou, S. Kelouwani, N. Henao, and S. S. Hosseini, "Coordination of smart home energy management systems in neighborhood areas: A systematic review," *IEEE Access*, vol. 9, pp. 36 417–36 443, 2021.

[55] S. Cajot, M. Peter, J.-M. Bahu, F. Guignet, A. Koch, and F. Maréchal, "Obstacles in energy planning at the urban scale," *Sustainable cities and society*, vol. 30, pp. 223–236, 2017.

[56] O. V. H. Van Cutsem, "On smart-buildings and their integration into the smart-grid," EPFL, Tech. Rep., 2020.

[57] I. A. T. Hashem, V. Chang, N. B. Anuar, *et al.*, "The role of big data in smart city," *International Journal of information management*, vol. 36, no. 5, pp. 748–758, 2016.

[58] J. Leprince, C. Miller, and W. Zeiler, "Data mining cubes for buildings, a generic framework for multidimensional analytics of building performance data," *Energy and Buildings*, vol. 248, p. 111 195, 2021.

[59] F. M. Orr, "2015 quadrennial technology review," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2015.

[60] R. Zacks, *The technology review ten*, Jan. 2001. [Online]. Available: `https://wp.technologyreview.com/wp-content/uploads/2021/02/MIT-Technology-Review-2001-TR10.pdf`.

[61] Clarivate, *Web of science core collection basic search.* 2020. [Online]. Available: `https://apps.%20webofknowledge.com/WOS_GeneralSearch_%20input.do?product=WOS&search_mode=GeneralSearch&SID=F5lKgrWCQYzjDto4GIO&preferencesSaved=`.

[62] M. Oded and R. Lior, "Data mining and knowledge discovery handbook," *Edited by Rokach Lior Maimon Oded. Second Edi. Israel: Springer*, 2010.

[63] M. M. Abdelrahman, S. Zhan, C. Miller, and A. Chong, "Data science for building energy efficiency: A comprehensive text-mining driven review of scientific literature," *Energy and Buildings*, vol. 242, p. 110 885, 2021, ISSN: 0378-7788. DOI: `https://doi.org/10.1016/j.enbuild.2021.110885`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0378778821001699`.

[64] F. Dalene, "Technology and information management for low-carbon building," *Journal of Renewable and Sustainable Energy*, vol. 4, no. 4, p. 041 402, 2012.

[65] M. Bourdeau, X. qiang Zhai, E. Nefzaoui, X. Guo, and P. Chatellier, "Modeling and forecasting building energy consumption: A review of data-driven techniques," *Sustainable Cities and Society*, vol. 48, p. 101 533, 2019.

[66] I. Ghalehkhondabi, E. Ardjmand, G. R. Weckman, and W. A. Young, "An overview of energy demand forecasting methods published in 2005–2015," *Energy Systems*, vol. 8, pp. 411–447, 2017.

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

[68] K. Amasyali and N. M. El-Gohary, "A review of data-driven building energy consumption prediction studies," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1192–1205, 2018.

[69] C. Fan, Y. Ding, and Y. Liao, "Analysis of hourly cooling load prediction accuracy with data-mining approaches on different training time scales," *Sustainable Cities and Society*, vol. 51, p. 101 717, 2019.

[70] A. Moradzadeh, A. Mansour-Saatloo, B. Mohammadi-Ivatloo, and A. Anvari-Moghaddam, "Performance evaluation of two machine learning techniques in heating and cooling loads forecasting of residential buildings," *Applied Sciences*, vol. 10, no. 11, p. 3829, 2020.

[71] P. K. Sharma, T. De, and S. Saha, "Iot based indoor environment data modelling and prediction," in *2018 10th International Conference on Communication Systems & Networks (COMSNETS)*, IEEE, 2018, pp. 537–539.

[72] C. Xu, H. Chen, J. Wang, Y. Guo, and Y. Yuan, "Improving prediction performance for indoor temperature in public buildings based on a novel deep learning method," *Building and Environment*, vol. 148, pp. 128–135, 2019.

[73] C. Miller and F. Meggers, "Mining electrical meter data to predict principal building use, performance class, and operations strategy for hundreds of non-residential buildings," *Energy and Buildings*, vol. 156, pp. 360–373, 2017.

[74] D. Chakraborty and H. Elzarka, "Advanced machine learning techniques for building performance simulation: A comparative analysis," *Journal of Building Performance Simulation*, vol. 12, no. 2, pp. 193–207, 2019.

[75] T. Ahmad, H. Chen, and J. Shair, "Water source heat pump energy demand prognosticate using disparate data-mining based approaches," *Energy*, vol. 152, pp. 788–803, 2018.

[76] C. Miller, Z. Nagy, and A. Schlueter, "Automated daily pattern filtering of measured building performance data," *Automation in Construction*, vol. 49, pp. 1–17, 2015.

[77] W. Kim and S. Katipamula, "A review of fault detection and diagnostics methods for building systems," *Science and Technology for the Built Environment*, vol. 24, no. 1, pp. 3–21, 2018.

[78] Z. Shi and W. O'Brien, "Development and implementation of automated fault detection and diagnostics for building systems: A review," *Automation in Construction*, vol. 104, pp. 215–229, 2019.

[79] Y. Zhao, T. Li, X. Zhang, and C. Zhang, "Artificial intelligence-based fault detection and diagnosis methods for building energy systems: Advantages, challenges and the future," *Renewable and Sustainable Energy Reviews*, vol. 109, pp. 85–101, 2019.

[80] J. Gray, S. Chaudhuri, A. Bosworth, *et al.*, "Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals," *Data mining and knowledge discovery*, vol. 1, pp. 29–53, 1997.

[81] R. Ramakrishnan and B.-C. Chen, "Exploratory mining in cube space," *Data Mining and Knowledge Discovery*, vol. 15, no. 1, pp. 29–54, 2007.

[82] Z. Yu, B. C. Fung, and F. Haghighat, "Extracting knowledge from building-related data—a data mining framework," in *Building Simulation*, Springer, vol. 6, 2013, pp. 207–222.

[83] C. Fan, F. Xiao, and S. Wang, "Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques," *Applied Energy*, vol. 127, pp. 1–10, 2014.

[84] I. 8.-1. 2019, *Date and time—representations for information interchange—part 1: Basic rules*, 2019.

[85] M. Kottek, J. Grieser, C. Beck, B. Rudolf, and F. Rubel, "World map of the köppen-geiger climate classification updated," 2006.

[86] A. Wagner, W. O'Brien, and B. Dong, "Exploring occupant behavior in buildings," *Wagner, A., O'Brien, W., Dong, B., Eds*, 2018.

[87] P. Jayathissa, M. Quintana, M. Abdelrahman, and C. Miller, "Humans-as-a-sensor for buildings—intensive longitudinal indoor comfort models," *Buildings*, vol. 10, no. 10, p. 174, 2020.

[88] A. Mahdavi and M. Taheri, "An ontology for building monitoring," *Journal of Building Performance Simulation*, vol. 10, no. 5-6, pp. 499–508, 2017.

[89] V. Charpenay, S. Käbisch, D. Anicic, and H. Kosch, "An ontology design pattern for iot device tagging systems," in *2015 5th International Conference on the Internet of Things (IOT)*, IEEE, 2015, pp. 138–145.

[90] J. Yang, C. Ning, C. Deb, *et al.*, "K-shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement," *Energy and Buildings*, vol. 146, pp. 27–37, 2017.

[91] Y. Wang, Q. Chen, C. Kang, M. Zhang, K. Wang, and Y. Zhao, "Load profiling and its application to demand response: A review," *Tsinghua Science and Technology*, vol. 20, no. 2, pp. 117–129, 2015.

[92] H. Wang, M. J. Bah, and M. Hammad, "Progress in outlier detection techniques: A survey," *Ieee Access*, vol. 7, pp. 107 964–108 000, 2019.

[93] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial intelligence review*, vol. 22, pp. 85–126, 2004.

[94] Y. Qin, S. Zhang, X. Zhu, J. Zhang, and C. Zhang, "Pop algorithm: Kernel-based imputation to treat missing values in knowledge discovery from databases," *Expert systems with applications*, vol. 36, no. 2, pp. 2794–2804, 2009.

[95] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen, "Methods for imputation of missing values in air quality data sets," *Atmospheric environment*, vol. 38, no. 18, pp. 2895–2907, 2004.

[96] G. Hawthorne, G. Hawthorne, and P. Elliott, "Imputing cross-sectional missing data: Comparison of common techniques," *Australian and New Zealand Journal of Psychiatry*, vol. 39, no. 7, pp. 583–590, 2005.

[97] A. Plaia and A. Bondi, "Single imputation method of missing values in environmental pollution data sets," *Atmospheric Environment*, vol. 40, no. 38, pp. 7316–7330, 2006.

[98] M. Di Zio, U. Guarnera, and O. Luzi, "Imputation through finite gaussian mixture models," *Computational Statistics & Data Analysis*, vol. 51, no. 11, pp. 5305–5316, 2007.

[99] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793.

[100] M. N. Ramli, A. Yahaya, N. Ramli, N. Yusof, and M. Abdullah, "Roles of imputation methods for filling the missing values: A review," *Advances in Environmental Biology*, vol. 7, no. 12 S2, pp. 3861–3870, 2013.

[101] S. Morgenthaler, "Exploratory data analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 1, pp. 33–44, 2009.

[102] M. Abzalov and M. Abzalov, "Exploratory data analysis," *Applied Mining Geology*, pp. 207–219, 2016.

[103] B. Yildiz, J. I. Bilbao, J. Dore, and A. B. Sproul, "Recent advances in the analysis of residential electricity consumption and applications of smart meter data," *Applied Energy*, vol. 208, pp. 402–427, 2017.

[104] P. Bobko and R. Karren, "The perception of pearson product moment correlations from bivariate scatterplots," *Personnel Psychology*, vol. 32, no. 2, pp. 313–325, 1979.

[105] A. Davison and S. Sardy, "The partial scatterplot matrix," *Journal of Computational and Graphical Statistics*, vol. 9, no. 4, pp. 750–758, 2000.

[106] C. Zhang, L. Cao, and A. Romagnoli, "On the feature engineering of building energy data mining," *Sustainable cities and society*, vol. 39, pp. 508–518, 2018.

[107] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[108] C. Fan, Y. Sun, F. Xiao, *et al.*, "Statistical investigations of transfer learning-based methodology for short-term building energy predictions," *Applied Energy*, vol. 262, p. 114 499, 2020, ISSN: 0306-2619. DOI: `https://doi.org/10.1016/j.apenergy.2020.114499`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0306261920300118`.

[109] S. Singh and A. Yassine, "Big data mining of energy time series for behavioral analytics and energy consumption forecasting," *Energies*, vol. 11, no. 2, 2018, ISSN: 1996-1073. DOI: `10.3390/en11020452`. [Online]. Available: `https://www.mdpi.com/1996-1073/11/2/452`.

[110] S. Visalakshi and V. Radha, "A literature review of feature selection techniques and applications: Review of feature selection in data mining," in *2014 IEEE International*

Conference on Computational Intelligence and Computing Research, 2014, pp. 1–6. DOI: 10.1109/ICCIC.2014.7238499.

