Universitat de Lleida

Statistical learning methods for energy assessment in buildings with applications at different geographic levels

Gerard Mor Martínez

Universitat de Lleida

PhD Thesis

Statistical learning methods for energy assessment in buildings with applications at different geographic levels

Gerard Mor Martínez

Thesis presented for the degree of Doctor by Universitat de Lleida
PhD program in Engineering and Information Technologies

Thesis supervisor
Daniel Chemisana Villegas

Academic supervisor
Daniel Chemisana Villegas

2021

Essentially, all models are wrong, but some are useful.

George E. P. Box

Acknowledgements

Over the last six years, I have been working as a researcher at CIMNE simultaneously with the development of this Thesis. In this time, I have been helped and met many people to whom I wish to express my gratitude.

First of all, I want to dedicate this Thesis to my parents, Josep and Maite, and my wife, Vanesa. Without you, it would have been impossible to get this far. Thank you for giving me always the best of you. I hope to return it in the same way. Indeed, I want to extend my gratitude to all my grandparents, aunts, uncles, cousins and in-law family. Especially to Lluïsa, who has always been like a second mother to me. Moreover, I want to express my gratefulness to my closest friends to accompany me during this process and pushing me to reach my full potential.

Secondly, I want to thank all my colleagues at CIMNE, BEE Group and BEE Data (Jordi and Xavi Cipriano, Jordi Carbonell, Stoyan Danov, Daniel Pérez, Benedetto Grillone, ...). This thesis would not have been finished without our teamwork and good fellowship. Special thanks to my boss, whom I can consider more like a friend, Dr Jordi Cipriano, who proposed me start this path at the end of 2015 and always gave me the necessary motivation, help, and kindness to do so. Furthermore, it is worth mentioning the good friendship we have in our Lleida's office. It has been an enormous pleasure to share my day-to-day with Jordi, Jorge, Chiara, Jose Santos, Meredith, Josep, Eloi, Jaume, Florencia, Gerard, Francesc, Edgar and Joel. You helped me more than you might think.

Thank you to my director Dr Daniel Chemisana, for your patience, guidance and help since the beginning of this 6-years path. I also want to thank the motivation that other PhD candidates gave me during this period (Laia, Gerard, Alex, Alberto, Jordi, ...). Besides, I extend my appreciation to Dr Francesc Solsona for the critical encouragement he gave me in several moments.

Last but not least, I would like to express my gratitude to my colleagues at JRC (Hans, Giacomo, Lorena and Francesco) and DTU (Peder and Henrik). Thank you for the stay abroad periods I did throughout my PhD. I have learned a lot from you and your teammates, both from data science and cultural point of view. It has been a pleasure to collaborate with you.

To all of you, I express not