[111] A. N. Akansu and R. A. Haddad, *Multiresolution signal decomposition: transforms, subbands, and wavelets*. Academic press, 2001.

[112] Y. Wei, X. Zhang, Y. Shi, *et al.*, "A review of data-driven approaches for prediction and classification of building energy consumption," *Renewable and Sustainable Energy Reviews*, vol. 82, pp. 1027–1047, 2018.

[113] S. B. Kotsiantis, I. Zaharakis, P. Pintelas, *et al.*, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007.

[114] D. Francisci and M. Collard, "Multi-criteria evaluation of interesting dependencies according to a data mining approach," in *The 2003 Congress on Evolutionary Computation, 2003. CEC'03.*, IEEE, vol. 3, 2003, pp. 1568–1574.

[115] M. Aruldoss, T. M. Lakshmi, and V. P. Venkatesan, "A survey on multi criteria decision making methods and its applications," *American Journal of Information Systems*, vol. 1, no. 1, pp. 31–43, 2013.

[116] I. P. Panapakidis and G. C. Christoforidis, "Optimal selection of clustering algorithm via multi-criteria decision analysis (mcda) for load profiling applications," *Applied Sciences*, vol. 8, no. 2, p. 237, 2018.

[117] J. P. Gouveia and J. Seixas, "Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys," *Energy and buildings*, vol. 116, pp. 666–676, 2016.

[118] J. Leprince and W. Zeiler, "A robust building energy pattern mining method and its application to demand forecasting," in *2020 International Conference on Smart Energy Systems and Technologies (SEST)*, IEEE, 2020, pp. 1–6.

[119] R. K. Pearson, "Outliers in process modeling and identification," *IEEE Transactions on control systems technology*, vol. 10, no. 1, pp. 55–63, 2002.

[120] D. Q. Goldin and P. C. Kanellakis, "On similarity queries for time-series data: Constraint specification and implementation," in *Principles and Practice of Constraint Programming—CP'95: First International Conference, CP'95 Cassis, France, September 19–22, 1995 Proceedings 1*, Springer, 1995, pp. 137–153.

[121] K. Techapichetvanich and A. Datta, "Interactive visualization for olap," in *Computational Science and Its Applications–ICCSA 2005: International Conference, Singapore, May 9-12, 2005, Proceedings, Part III 5*, Springer Berlin Heidelberg, 2005, pp. 206–214.

[122] J. Leprince, H. Madsen, C. Miller, *et al.*, "Fifty shades of grey: Automated stochastic model identification of building heat dynamics," *Energy and Buildings*, vol. 266, p. 112 095, 2022.

[123] J. Leprince, C. Miller, H. Madsen, K. Basu, R. van der Vlist, and W. Zeiler, "Greybrick buildings, an open data set of calibrated rc models of dutch residential building

heat dynamics," in *Proceedings of the Twentieth ACM Conference on Embedded Networked Sensor Systems*, 2022, pp. 1067–1071.

[124] Z. Wang and T. Hong, "Reinforcement learning for building controls: The opportunities and challenges," *Applied Energy*, vol. 269, p. 115 036, 2020.

[125] M. Bourdeau, X. qiang Zhai, E. Nefzaoui, X. Guo, and P. Chatellier, "Modeling and forecasting building energy consumption: A review of data-driven techniques," *Sustainable Cities and Society*, vol. 48, p. 101 533, 2019, ISSN: 2210-6707. DOI: https://doi.org/10.1016/j.scs.2019.101533. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210670718323862.

[126] D. B. Crawley, L. K. Lawrie, C. O. Pedersen, and F. C. Winkelmann, "Energy plus: Energy simulation program," *ASHRAE journal*, vol. 42, no. 4, pp. 49–56, 2000.

[127] S. E. Mattsson, H. Elmqvist, and M. Otter, "Physical system modeling with modelica," *Control Engineering Practice*, vol. 6, no. 4, pp. 501–510, 1998.

[128] X. Yu, L. Georges, and L. Imsland, "Data Pre-Processing and Optimization Techniques for Stochastic and Deterministic Low-Order Grey-box Models of Residential Buildings," *Energy and Buildings*, vol. 236, p. 110 775, Feb. 2021, ISSN: 03787788. DOI: 10.1016/j.enbuild.2021.110775.

[129] L. Berti-Equille, "Measuring and modelling data quality for quality-awareness in data mining," in *Quality measures in data mining*, Springer, 2007, pp. 101–126.

[130] J. A. Crabb, N. Murdoch, and J. M. Penman, "A simplified thermal response model," *Building Services Engineering Research and Technology*, vol. 8, no. 1, pp. 13–19, 1987. DOI: 10.1177/014362448700800104. eprint: https://doi.org/10.1177/014362448700800104. [Online]. Available: https://doi.org/10.1177/014362448700800104.

[131] A. Afram and F. Janabi-Sharifi, "Review of modeling methods for HVAC systems," *Applied Thermal Engineering*, vol. 67, no. 1-2, pp. 507–519, 2014, ISSN: 13594311. DOI: 10.1016/j.applthermaleng.2014.03.055.

[132] P. Bacher and H. Madsen, "Identifying suitable models for the heat dynamics of buildings," *Energy and Buildings*, vol. 43, no. 7, pp. 1511–1522, Jul. 2011, ISSN: 03787788. DOI: 10.1016/j.enbuild.2011.02.005.

[133] M. M. Hossain, T. Zhang, and O. Ardakanian, "Identifying grey-box thermal models with Bayesian neural networks," *Energy and Buildings*, vol. 238, p. 110 836, 2021, ISSN: 03787788. DOI: 10.1016/j.enbuild.2021.110836. arXiv: 2009.05889. [Online]. Available: https://doi.org/10.1016/j.enbuild.2021.110836.

[134] A.-H. Deconinck and S. Roels, "Is stochastic grey-box modelling suited for physical properties estimation of building components from on-site measurements?" *Journal of Building Physics*, vol. 40, no. 5, pp. 444–471, 2017. DOI: 10.1177/1744259116688384. eprint: https://doi.org/10.1177/1744259116688384. [Online]. Available: https://doi.org/10.1177/1744259116688384.

[135] C. Rasmussen, P. Bacher, D. Calì, H. A. Nielsen, and H. Madsen, "Method for scalable and automatised thermal building performance documentation and screening," *Energies*, vol. 13, no. 15, 2020, ISSN: 1996-1073. DOI: 10.3390/en13153866. [Online]. Available: https://www.mdpi.com/1996-1073/13/15/3866.

[136] H. Madsen and J. Holst, "Estimation of continuous-time models for the heat dynamics of a building," *Energy and Buildings*, vol. 22, pp. 67–79, 1995.

[137] N. R. Kristensen, H. Madsen, and S. B. Jørgensen, "Parameter estimation in stochastic grey-box models," *Automatica*, vol. 40, no. 2, pp. 225–237, 2004, ISSN: 0005-1098. DOI: https://doi.org/10.1016/j.automatica.2003.10.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S000510980300298X.

[138] J. Neyman and E. S. Pearson, "Ix. on the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, no. 694-706, pp. 289–337, 1933.

[139] H. Madsen, *Time series analysis*. Chapman and Hall/CRC, 2007.

[140] P. J. Brockwell and R. A. Davis, *Time series: theory and methods*. Springer Series in Statistics, 1991.

[141] D. Chen and H. W. Chen, "Using the köppen classification to quantify climate variation and change: An example for 1901-2010," *Environmental Development*, vol. 6, pp. 69–79, 2013, ISSN: 2211-4645. DOI: https://doi.org/10.1016/j.envdev.2013.03.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2211464513000328.

[142] KNMI, *Hourly values of weather stations*, Aug. 2021. [Online]. Available: https://www.daggegevens.knmi.nl/klimatologie/uurgegevens.

[143] R. Juhl, J. K. Møller, and H. Madsen, *Ctsmr - continuous time stochastic modeling in r*, 2016. arXiv: 1606.00242 [stat.CO].

[144] J. Leprince, C. Miller, M. Frei, H. Madsen, and W. Zeiler, "Fifty shades of black: Uncovering physical models from symbolic regressions for scalable building heat dynamics identification," in *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, ser. BuildSys '21, Coimbra, Portugal: Association for Computing Machinery, Nov. 2021, pp. 345–348.

[145] "Thermal performance of buildings — Transmission and ventilation heat transfer coefficients — Calculation method," International Organization for Standardization, Geneva, CH, Standard, Jun. 2017.

[146] M. Li, D. Allinson, and K. Lomas, "Estimation of building heat transfer coefficients from in-use data: Impacts of unmonitored energy flows," *International Journal of Building Pathology and Adaptation*, 2019.

[147] "Energy performance of buildings — Energy needs for heating and cooling, internal temperatures and sensible and latent heat loads," International Organization for Standardization, Geneva, CH, Standard, Jun. 2017.

[148] G. Feng, S. Sha, and X. Xu, "Analysis of the building envelope influence to building energy consumption in the cold regions," *Procedia Engineering*, vol. 146, pp. 244–250, 2016.

[149] K. Lylykangas, "Shape factor as an indicator of heating energy demand," in *Proceedings of the 15th International Wood Construction Conference (IHF)*, 2009.

[150] Y.-Y. Zhang, Z.-Z. Hu, J.-R. Lin, and J.-P. Zhang, "Linking data model and formula to automate kpi calculation for building performance benchmarking," *Energy Reports*, vol. 7, pp. 1326–1337, 2021, ISSN: 2352-4847. DOI: `https://doi.org/10.1016/j.egyr.2021.02.044`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S2352484721001426`.

[151] J. Palmer Real, C. Rasmussen, R. Li, *et al.*, "Characterisation of thermal energy dynamics of residential buildings with scarce data," *Energy and Buildings*, vol. 230, p. 110 530, 2021, ISSN: 0378-7788. DOI: `https://doi.org/10.1016/j.enbuild.2020.110530`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0378778820317503`.

[152] Y. Li, J. O'Donnell, R. García-Castro, and S. Vega-Sánchez, "Identifying stakeholders and key performance indicators for district and building energy performance analysis," *Energy and Buildings*, vol. 155, pp. 1–15, 2017, ISSN: 03787788. DOI: `10.1016/j.enbuild.2017.09.003`. [Online]. Available: `https://doi.org/10.1016/j.enbuild.2017.09.003`.

[153] C. Miller and F. Meggers, "Mining electrical meter data to predict principal building use, performance class, and operations strategy for hundreds of non-residential buildings," *Energy Build.*, vol. 156, pp. 360–373, Dec. 2017.

[154] C. F. Reinhart and C. Cerezo Davila, "Urban building energy modeling – a review of a nascent field," *Build. Environ.*, vol. 97, pp. 196–202, Feb. 2016.

[155] J. Sokol, C. Cerezo Davila, and C. F. Reinhart, "Validation of a bayesian-based method for defining residential archetypes in urban building energy models," *Energy Build.*, vol. 134, pp. 11–24, Jan. 2017.

[156] J. Roth, A. Martin, C. Miller, and R. K. Jain, "SynCity: Using open data to create a synthetic city of hourly building energy estimates by integrating data-driven and physics-based methods," *Appl. Energy*, vol. 280, p. 115 981, Dec. 2020.

[157] T. Kerdphol, K. Fuji, Y. Mitani, M. Watanabe, and Y. Qudaih, "Optimization of a battery energy storage system using particle swarm optimization for stand-alone microgrids," *Int. J. Electr. Power Energy Syst.*, vol. 81, pp. 32–39, Oct. 2016.

[158] B. Howard, L. Parshall, J. Thompson, S. Hammer, J. Dickinson, and V. Modi, "Spatial distribution of urban building energy consumption by end use," *Energy Build.*, vol. 45, pp. 141–151, Feb. 2012.

[159] R. K. Jain, K. M. Smith, P. J. Culligan, and J. E. Taylor, "Forecasting energy consumption of multi-family residential buildings using support vector regression: In-

vestigating the impact of temporal and spatial monitoring granularity on performance accuracy," *Appl. Energy*, vol. 123, pp. 168–178, Jun. 2014.

[160] M. C. Silva, I. M. Horta, V. Leal, and V. Oliveira, "A spatially-explicit methodological framework based on neural networks to assess the effect of urban form on energy demand," *Appl. Energy*, vol. 202, pp. 386–398, Sep. 2017.

[161] J. Leprince, H. Madsen, J. K. Møller, and W. Zeiler, *Hierarchical learning, forecasting coherent spatio-temporal individual and aggregated building loads*, 2023. DOI: `10.48550/ARXIV.2301.12967`. [Online]. Available: `https://arxiv.org/abs/2301.12967`.

[162] J. Leprince, W. Khan, H. Madsen, J. K. Møller, and W. Zeiler, *Structural hierarchical learning for energy networks*, 2023. DOI: `10.48550/ARXIV.2302.03978`. [Online]. Available: `https://arxiv.org/abs/2302.03978`.

[163] P. Nystrup, E. Lindstrøm, P. Pinson, and H. Madsen, "Temporal hierarchies with autocorrelation for load forecasting," *European Journal of Operational Research*, vol. 280, no. 3, pp. 876–888, 2020.

[164] N. Kourentzes and G. Athanasopoulos, "Cross-temporal coherent forecasts for australian tourism," *Annals of Tourism Research*, vol. 75, pp. 393–409, 2019.

[165] G. Athanasopoulos, R. A. Ahmed, and R. J. Hyndman, "Hierarchical forecasts for australian domestic tourism," *International Journal of Forecasting*, vol. 25, no. 1, pp. 146–166, 2009.

[166] M. Kremer, E. Siemsen, and D. J. Thomas, "The sum and its parts: Judgmental hierarchical forecasting," *Management Science*, vol. 62, no. 9, pp. 2745–2764, 2016.

[167] E. Spiliotis, M. Abolghasemi, R. J. Hyndman, F. Petropoulos, and V. Assimakopoulos, "Hierarchical forecast reconciliation with machine learning," *Applied Soft Computing*, vol. 112, p. 107 756, 2021.

[168] S. B. Taieb, J. W. Taylor, and R. J. Hyndman, "Hierarchical probabilistic forecasting of electricity demand with smart meter data," *Journal of the American Statistical Association*, vol. 116, no. 533, pp. 27–43, 2021.

[169] T. Ahmad, H. Zhang, and B. Yan, "A review on renewable energy and electricity requirement forecasting models for smart grid and buildings," *Sustainable Cities and Society*, vol. 55, p. 102 052, 2020.

[170] J. Peng, A. Kimmig, Z. Niu, J. Wang, X. Liu, and J. Ovtcharova, "A flexible potential-flow model based high resolution spatiotemporal energy demand forecasting framework," *Applied Energy*, vol. 299, p. 117 321, 2021.

[171] E. Spiliotis, F. Petropoulos, N. Kourentzes, and V. Assimakopoulos, "Cross-temporal aggregation: Improving the forecast accuracy of hierarchical electricity consumption," *Applied Energy*, vol. 261, p. 114 339, 2020.

[172] R. Stone, D. G. Champernowne, and J. E. Meade, "The precision of national income estimates," *The Review of Economic Studies*, vol. 9, no. 2, pp. 111–125, 1942.

[173] M. Weale, "The reconciliation of values, volumes and prices in the national accounts," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 151, no. 1, pp. 211–221, 1988.

[174] R. J. Hyndman, R. A. Ahmed, G. Athanasopoulos, and H. L. Shang, "Optimal combination forecasts for hierarchical time series," *Computational statistics & data analysis*, vol. 55, no. 9, pp. 2579–2589, 2011.

[175] Y. Zhang and J. Dong, "Least squares-based optimal reconciliation method for hierarchical forecasts of wind power generation," *IEEE Transactions on Power Systems*, pp. 1–1, 2018. DOI: 10.1109/TPWRS.2018.2868175.

[176] G. Athanasopoulos, R. J. Hyndman, N. Kourentzes, and F. Petropoulos, "Forecasting with temporal hierarchies," *European Journal of Operational Research*, vol. 262, no. 1, pp. 60–74, 2017, ISSN: 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2017.02.046. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377221717301911.

[177] E. Spiliotis, F. Petropoulos, and V. Assimakopoulos, "Improving the forecasting performance of temporal hierarchies," *Plos one*, vol. 14, no. 10, e0223422, 2019.

[178] H. G. Bergsteinsson, J. K. Møller, P. Nystrup, Ó. P. Pálsson, D. Guericke, and H. Madsen, "Heat load forecasting using adaptive temporal hierarchies," *Applied Energy*, vol. 292, p. 116 872, 2021.

[179] P. Nystrup, E. Lindström, J. K. Møller, and H. Madsen, "Dimensionality reduction in forecasting with temporal hierarchies," *International Journal of Forecasting*, vol. 37, no. 3, pp. 1127–1146, 2021.

[180] D. Yang, H. Quan, V. R. Disfani, and C. D. Rodríguez-Gallegos, "Reconciling solar forecasts: Temporal hierarchy," *Solar Energy*, vol. 158, pp. 332–346, 2017.

[181] S. L. Wickramasuriya, G. Athanasopoulos, and R. J. Hyndman, "Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization," *Journal of the American Statistical Association*, vol. 114, no. 526, pp. 804–819, 2019.

[182] T. van Erven and J. Cugliari, "Game-theoretically optimal reconciliation of contemporaneous hierarchical time series forecasts," in *Modeling and Stochastic Learning for Forecasting in High Dimensions*, A. Antoniadis, J.-M. Poggi, and X. Brossat, Eds., Cham: Springer International Publishing, 2015, pp. 297–317, ISBN: 978-3-319-18732-7.

[183] A. Timmermann, "Chapter 4 forecast combinations," in ser. Handbook of Economic Forecasting, G. Elliott, C. Granger, and A. Timmermann, Eds., vol. 1, Elsevier, 2006, pp. 135–196. DOI: https://doi.org/10.1016/S1574-0706(05)01004-9. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574070605010049.

[184] J. B. Edwards and G. H. Orcutt, "Should aggregation prior to estimation be the rule?" *The Review of Economics and Statistics*, pp. 409–420, 1969.

[185] Y. Grunfeld and Z. Griliches, "Is aggregation necessarily bad?" *The Review of Economics and Statistics*, pp. 1–13, 1960.

[186] N. Kourentzes, D. Barrow, and F. Petropoulos, "Another look at forecast selection and combination: Evidence from forecast pooling," *International Journal of Production Economics*, vol. 209, pp. 226–235, 2019.

[187] J. Bien, F. Bunea, and L. Xiao, "Convex banding of the covariance matrix," *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 834–845, 2016, PMID: 28042189. DOI: 10.1080/01621459.2015.1058265. eprint: https://doi.org/10.1080/01621459.2015.1058265. [Online]. Available: https://doi.org/10.1080/01621459.2015.1058265.

[188] R. J. Hyndman, A. J. Lee, and E. Wang, "Fast computation of reconciled forecasts for hierarchical and grouped time series," *Computational Statistics & Data Analysis*, vol. 97, pp. 16–32, 2016, ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2015.11.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016794731500290X.

[189] A. Sagheer, H. Hamdoun, and H. Youness, "Deep lstm-based transfer learning approach for coherent forecasts in hierarchical time series," *Sensors*, vol. 21, no. 13, p. 4379, 2021.

[190] P. Mancuso, V. Piccialli, and A. M. Sudoso, "A machine learning approach for forecasting hierarchical time series," *Expert Systems with Applications*, vol. 182, p. 115 102, 2021.

[191] N. Kourentzes and F. Petropoulos, "Forecasting with multivariate temporal aggregation: The case of promotional modelling," *International Journal of Production Economics*, vol. 181, pp. 145–153, 2016, SI: ISIR 2014, ISSN: 0925-5273. DOI: https://doi.org/10.1016/j.ijpe.2015.09.011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925527315003382.

[192] R. L. Winkler and R. T. Clemen, "Sensitivity of weights in combining forecasts," *Operations research*, vol. 40, no. 3, pp. 609–614, 1992.

[193] S. Punia, S. P. Singh, and J. K. Madaan, "A cross-temporal hierarchical framework and deep learning for supply chain forecasting," *Computers & Industrial Engineering*, vol. 149, p. 106 796, 2020.

[194] T. Di Fonzo and D. Girolimetto, "Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives," *International Journal of Forecasting*, 2021.

[195] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of multivariate analysis*, vol. 88, no. 2, pp. 365–411, 2004.

[196] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical applications in genetics and molecular biology*, vol. 4, no. 1, 2005.

[197] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1796–1804.

[198] Z. Jia, X. Huang, I. Eric, C. Chang, and Y. Xu, "Constrained deep weak supervision for histopathology image segmentation," *IEEE transactions on medical imaging*, vol. 36, no. 11, pp. 2376–2388, 2017.

[199] Y. Liu, A. W. K. Kong, and C. K. Goh, "A constrained deep neural network for ordinal regression," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 831–839.

[200] J. Drgoňa, A. R. Tuor, V. Chandan, and D. L. Vrabie, "Physics-constrained deep learning of multi-zone building thermal dynamics," *Energy and Buildings*, vol. 243, p. 110 992, 2021.

[201] P. Márquez-Neila, M. Salzmann, and P. Fua, "Imposing hard constraints on deep networks: Promises and limitations," *arXiv preprint arXiv:1706.02025*, 2017.

[202] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv preprint arXiv:1109.2378*, 2011.

[203] D. N. Reshef, Y. A. Reshef, H. K. Finucane, *et al.*, "Detecting novel associations in large data sets," *science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[204] C. Miller, "What's in the box?! towards explainable machine learning applied to non-residential building smart meter classification," *Energy and Buildings*, vol. 199, pp. 523–536, 2019.

[205] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[206] H. Hewamalage, K. Ackermann, and C. Bergmeir, "Forecast evaluation for data scientists: Common pitfalls and best practices," *arXiv preprint arXiv:2203.10716*, 2022.

[207] N. Bjorck, C. P. Gomes, B. Selman, and K. Q. Weinberger, "Understanding batch normalization," *Advances in neural information processing systems*, vol. 31, 2018.

[208] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[209] C. Garbin, X. Zhu, and O. Marques, "Dropout vs. batch normalization: An empirical study of their impact to deep learning," *Multimedia Tools and Applications*, vol. 79, no. 19, pp. 12 777–12 815, 2020.

[210] Martín Abadi, Ashish Agarwal, Paul Barham, *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: `https://www.tensorflow.org/`.

[211] A. Saad, T. Youssef, A. T. Elsayed, A. Amin, O. H. Abdalla, and O. Mohammed, "Data-centric hierarchical distributed model predictive control for smart grid energy management," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4086–4098, 2018.

[212] B. Jiang and Y. Fei, "Smart home in smart microgrid: A cost-effective energy ecosystem with intelligent hierarchical agents," *IEEE Transactions on Smart Grid*, vol. 6, no. 1, pp. 3–13, 2014.

[213] J. Leprince, A. Schledorn, D. Guericke, D. F. Dominkovic, H. Madsen, and W. Zeiler, *Can occupant behaviors affect urban energy planning? distributed stochastic optimization for energy communities*, 2023. DOI: 10.48550/ARXIV.2303.03006. [Online]. Available: https://arxiv.org/abs/2303.03006.

[214] IEA, *Global energy-related co2 emissions by sector, 2020*, 2020. [Online]. Available: https://www.iea.org/data-and-statistics/charts/global-energy-related-co2-emissions-by-sector-2020.

[215] S. Zhang, K. Wang, W. Xu, *et al.*, "Policy recommendations for the zero energy building promotion towards carbon neutral in asia-pacific region," *Energy Policy*, vol. 159, p. 112 661, 2021.

[216] Y. Lu, X. Zhu, and Q. Cui, "Effectiveness and equity implications of carbon policies in the united states construction industry," *Building and Environment*, vol. 49, pp. 259–269, 2012.

[217] M. Economidou, V. Todeschi, P. Bertoldi, D. D'Agostino, P. Zangheri, and L. Castellazzi, "Review of 50 years of eu energy efficiency policies for buildings," *Energy and Buildings*, vol. 225, p. 110 322, 2020.

[218] D. D'agostino, P. Zangheri, and L. Castellazzi, "Towards nearly zero energy buildings in europe: A focus on retrofit in non-residential buildings," *Energies*, vol. 10, no. 1, p. 117, 2017.

[219] D. Eltigani and S. Masri, "Challenges of integrating renewable energy sources to smart grids: A review," *Renewable and Sustainable Energy Reviews*, vol. 52, pp. 770–780, 2015.

[220] P. Palensky and D. Dietrich, "Demand side management: Demand response, intelligent energy systems, and smart loads," *IEEE transactions on industrial informatics*, vol. 7, no. 3, pp. 381–388, 2011.

[221] E. Commission, *Proposal for a directive of the european parliament and of the council on common rules for the internal market in electricity*, 2017.

[222] M.-H. Kim, D. Kim, J. Heo, and D.-W. Lee, "Techno-economic analysis of hybrid renewable energy system with solar district heating for net zero energy community," *Energy*, vol. 187, p. 115 916, 2019.

[223] H. ur Rehman, F. Reda, S. Paiho, and A. Hasan, "Towards positive energy communities at high latitudes," *Energy conversion and management*, vol. 196, pp. 175–195, 2019.

[224] V. Z. Gjorgievski, S. Cundeva, and G. E. Georghiou, "Social arrangements, technical designs and impacts of energy communities: A review," *Renewable Energy*, vol. 169, pp. 1138–1156, 2021.

[225] K. Orehounig, R. Evins, and V. Dorer, "Integration of decentralized energy systems in neighbourhoods using the energy hub approach," *Applied Energy*, vol. 154, pp. 277–289, 2015.

[226] M. Mohammadi, Y. Noorollahi, B. Mohammadi-ivatloo, M. Hosseinzadeh, H. Yousefi, and S. T. Khorasani, "Optimal management of energy hubs and smart energy hubs–a review," *Renewable and Sustainable Energy Reviews*, vol. 89, pp. 33–50, 2018.

[227] A. Maroufmashat, A. Elkamel, M. Fowler, *et al.*, "Modeling and optimization of a network of energy hubs to improve economic and emission considerations," *Energy*, vol. 93, pp. 2546–2558, 2015.

[228] S. Moret, "Strategic energy planning under uncertainty," EPFL, Tech. Rep., 2017.

[229] A. Haurie, M. Tavoni, and B. Van der Zwaan, "Modeling uncertainty and the economics of climate change: Recommendations for robust energy policy," *Environmental Modeling & Assessment*, vol. 17, no. 1, pp. 1–5, 2012.

[230] M. J. Santos, P. Ferreira, and M. Araújo, "A methodology to incorporate risk and uncertainty in electricity power planning," *Energy*, vol. 115, pp. 1400–1411, 2016.

[231] W. Usher and N. Strachan, "An expert elicitation of climate, energy and economic uncertainties," *Energy policy*, vol. 61, pp. 811–821, 2013.

[232] J. Virote and R. Neves-Silva, "Stochastic models for building energy prediction based on occupant behavior assessment," *Energy and buildings*, vol. 53, pp. 183–193, 2012.

[233] G. Happle, J. A. Fonseca, and A. Schlueter, "A review on occupant behavior in urban building energy models," *Energy and Buildings*, vol. 174, pp. 276–292, 2018.

[234] P. Stadler, A. Ashouri, and F. Maréchal, "Model-based optimization of distributed and renewable energy systems in buildings," *Energy and Buildings*, vol. 120, pp. 103–113, 2016.

[235] P. Stadler, L. Girardin, A. Ashouri, and F. Maréchal, "Contribution of model predictive control in the integration of renewable energy sources within the built environment," *Frontiers in Energy Research*, vol. 6, p. 22, 2018.

[236] A. Ashouri, S. S. Fux, M. J. Benz, and L. Guzzella, "Optimal design and operation of building services using mixed-integer linear programming techniques," *Energy*, vol. 59, pp. 365–376, 2013.

[237] T. Schütz, L. Schiffer, H. Harb, M. Fuchs, and D. Müller, "Optimal design of energy conversion units and envelopes for residential building retrofits using a comprehensive milp model," *Applied Energy*, vol. 185, pp. 1–15, 2017.

[238] S. Fazlollahi and F. Maréchal, "Multi-objective, multi-period optimization of biomass conversion technologies using evolutionary algorithms and mixed integer linear programming (milp)," *Applied thermal engineering*, vol. 50, no. 2, pp. 1504–1513, 2013.

[239] J. E. Braun, "Load control using building thermal mass," *J. Sol. Energy Eng.*, vol. 125, no. 3, pp. 292–301, 2003.

[240] A. Fazeli, E. Christopher, C. M. Johnson, M. Gillott, and M. Sumner, "Investigating the effects of dynamic demand side management within intelligent smart energy

communities of future decentralized power system," in *2011 2nd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies*, IEEE, 2011, pp. 1–8.

[241] A. Sheikhi, M. Rayati, S. Bahrami, and A. M. Ranjbar, "Integrated demand side management game in smart energy hubs," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 675–683, 2015.

[242] S. Rahgozar, A. Z. G. Seyyedi, and P. Siano, "A resilience-oriented planning of energy hub by considering demand response program and energy storage systems," *Journal of Energy Storage*, vol. 52, p. 104 841, 2022.

[243] P. Bacher and H. Madsen, "Identifying suitable models for the heat dynamics of buildings," *Energy and Buildings*, vol. 43, no. 7, pp. 1511–1522, 2011.

[244] I. Sarbu and C. Sebarchievici, "General review of ground-source heat pump systems for heating and cooling of buildings," *Energy and buildings*, vol. 70, pp. 441–454, 2014.

[245] P. Huang, M. Lovati, X. Zhang, *et al.*, "Transforming a residential building cluster into electricity prosumers in sweden: Optimal design of a coupled pv-heat pump-thermal storage-electric vehicle system," *Applied Energy*, vol. 255, p. 113 864, 2019.

[246] S. Smith, "Solar panels set to be mandatory on all new buildings under eu plan," *Independent*, May 18, 2022. [Online]. Available: `https://www.independent.co.uk/climate-change/news/solar-panels-new-buildings-eu-mandatory-b2081732.html` (visited on 01/13/2023).

[247] P. Murray, K. Orehounig, D. Grosspietsch, and J. Carmeliet, "A comparison of storage systems in neighbourhood decentralized energy system applications from 2015 to 2050," *Applied Energy*, vol. 231, pp. 1285–1306, 2018.

[248] P. Gabrielli, M. Gazzani, E. Martelli, and M. Mazzotti, "Optimal design of multi-energy systems with seasonal storage," *Applied Energy*, vol. 219, pp. 408–424, 2018.

[249] G. Darivianakis, A. Eichler, R. S. Smith, and J. Lygeros, "A data-driven stochastic optimization approach to the seasonal storage energy management," *IEEE control systems letters*, vol. 1, no. 2, pp. 394–399, 2017.

[250] V. Bansal, J. D. Perkins, and E. N. Pistikopoulos, "A case study in simultaneous design and control using rigorous, mixed-integer dynamic optimization models," *Industrial & Engineering Chemistry Research*, vol. 41, no. 4, pp. 760–778, 2002.

[251] O. Van Cutsem, D. H. Dac, P. Boudou, and M. Kayal, "Cooperative energy management of a community of smart-buildings: A blockchain approach," *International Journal of electrical power & energy systems*, vol. 117, p. 105 643, 2020.

[252] J. M. F. Rager, "Urban energy system design from the heat perspective using mathematical programming including thermal storage," EPFL, Tech. Rep., 2015.

[253] A. E. Dudek, J. Leśkow, E. Paparoditis, and D. N. Politis, "A generalized block bootstrap for seasonal time series," *Journal of Time Series Analysis*, vol. 35, no. 2, pp. 89–114, 2014.

[254] D. Guericke, A. Schledorn, and H. Madsen, *A generic stochastic network flow formulation for production optimization of district heating systems*, 2022. DOI: 10.48550/ARXIV.2211.00934. [Online]. Available: https://arxiv.org/abs/2211.00934.

[255] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert systems with applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[256] I. Blanco, D. Guericke, A. N. Andersen, and H. Madsen, "Operational planning and bidding for district heating systems with uncertain renewable energy production," *Energies*, vol. 11, no. 12, 2018, ISSN: 1996-1073. DOI: 10.3390/en11123310. [Online]. Available: https://www.mdpi.com/1996-1073/11/12/3310.

[257] G. Mavromatidis, K. Orehounig, and J. Carmeliet, "Design of distributed energy systems under uncertainty: A two-stage stochastic programming approach," *Applied energy*, vol. 222, pp. 932–950, 2018.

[258] Z. Zhou, J. Zhang, P. Liu, Z. Li, M. C. Georgiadis, and E. N. Pistikopoulos, "A two-stage stochastic programming model for the optimal design of distributed energy systems," *Applied Energy*, vol. 103, pp. 135–144, 2013.

[259] A. J. Conejo, M. Carrión, J. M. Morales, *et al.*, *Decision making under uncertainty in electricity markets*. Springer, 2010, vol. 1.

[260] J. Morio, "Global and local sensitivity analysis methods for a physical system," *European journal of physics*, vol. 32, no. 6, p. 1577, 2011.

[261] S. Kucherenko, M. Rodriguez-Fernandez, C. Pantelides, and N. Shah, "Monte carlo evaluation of derivative-based global sensitivity measures," *Reliability Engineering & System Safety*, vol. 94, no. 7, pp. 1135–1148, 2009.

[262] S. Moret, V. C. Gironès, M. Bierlaire, and F. Maréchal, "Characterization of input uncertainties in strategic energy planning models," *Applied energy*, vol. 202, pp. 597–617, 2017.

[263] R. Amaral Lopes, R. GrÃ¸nborg Junker, J. Martins, J. Murta-Pina, G. Reynders, and H. Madsen, "Characterisation and use of energy flexibility in water pumping and storage systems," *Applied Energy*, vol. 277, p. 115 587, 2020, ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2020.115587. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261920310989.

[264] C. C. Marín-Cano, J. E. Sierra-Aguilar, J. M. López-Lezama, Á. Jaramillo-Duque, and J. G. Villegas, "A novel strategy to reduce computational burden of the stochastic security constrained unit commitment problem," *Energies*, vol. 13, no. 15, p. 3777, 2020.

[265] J.-F. Toubeau, Z. De Grève, and F. Vallèe, "Medium-term multimarket optimization for virtual power plants: A stochastic-based decision environment," *IEEE Transactions on Power Systems*, vol. 33, no. 2, pp. 1399–1410, 2017.

[266] Q. P. Zheng, J. Wang, and A. L. Liu, "Stochastic optimization for unit commitment—a review," *IEEE Transactions on Power Systems*, vol. 30, no. 4, pp. 1913–1924, 2014.

[267] G. T. Costanzo, O. Gehrke, D. E. M. Bondy, *et al.*, "A coordination scheme for distributed model predictive control: Integration of flexible ders," in *IEEE PES ISGT Europe 2013*, IEEE, 2013, pp. 1–5.

[268] M. Stephant, D. Abbes, K. Hassam-Ouari, A. Labrunie, and B. Robyns, "Distributed optimization of energy profiles to improve photovoltaic self-consumption on a local energy community," *Simulation Modelling Practice and Theory*, vol. 108, p. 102 242, 2021.

[269] P. Stadler, A. Ashouri, and F. Maréchal, "Distributed model predictive control of energy systems in microgrids," in *2016 Annual IEEE Systems Conference (SysCon)*, IEEE, 2016, pp. 1–6.

[270] S. Mitchell, M. OSullivan, and I. Dunning, "Pulp: A linear programming toolkit for python," *The University of Auckland, Auckland, New Zealand*, vol. 65, 2011.

[271] Gurobi Optimization, LLC, *Gurobi Optimizer Reference Manual*, 2023. [Online]. Available: `https://www.gurobi.com`.

[272] ENTSO-E, *Electricity day-ahead prices*, Jan. 2023. [Online]. Available: `https://transparency.entsoe.eu/transmission-domain/r2/dayAheadPrices/show`.

[273] eurostat, *Natural gas price statistics*, Jan. 2023. [Online]. Available: `https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Natural_gas_price_statistics`.

[274] Belastingdienst, *Tables rates environmental taxes*, 2021. [Online]. Available: `https://www.belastingdienst.nl/wps/wcm/connect/bldcontentnl/belastingdienst/zakelijk/overige_belastingen/belastingen_op_milieugrondslag/tarieven_milieubelastingen/tabellen_tarieven_milieubelastingen`.

[275] N. A. for Consumers and Markets, *Decision setting the maximum electricity distribution tariffs with effect from january 1st, 2021 for saba electric company n.v*, 2021. [Online]. Available: `https://www.acm.nl/sites/default/files/documents/distribution-tariffs-sec-electricity-2021.pdf`.

[276] D. E. Agency, *Technology data*, Jan. 2023. [Online]. Available: `https://ens.dk/en/our-services/projections-and-models/technology-data`.

[277] *Ashrae - great energy predictor iii*, 2020. [Online]. Available: `https://www.kaggle.com/c/ashrae-energy-prediction`.

[278] *The citylearn challenge - multi-agend reinforcement learning for intelligent energy management*, 2021. [Online]. Available: `https://sites.google.com/view/citylearnchallenge`.

[279] Q. Qu, X. Li, Y. Zhou, *et al.*, *An improved reinforcement learning algorithm for learning to branch*, 2022. DOI: `10.48550/ARXIV.2201.06213`. [Online]. Available: `https://arxiv.org/abs/2201.06213`.

[280] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st international conference on neural information processing systems*, 2017, pp. 4768–4777.

[281] B. McKay, M. Willis, and G. Barton, "Using a tree structured genetic algorithm to perform symbolic regression," in *First International Conference on Genetic Algorithms in Engineering Systems: Innovations and Applications*, 1995, pp. 487–492. DOI: 10.1049/cp:19951096.

[282] M. Cranmer, A. Sanchez-Gonzalez, P. Battaglia, *et al.*, *Discovering symbolic models from deep learning with inductive biases*, 2020. arXiv: 2006.11287 [cs.LG].

Bouwstenen is een publicatiereeks
van de Faculteit Bouwkunde,
Technische Universiteit Eindhoven.
Zij presenteert resultaten van
onderzoek en andere activiteiten op
het vakgebied der Bouwkunde,
uitgevoerd in het kader van deze
Faculteit.

Bouwstenen en andere proefschriften van de
TU/e zijn online beschikbaar via:
https://research.tue.nl/

Reeds verschenen in de serie
Bouwstenen

nr 1
Elan: A Computer Model for Building Energy Design: Theory and Validation
Martin H. de Wit
H.H. Driessen
R.M.M. van der Velden

nr 2
Kwaliteit, Keuzevrijheid en Kosten: Evaluatie van Experiment Klarendal, Arnhem
J. Smeets
C. le Nobel
M. Broos
J. Frenken
A. v.d. Sanden

nr 3
Crooswijk: Van 'Bijzonder' naar 'Gewoon'
Vincent Smit
Kees Noort

nr 4
Staal in de Woningbouw
Edwin J.F. Delsing

nr 5
Mathematical Theory of Stressed Skin Action in Profiled Sheeting with Various Edge Conditions
Andre W.A.M.J. van den Bogaard

nr 6
Hoe Berekenbaar en Betrouwbaar is de Coëfficiënt k in x-ksigma en x-ks?
K.B. Lub
A.J. Bosch

nr 7
Het Typologisch Gereedschap: Een Verkennende Studie Omtrent Typologie en Omtrent de Aanpak van Typologisch Onderzoek
J.H. Luiten

nr 8
Informatievoorziening en Beheerprocessen
A. Nauta
Jos Smeets (red.)
Helga Fassbinder (projectleider)
Adrie Proveniers
J. v.d. Moosdijk

nr 9
Strukturering en Verwerking van Tijdgegevens voor de Uitvoering van Bouwwerken
ir. W.F. Schaefer
P.A. Erkelens

nr 10
Stedebouw en de Vorming van een Speciale Wetenschap
K. Doevendans

nr 11
Informatica en Ondersteuning van Ruimtelijke Besluitvorming
G.G. van der Meulen

nr 12
Staal in de Woningbouw, Korrosie-Bescherming van de Begane Grondvloer
Edwin J.F. Delsing

nr 13
Een Thermisch Model voor de Berekening van Staalplaatbetonvloeren onder Brandomstandigheden
A.F. Hamerlinck

nr 14
De Wijkgedachte in Nederland: Gemeenschapsstreven in een Stedebouwkundige Context
K. Doevendans
R. Stolzenburg

nr 15
Diaphragm Effect of Trapezoidally Profiled Steel Sheets: Experimental Research into the Influence of Force Application
Andre W.A.M.J. van den Bogaard

nr 16
Versterken met Spuit-Ferrocement: Het Mechanische Gedrag van met Spuit-Ferrocement Versterkte Gewapend Betonbalken
K.B. Lubir
M.C.G. van Wanroy

nr 17
De Tractaten van Jean Nicolas Louis Durand
G. van Zeyl

nr 18
Wonen onder een Plat Dak: Drie Opstellen over Enkele Vooronderstellingen van de Stedebouw
K. Doevendans

nr 19
Supporting Decision Making Processes: A Graphical and Interactive Analysis of Multivariate Data
W. Adams

nr 20
Self-Help Building Productivity: A Method for Improving House Building by Low-Income Groups Applied to Kenya 1990-2000
P. A. Erkelens

nr 21
De Verdeling van Woningen: Een Kwestie van Onderhandelen
Vincent Smit

nr 22
Flexibiliteit en Kosten in het Ontwerpproces: Een Besluitvormingondersteunend Model
M. Prins

nr 23
Spontane Nederzettingen Begeleid: Voorwaarden en Criteria in Sri Lanka
Po Hin Thung

nr 24
Fundamentals of the Design of Bamboo Structures
Oscar Arce-Villalobos

nr 25
Concepten van de Bouwkunde
M.F.Th. Bax (red.)
H.M.G.J. Trum (red.)

nr 26
Meaning of the Site
Xiaodong Li

nr 27
Het Woonmilieu op Begrip Gebracht: Een Speurtocht naar de Betekenis van het Begrip 'Woonmilieu'
Jaap Ketelaar

nr 28
Urban Environment in Developing Countries
editors: Peter A. Erkelens
George G. van der Meulen (red.)

nr 29
Stategische Plannen voor de Stad: Onderzoek en Planning in Drie Steden
prof.dr. H. Fassbinder (red.)
H. Rikhof (red.)

nr 30
Stedebouwkunde en Stadsbestuur
Piet Beekman

nr 31
De Architectuur van Djenné: Een Onderzoek naar de Historische Stad
P.C.M. Maas

nr 32
Conjoint Experiments and Retail Planning
Harmen Oppewal

nr 33
Strukturformen Indonesischer Bautechnik: Entwicklung Methodischer Grundlagen für eine 'Konstruktive Pattern Language' in Indonesien
Heinz Frick arch. SIA

nr 34
Styles of Architectural Designing: Empirical Research on Working Styles and Personality Dispositions
Anton P.M. van Bakel

nr 35
Conjoint Choice Models for Urban Tourism Planning and Marketing
Benedict Dellaert

nr 36
Stedelijke Planvorming als Co-Produktie
Helga Fassbinder (red.)

nr 37
Design Research in the Netherlands
editors: R.M. Oxman
M.F.Th. Bax
H.H. Achten

nr 38
Communication in the Building Industry
Bauke de Vries

nr 39
Optimaal Dimensioneren van Gelaste Plaatliggers
J.B.W. Stark
F. van Pelt
L.F.M. van Gorp
B.W.E.M. van Hove

nr 40
Huisvesting en Overwinning van Armoede
P.H. Thung
P. Beekman (red.)

nr 41
Urban Habitat: The Environment of Tomorrow
George G. van der Meulen
Peter A. Erkelens

nr 42
A Typology of Joints
John C.M. Olie

nr 43
Modeling Constraints-Based Choices for Leisure Mobility Planning
Marcus P. Stemerding

nr 44
Activity-Based Travel Demand Modeling
Dick Ettema

nr 45
Wind-Induced Pressure Fluctuations on Building Facades
Chris Geurts

nr 46
Generic Representations
Henri Achten

nr 47
Johann Santini Aichel: Architectuur en Ambiguiteit
Dirk De Meyer

nr 48
Concrete Behaviour in Multiaxial Compression
Erik van Geel

nr 49
Modelling Site Selection
Frank Witlox

nr 50
Ecolemma Model
Ferdinand Beetstra

nr 51
Conjoint Approaches to Developing Activity-Based Models
Donggen Wang

nr 52
On the Effectiveness of Ventilation
Ad Roos

nr 53
Conjoint Modeling Approaches for Residential Group preferences
Eric Molin

nr 54
Modelling Architectural Design Information by Features
Jos van Leeuwen

nr 55
A Spatial Decision Support System for the Planning of Retail and Service Facilities
Theo Arentze

nr 56
Integrated Lighting System Assistant
Ellie de Groot

nr 57
Ontwerpend Leren, Leren Ontwerpen
J.T. Boekholt

nr 58
Temporal Aspects of Theme Park Choice Behavior
Astrid Kemperman

nr 59
Ontwerp van een Geïndustrialiseerde Funderingswijze
Faas Moonen

nr 60
Merlin: A Decision Support System for Outdoor Leisure Planning
Manon van Middelkoop

nr 61
The Aura of Modernity
Jos Bosman

nr 62
Urban Form and Activity-Travel Patterns
Daniëlle Snellen

nr 63
Design Research in the Netherlands 2000
Henri Achten

nr 64
Computer Aided Dimensional Control in Building Construction
Rui Wu

nr 65
Beyond Sustainable Building
editors: Peter A. Erkelens
Sander de Jonge
August A.M. van Vliet
co-editor: Ruth J.G. Verhagen

nr 66
Das Globalrecyclingfähige Haus
Hans Löfflad

nr 67
Cool Schools for Hot Suburbs
René J. Dierkx

nr 68
A Bamboo Building Design Decision Support Tool
Fitri Mardjono

nr 69
Driving Rain on Building Envelopes
Fabien van Mook

nr 70
Heating Monumental Churches
Henk Schellen

nr 71
Van Woningverhuurder naar Aanbieder van Woongenot
Patrick Dogge

nr 72
Moisture Transfer Properties of Coated Gypsum
Emile Goossens

nr 73
Plybamboo Wall-Panels for Housing
Guillermo E. González-Beltrán

nr 74
The Future Site-Proceedings
Ger Maas
Frans van Gassel

nr 75
Radon transport in Autoclaved Aerated Concrete
Michel van der Pal

nr 76
The Reliability and Validity of Interactive Virtual Reality Computer Experiments
Amy Tan

nr 77
Measuring Housing Preferences Using Virtual Reality and Belief Networks
Maciej A. Orzechowski

nr 78
Computational Representations of Words and Associations in Architectural Design
Nicole Segers

nr 79
Measuring and Predicting Adaptation in Multidimensional Activity-Travel Patterns
Chang-Hyeon Joh

nr 80
Strategic Briefing
Fayez Al Hassan

nr 81
Well Being in Hospitals
Simona Di Cicco

nr 82
Solares Bauen: Implementierungs- und Umsetzungs-Aspekte in der Hochschulausbildung in Österreich
Gerhard Schuster

nr 83
Supporting Strategic Design of Workplace Environments with Case-Based Reasoning
Shauna Mallory-Hill

nr 84
ACCEL: A Tool for Supporting Concept Generation in the Early Design Phase
Maxim Ivashkov

nr 85
Brick-Mortar Interaction in Masonry under Compression
Ad Vermeltfoort

nr 86
Zelfredzaam Wonen
Guus van Vliet

nr 87
Een Ensemble met Grootstedelijke Allure
Jos Bosman
Hans Schippers

nr 88
On the Computation of Well-Structured Graphic Representations in Architectural Design
Henri Achten

nr 89
De Evolutie van een West-Afrikaanse Vernaculaire Architectuur
Wolf Schijns

nr 90
ROMBO Tactiek
Christoph Maria Ravesloot

nr 91
External Coupling between Building Energy Simulation and Computational Fluid Dynamics
Ery Djunaedy

nr 92
Design Research in the Netherlands 2005
editors: Henri Achten
Kees Dorst
Pieter Jan Stappers
Bauke de Vries

nr 93
Ein Modell zur Baulichen Transformation
Jalil H. Saber Zaimian

nr 94
Human Lighting Demands: Healthy Lighting in an Office Environment
Myriam Aries

nr 95
A Spatial Decision Support System for the Provision and Monitoring of Urban Greenspace
Claudia Pelizaro

nr 96
Leren Creëren
Adri Proveniers

nr 97
Simlandscape
Rob de Waard

nr 98
Design Team Communication
Ad den Otter

nr 99
Humaan-Ecologisch Georiënteerde Woningbouw
Juri Czabanowski

nr 100
Hambase
Martin de Wit

nr 101
Sound Transmission through Pipe Systems and into Building Structures
Susanne Bron-van der Jagt

nr 102
Het Bouwkundig Contrapunt
Jan Francis Boelen

nr 103
A Framework for a Multi-Agent Planning Support System
Dick Saarloos

nr 104
Bracing Steel Frames with Calcium Silicate Element Walls
Bright Mweene Ng'andu

nr 105
Naar een Nieuwe Houtskeletbouw
F.N.G. De Medts

nr 106 and 107
Niet gepubliceerd

nr 108
Geborgenheid
T.E.L. van Pinxteren

nr 109
Modelling Strategic Behaviour in Anticipation of Congestion
Qi Han

nr 110
Reflecties op het Woondomein
Fred Sanders

nr 111
On Assessment of Wind Comfort by Sand Erosion
Gábor Dezsö

nr 112
Bench Heating in Monumental Churches
Dionne Limpens-Neilen

nr 113
RE. Architecture
Ana Pereira Roders

nr 114
Toward Applicable Green Architecture
Usama El Fiky

nr 115
Knowledge Representation under Inherent Uncertainty in a Multi-Agent System for Land Use Planning
Liying Ma

nr 116
Integrated Heat Air and Moisture Modeling and Simulation
Jos van Schijndel

nr 117
Concrete Behaviour in Multiaxial Compression
J.P.W. Bongers

nr 118
The Image of the Urban Landscape
Ana Moya Pellitero

nr 119
The Self-Organizing City in Vietnam
Stephanie Geertman

nr 120
A Multi-Agent Planning Support System for Assessing Externalities of Urban Form Scenarios
Rachel Katoshevski-Cavari

nr 121
Den Schulbau Neu Denken, Fühlen und Wollen
Urs Christian Maurer-Dietrich

nr 122
Peter Eisenman Theories and Practices
Bernhard Kormoss

nr 123
User Simulation of Space Utilisation
Vincent Tabak

nr 125
In Search of a Complex System Model
Oswald Devisch

nr 126
Lighting at Work: Environmental Study of Direct Effects of Lighting Level and Spectrum on Psycho-Physiological Variables
Grazyna Górnicka

nr 127
Flanking Sound Transmission through Lightweight Framed Double Leaf Walls
Stefan Schoenwald

nr 128
Bounded Rationality and Spatio-Temporal Pedestrian Shopping Behavior
Wei Zhu

nr 129
Travel Information: Impact on Activity Travel Pattern
Zhongwei Sun

nr 130
Co-Simulation for Performance Prediction of Innovative Integrated Mechanical Energy Systems in Buildings
Marija Trčka

nr 131
Niet gepubliceerd

nr 132
Architectural Cue Model in Evacuation Simulation for Underground Space Design
Chengyu Sun

nr 133
Uncertainty and Sensitivity Analysis in Building Performance Simulation for Decision Support and Design Optimization
Christina Hopfe

nr 134
Facilitating Distributed Collaboration in the AEC/FM Sector Using Semantic Web Technologies
Jacob Beetz

nr 135
Circumferentially Adhesive Bonded Glass Panes for Bracing Steel Frame in Façades
Edwin Huveners

nr 136
Influence of Temperature on Concrete Beams Strengthened in Flexure with CFRP
Ernst-Lucas Klamer

nr 137
Sturen op Klantwaarde
Jos Smeets

nr 139
Lateral Behavior of Steel Frames with Discretely Connected Precast Concrete Infill Panels
Paul Teewen

nr 140
Integral Design Method in the Context of Sustainable Building Design
Perica Savanović

nr 141
Household Activity-Travel Behavior: Implementation of Within-Household Interactions
Renni Anggraini

nr 142
Design Research in the Netherlands 2010
Henri Achten

nr 143
Modelling Life Trajectories and Transport Mode Choice Using Bayesian Belief Networks
Marloes Verhoeven

nr 144
Assessing Construction Project Performance in Ghana
William Gyadu-Asiedu

nr 145
Empowering Seniors through Domotic Homes
Masi Mohammadi

nr 146
An Integral Design Concept for Ecological Self-Compacting Concrete
Martin Hunger

nr 147
Governing Multi-Actor Decision Processes in Dutch Industrial Area Redevelopment
Erik Blokhuis

nr 148
A Multifunctional Design Approach for Sustainable Concrete
Götz Hüsken

nr 149
Quality Monitoring in Infrastructural Design-Build Projects
Ruben Favié

nr 150
Assessment Matrix for Conservation of Valuable Timber Structures
Michael Abels

nr 151
Co-simulation of Building Energy Simulation and Computational Fluid Dynamics for Whole-Building Heat, Air and Moisture Engineering
Mohammad Mirsadeghi

nr 152
External Coupling of Building Energy Simulation and Building Element Heat, Air and Moisture Simulation
Daniel Cóstola

nr 153
Adaptive Decision Making In Multi-Stakeholder Retail Planning
Ingrid Janssen

nr 154
Landscape Generator
Kymo Slager

nr 155
Constraint Specification in Architecture
Remco Niemeijer

nr 156
A Need-Based Approach to Dynamic Activity Generation
Linda Nijland

nr 157
Modeling Office Firm Dynamics in an Agent-Based Micro Simulation Framework
Gustavo Garcia Manzato

nr 158
Lightweight Floor System for Vibration Comfort
Sander Zegers

nr 159
Aanpasbaarheid van de Draagstructuur
Roel Gijsbers

nr 160
'Village in the City' in Guangzhou, China
Yanliu Lin

nr 161
Climate Risk Assessment in Museums
Marco Martens

nr 162
Social Activity-Travel Patterns
Pauline van den Berg

nr 163
Sound Concentration Caused by Curved Surfaces
Martijn Vercammen

nr 164
Design of Environmentally Friendly Calcium Sulfate-Based Building Materials: Towards an Improved Indoor Air Quality
Qingliang Yu

nr 165
Beyond Uniform Thermal Comfort on the Effects of Non-Uniformity and Individual Physiology
Lisje Schellen

nr 166
Sustainable Residential Districts
Gaby Abdalla

nr 167
Towards a Performance Assessment Methodology using Computational Simulation for Air Distribution System Designs in Operating Rooms
Mônica do Amaral Melhado

nr 168
Strategic Decision Modeling in Brownfield Redevelopment
Brano Glumac

nr 169
Pamela: A Parking Analysis Model for Predicting Effects in Local Areas
Peter van der Waerden

nr 170
A Vision Driven Wayfinding Simulation-System Based on the Architectural Features Perceived in the Office Environment
Qunli Chen

nr 171
Measuring Mental Representations Underlying Activity-Travel Choices
Oliver Horeni

nr 172
Modelling the Effects of Social Networks on Activity and Travel Behaviour
Nicole Ronald

nr 173
Uncertainty Propagation and Sensitivity Analysis Techniques in Building Performance Simulation to Support Conceptual Building and System Design
Christian Struck

nr 174
Numerical Modeling of Micro-Scale Wind-Induced Pollutant Dispersion in the Built Environment
Pierre Gousseau

nr 175
Modeling Recreation Choices over the Family Lifecycle
Anna Beatriz Grigolon

nr 176
Experimental and Numerical Analysis of Mixing Ventilation at Laminar, Transitional and Turbulent Slot Reynolds Numbers
Twan van Hooff

nr 177
Collaborative Design Support: Workshops to Stimulate Interaction and Knowledge Exchange Between Practitioners
Emile M.C.J. Quanjel

nr 178
Future-Proof Platforms for Aging-in-Place
Michiel Brink

nr 179
Motivate: A Context-Aware Mobile Application for Physical Activity Promotion
Yuzhong Lin

nr 180
Experience the City: Analysis of Space-Time Behaviour and Spatial Learning
Anastasia Moiseeva

nr 181
Unbonded Post-Tensioned Shear Walls of Calcium Silicate Element Masonry
Lex van der Meer

nr 182
Construction and Demolition Waste Recycling into Innovative Building Materials for Sustainable Construction in Tanzania
Mwita M. Sabai

nr 183
Durability of Concrete with Emphasis on Chloride Migration
Przemysław Spiesz

nr 184
Computational Modeling of Urban Wind Flow and Natural Ventilation Potential of Buildings
Rubina Ramponi

nr 185
A Distributed Dynamic Simulation Mechanism for Buildings Automation and Control Systems
Azzedine Yahiaoui

nr 186
Modeling Cognitive Learning of Urban Networks in Daily Activity-Travel Behavior
Şehnaz Cenani Durmazoğlu

nr 187
Functionality and Adaptability of Design Solutions for Public Apartment Buildings in Ghana
Stephen Agyefi-Mensah

nr 188
A Construction Waste Generation Model for Developing Countries
Lilliana Abarca-Guerrero

nr 189
Synchronizing Networks: The Modeling of Supernetworks for Activity-Travel Behavior
Feixiong Liao

nr 190
Time and Money Allocation Decisions in Out-of-Home Leisure Activity Choices
Gamze Zeynep Dane

nr 191
How to Measure Added Value of CRE and Building Design
Rianne Appel-Meulenbroek

nr 192
Secondary Materials in Cement-Based Products: Treatment, Modeling and Environmental Interaction
Miruna Florea

nr 193
Concepts for the Robustness Improvement of Self-Compacting Concrete: Effects of Admixtures and Mixture Components on the Rheology and Early Hydration at Varying Temperatures
Wolfram Schmidt

nr 194
Modelling and Simulation of Virtual Natural Lighting Solutions in Buildings
Rizki A. Mangkuto

nr 195
Nano-Silica Production at Low Temperatures from the Dissolution of Olivine - Synthesis, Tailoring and Modelling
Alberto Lazaro Garcia

nr 196
Building Energy Simulation Based Assessment of Industrial Halls for Design Support
Bruno Lee

nr 197
Computational Performance Prediction of the Potential of Hybrid Adaptable Thermal Storage Concepts for Lightweight Low-Energy Houses
Pieter-Jan Hoes

nr 198
Application of Nano-Silica in Concrete
George Quercia Bianchi

nr 199
Dynamics of Social Networks and Activity Travel Behaviour
Fariya Sharmeen

nr 200
Building Structural Design Generation and Optimisation including Spatial Modification
Juan Manuel Davila Delgado

nr 201
Hydration and Thermal Decomposition of Cement/Calcium-Sulphate Based Materials
Ariën de Korte

nr 202
Republiek van Beelden:
De Politieke Werkingen van het Ontwerp in Regionale Planvorming
Bart de Zwart

nr 203
Effects of Energy Price Increases on Individual Activity-Travel Repertoires and Energy Consumption
Dujuan Yang

nr 204
Geometry and Ventilation:
Evaluation of the Leeward Sawtooth Roof Potential in the Natural Ventilation of Buildings
Jorge Isaac Perén Montero

nr 205
Computational Modelling of Evaporative Cooling as a Climate Change Adaptation Measure at the Spatial Scale of Buildings and Streets
Hamid Montazeri

nr 206
Local Buckling of Aluminium Beams in Fire Conditions
Ronald van der Meulen

nr 207
Historic Urban Landscapes:
Framing the Integration of Urban and Heritage Planning in Multilevel Governance
Loes Veldpaus

nr 208
Sustainable Transformation of the Cities: Urban Design Pragmatics to Achieve a Sustainable City
Ernesto Antonio Zumelzu Scheel

nr 209
Development of Sustainable Protective Ultra-High Performance Fibre Reinforced Concrete (UHPFRC):
Design, Assessment and Modeling
Rui Yu

nr 210
Uncertainty in Modeling Activity-Travel Demand in Complex Uban Systems
Soora Rasouli

nr 211
Simulation-based Performance Assessment of Climate Adaptive Greenhouse Shells
Chul-sung Lee

nr 212
Green Cities:
Modelling the Spatial Transformation of the Urban Environment using Renewable Energy Technologies
Saleh Mohammadi

nr 213
A Bounded Rationality Model of Short and Long-Term Dynamics of Activity-Travel Behavior
Ifigeneia Psarra

nr 214
Effects of Pricing Strategies on Dynamic Repertoires of Activity-Travel Behaviour
Elaheh Khademi

nr 215
Handstorm Principles for Creative and Collaborative Working
Frans van Gassel

nr 216
Light Conditions in Nursing Homes: Visual Comfort and Visual Functioning of Residents
Marianne M. Sinoo

nr 217
Woonsporen: De Sociale en Ruimtelijke Biografie van een Stedelijk Bouwblok in de Amsterdamse Transvaalbuurt
Hüseyin Hüsnü Yegenoglu

nr 218
Studies on User Control in Ambient Intelligent Systems
Berent Willem Meerbeek

nr 219
Daily Livings in a Smart Home: Users' Living Preference Modeling of Smart Homes
Erfaneh Allameh

nr 220
Smart Home Design: Spatial Preference Modeling of Smart Homes
Mohammadali Heidari Jozam

nr 221
Wonen: Discoursen, Praktijken, Perspectieven
Jos Smeets

nr 222
Personal Control over Indoor Climate in Offices: Impact on Comfort, Health and Productivity
Atze Christiaan Boerstra

nr 223
Personalized Route Finding in Multimodal Transportation Networks
Jianwe Zhang

nr 224
The Design of an Adaptive Healing Room for Stroke Patients
Elke Daemen

nr 225
Experimental and Numerical Analysis of Climate Change Induced Risks to Historic Buildings and Collections
Zara Huijbregts

nr 226
Wind Flow Modeling in Urban Areas Through Experimental and Numerical Techniques
Alessio Ricci

nr 227
Clever Climate Control for Culture: Energy Efficient Indoor Climate Control Strategies for Museums Respecting Collection Preservation and Thermal Comfort of Visitors
Rick Kramer

nr 228
Fatigue Life Estimation of Metal Structures Based on Damage Modeling
Sarmediran Silitonga

nr 229
A multi-agents and occupancy based strategy for energy management and process control on the room-level
Timilehin Moses Labeodan

nr 230
Environmental assessment of Building Integrated Photovoltaics: Numerical and Experimental Carrying Capacity Based Approach
Michiel Ritzen

nr 231
Performance of Admixture and Secondary Minerals in Alkali Activated Concrete: Sustaining a Concrete Future
Arno Keulen

nr 232
World Heritage Cities and Sustainable Urban Development: Bridging Global and Local Levels in Monitoring the Sustainable Urban Development of World Heritage Cities
Paloma C. Guzman Molina

nr 233
Stage Acoustics and Sound Exposure in Performance and Rehearsal Spaces for Orchestras: Methods for Physical Measurements
Remy Wenmaekers

nr 234
Municipal Solid Waste Incineration (MSWI) Bottom Ash: From Waste to Value Characterization, Treatments and Application
Pei Tang

nr 235
Large Eddy Simulations Applied to Wind Loading and Pollutant Dispersion
Mattia Ricci

nr 236
Alkali Activated Slag-Fly Ash Binders: Design, Modeling and Application
Xu Gao

nr 237
Sodium Carbonate Activated Slag: Reaction Analysis, Microstructural Modification & Engineering Application
Bo Yuan

nr 238
Shopping Behavior in Malls
Widiyani

nr 239
Smart Grid-Building Energy Interactions: Demand Side Power Flexibility in Office Buildings
Kennedy Otieno Aduda

nr 240
Modeling Taxis Dynamic Behavior in Uncertain Urban Environments
Zheng Zhong

nr 241
Gap-Theoretical Analyses of Residential Satisfaction and Intention to Move
Wen Jiang

nr 242
Travel Satisfaction and Subjective Well-Being: A Behavioral Modeling Perspective
Yanan Gao

nr 243
Building Energy Modelling to Support the Commissioning of Holistic Data Centre Operation
Vojtech Zavrel

nr 244
Regret-Based Travel Behavior Modeling: An Extended Framework
Sunghoon Jang

nr 245
Towards Robust Low-Energy Houses: A Computational Approach for Performance Robustness Assessment using Scenario Analysis
Rajesh Reddy Kotireddy

nr 246
Development of sustainable and functionalized inorganic binder-biofiber composites
Guillaume Doudart de la Grée

nr 247
A Multiscale Analysis of the Urban Heat Island Effect: From City Averaged Temperatures to the Energy Demand of Individual Buildings
Yasin Toparlar

nr 248
Design Method for Adaptive Daylight Systems for buildings covered by large (span) roofs
Florian Heinzelmann

nr 249
Hardening, high-temperature resistance and acid resistance of one-part geopolymers
Patrick Sturm

nr 250
Effects of the built environment on dynamic repertoires of activity-travel behaviour
Aida Pontes de Aquino

nr 251
Modeling for auralization of urban environments: Incorporation of directivity in sound propagation and analysis of a framework for auralizing a car pass-by
Fotis Georgiou

nr 252
Wind Loads on Heliostats and Photovoltaic Trackers
Andreas Pfahl

nr 253
Approaches for computational performance optimization of innovative adaptive façade concepts
Roel Loonen

nr 254
Multi-scale FEM-DEM Model for Granular Materials: Micro-scale boundary conditions, Statics, and Dynamics
Jiadun Liu

nr 255
Bending Moment - Shear Force Interaction of Rolled I-Shaped Steel Sections
Rianne Willie Adriana Dekker

nr 256
Paralympic tandem cycling and handcycling: Computational and wind tunnel analysis of aerodynamic performance
Paul Fionn Mannion

nr 257
Experimental characterization and numerical modelling of 3D printed concrete: Controlling structural behaviour in the fresh and hardened state
Robert Johannes Maria Wolfs

nr 258
Requirement checking in the building industry: Enabling modularized and extensible requirement checking systems based on semantic web technologies
Chi Zhang

nr 259
A Sustainable Industrial Site Redevelopment Planning Support System
Tong Wang

nr 260
Efficient storage and retrieval of detailed building models: Multi-disciplinary and long-term use of geometric and semantic construction information
Thomas Ferdinand Krijnen

nr 261
The users' value of business center concepts for knowledge sharing and networking behavior within and between organizations
Minou Weijs-Perrée

nr 262
Characterization and improvement of aerodynamic performance of vertical axis wind turbines using computational fluid dynamics (CFD)
Abdolrahim Rezaeiha

nr 263
In-situ characterization of the acoustic impedance of vegetated roofs
Chang Liu

nr 264
Occupancy-based lighting control: Developing an energy saving strategy that ensures office workers' comfort
Christel de Bakker

nr 265
Stakeholders-Oriented Spatial Decision Support System
Cahyono Susetyo

nr 266
Climate-induced damage in oak museum objects
Rianne Aleida Luimes

nr 267
Towards individual thermal comfort: Model predictive personalized control of heating systems
Katarina Katic

nr 268
Modelling and Measuring Quality of Urban Life: Housing, Neighborhood, Transport and Job
Lida Aminian

nr 269
Optimization of an aquifer thermal energy storage system through integrated modelling of aquifer, HVAC systems and building
Basar Bozkaya

nr 270
Numerical modeling for urban sound propagation: developments in wave-based and energy-based methods
Raúl Pagán Muñoz

nr 271
Lighting in multi-user office environments: improving employee wellbeing through personal control
Sanae van der Vleuten-Chraibi

nr 272
A strategy for fit-for-purpose occupant behavior modelling in building energy and comfort performance simulation
Isabella I. Gaetani dell'Aquila d'Aragona

nr 273
Een architectuurhistorische waardestelling van naoorlogse woonwijken in Nederland: Het voorbeeld van de Westelijke Tuinsteden in Amsterdam
Eleonore Henriette Marie Mens

nr 274
Job-Housing Co-Dependent Mobility Decisions in Life Trajectories
Jia Guo

nr 275
A user-oriented focus to create healthcare facilities: decision making on strategic values
Emilia Rosalia Catharina Maria Huisman

nr 276
Dynamics of plane impinging jets at moderate Reynolds numbers – with applications to air curtains
Adelya Khayrullina

nr 277
Valorization of Municipal Solid Waste Incineration Bottom Ash - Chemical Nature, Leachability and Treatments of Hazardous Elements
Qadeer Alam

nr 278
Treatments and valorization of MSWI bottom ash - application in cement-based materials
Veronica Caprai

nr 279
Personal lighting conditions of office workers - input for intelligent systems to optimize subjective alertness
Juliëtte van Duijnhoven

nr 280
Social influence effects in tourism travel: air trip itinerary and destination choices
Xiaofeng Pan

nr 281
Advancing Post-War Housing: Integrating Heritage Impact, Environmental Impact, Hygrothermal Risk and Costs in Renovation Design Decisions
Lisanne Claartje Havinga

nr 282
Impact resistant ultra-high performance fibre reinforced concrete: materials, components and properties
Peipeng Li

nr 283
Demand-driven Science Parks: The Perceived Benefits and Trade-offs of Tenant Firms with regard to Science Park Attributes
Wei Keat Benny Ng

nr 284
Raise the lantern; how light can help to maintain a healthy and safe hospital environment focusing on nurses
Maria Petronella Johanna Aarts

nr 285
Modelling Learning and Dynamic Route and Parking Choice Behaviour under Uncertainty
Elaine Cristina Schneider de Carvalho

nr 286
Identifying indoor local microclimates for safekeeping of cultural heritage
Karin Kompatscher

nr 287
Probabilistic modeling of fatigue resistance for welded and riveted bridge details. Resistance models and estimation of uncertainty.
Davide Leonetti

nr 288
Performance of Layered UHPFRC under Static and Dynamic Loads: Effects of steel fibers, coarse aggregates and layered structures
Yangyueye Cao

nr 289
Photocatalytic abatement of the nitrogen oxide pollution: synthesis, application and long-term evaluation of titania-silica composites
Yuri Hendrix

nr 290
Assessing knowledge adoption in post-disaster reconstruction: Understanding the impact of hazard-resistant construction knowledge on reconstruction processes of self-recovering communities in Nepal and the Philippines
Eefje Hendriks

nr 291
Locating electric vehicle charging stations: A multi-agent based dynamic simulation
Seheon Kim

nr 292
De invloed van Lean Management op de beheersing van het bouwproces
Wim van den Bouwhuijsen

nr 293
Neighborhood Environment and Physical Activity of Older Adults
Zhengying Liu

nr 294
Practical and continuous luminance distribution measurements for lighting quality
Thijs Willem Kruisselbrink

nr 295
Auditory Distraction in Open-Plan Study Environments in Higher Education
Pieternella Elizabeth Braat-Eggen

nr 296
Exploring the effect of the sound environment on nurses' task performance: an applied approach focusing on prospective memory
Jikke Reinten

nr 297
Design and performance of water resistant cementitious materials– Mechanisms, evaluation and applications
Zhengyao Qu

nr 298
Design Optimization of Seasonal Thermal Energy Storage Integrated District Heating and Cooling System: A Modeling and Simulation Approach
Luyi Xu

nr 299
Land use and transport: Integrated approaches for planning and management
Zhongqi Wang

nr 300
Multi-disciplinary optimization of building spatial designs: co-evolutionary design process simulations, evolutionary algorithms, hybrid approaches
Sjonnie Boonstra

nr 301
Modeling the spatial and temporal relation between urban land use, temperature, and energy demand
Hung-Chu Chen

nr 302
Seismic retrofitting of masonry walls with flexible deep mounted CFRP strips
Ömer Serhat Türkmen

nr 303
Coupled Aerostructural Shape and Topology Optimization of Horizontal-Axis Wind Turbine Rotor Blades
Zhijun Wang

nr 304
Valorization of Recycled Waste Glass and Converter Steel Slag as Ingredients for Building Materials: Hydration and Carbonation Studies
Gang Liu

nr 305
Low-Carbon City Development based on Land Use Planning
Gengzhe Wang

nr 306
Sustainable energy transition scenario analysis for buildings and neighborhoods - Data driven optimization
Shalika Saubhagya Wickramarachchi Walker

nr 307
In-between living and manufactured: an exploratory study on biobuilding components for building design
Berrak Kirbas Akyurek

nr 308
Development of alternative cementitious binders and functionalized materials: design, performance and durability
Anna Monika Kaja

nr 309
Development a morphological approach for interactive kinetic façade design: Improving multiple occupants' visual comfort
Seyed Morteza Hosseini

nr 310
PV in urban context: modeling and simulation strategies for analyzing the performance of shaded PV systems
Ádám Bognár

nr 311
Life Trajectory, Household Car Ownership Dynamics and Home Renewable Energy Equipment Adoption
Gaofeng Gu

nr 312
Impact of Street-Scale Built Environment on Walking/Cycling around Metro Stations
Yanan Liu

nr 313
Advances in Urban Traffic Network Equilibrium Models and Algorithms
Dong Wang

nr 314
Development of an uncertainty analysis framework for model-based consequential life cycle assessment: application to activity-based modelling and life cycle assessment of multimodal mobility
Paul Martin Baustert

nr 315
Variable stiffness and damping structural joints for semi-active vibration control
Qinyu Wang

nr 316
Understanding Carsharing-Facilitating Neighborhood Preferences
Juan Wang

nr 317
Dynamic alignment of Corporate Real Estate to business strategies: An empirical analysis using historical data and in-depth modelling of decision making
Howard Cooke

nr 318
Local People Matter: Towards participatory governance of cultural heritage in China
Ji Li

nr 319
Walkability and Walkable Healthy Neighborhoods
Bojing Liao

nr 320
Light directionality in design of healthy offices: exploration of two methods
Parisa Khademagha

nr 321
Room acoustic modeling with the time-domain discontinuous Galerkin method
Huiqing Wang

nr 322
Sustainable insulating lightweight materials for enhancing indoor building performance: miscanthus, aerogel and nano-silica
Yuxuan Chen

nr 323
Computational analysis of the impact of façade geometrical details on wind flow and pollutant dispersion
Xing Zheng

nr 324
Analysis of urban wind energy potential around high-rise buildings in close proximity using computational fluid dynamics
Yu-Hsuan Jang

nr 325
A new approach to automated energy performance and fault detection and diagnosis of HVAC systems: Development of the 4S3F method
Arie Taal

nr 326
Innovative Admixtures for Modifying Viscosity and Volume Change of Cement Composites
Hossein Karimi

nr 327
Towards houses with low grid dependency: A simulation-based design optimization approach
Zahra Mohammadi

nr 328
Activation of demand flexibility for heating systems in buildings: Real-life demonstration of optimal control for power-to-heat and thermal energy storage
Christian Finck

nr 329
A computational framework for analysis and optimisation of automated solar shading systems
Samuel B. de Vries

nr 330
Challenges and potential solutions for cultural heritage adaptive reuse: a comparative study employing the Historic Urban Landscape approach
Nadia Pintossi

nr 331
Shared control in office lighting systems
Tatiana Aleksandrovna Lashina

nr 332
Comfort in Urban Public Spaces
You Peng

nr 333
Numerical modelling of metal soap formation in historical oil paintings
Gerardus Johannes Anna Maria Eumelen

nr 334
A transdisciplinary decision-making approach to food-water-energy nexus: A guide towards sustainable development
Maryam Ghodsvali

nr 335
Numerical modelling of transient low-frequency sound propagation and vibration in buildings
Indra Sihar

nr 336
Characterization of impact sound from lightweight joist floors
Yi Qin

nr 337
Cities for Children: Supporting Children and Caregivers in Participatory Urban Planning
Özlemnur Ataol

nr 338
Engaging the unengaged: Exploring citizen participation in nature-based solutions in China
Li Dai

nr 339
Municipal Solid Waste Incineration Residues: analysis, treatments, and applications
Ekaterina Loginova

nr 340
Enhancing the Uptake of Nature-Based Solutions in Urban Settings: An Information Systems Approach
Shahryar Ershad Sarabi

nr 341
Work Schedule Arrangements in Two-Adult Households with Children
Bilin Han

nr 342
Increasing awareness of urban cultural heritage using digital technologies: empirical design and analysis of a new multi-media web platform
Benshuo Wang

nr 343
Mechanical and physical properties of fibre-cement composites using alternative natural fibres
Katerina Kochova

nr 344
Numerical and experimental investigation of urban microclimate in a real compact heterogeneous urban area
Nestoras Antoniou

nr 345
Examining in-class activities to facilitate academic achievement in higher education: A framework for optimal indoor environmental conditions
Henk W. Brink

nr 346
High-temperature resistant geopolymers: composition, microstructure and performance
Kinga Malgorzata Klima

nr 347
Individual and household decision-making in shared parking
Qianqian Yan

nr 348
In-situ formation of LDHs in Alkali activated binders
Tao Liu

nr 349
Condition assessment of concrete sewer pipes through an integrated experimental-numerical approach
Irene C. Scheperboer

nr 350
In situ PU-based characterization of sound absorbing materials for room acoustic modeling purposes
Baltazar Briere de La Hosseraye

nr 351
Uncertainty analysis and management in building energy data mining: A bottom-up approach considering the temporal and spatial aspect of data
Waqas Khan

nr 352
Personalized Heating Control Systems to improve thermal comfort and reduce energy consumption
Michal Veselý

nr 353
Restorative value of the urban greenscape: Urban residential streets as restorative environments
Robert P. van Dongen

nr 354
Urban ventilation and the compact Mediterranean city: numerical investigations of the dynamic relationships between density, morphology and wind flow
Olga Palusci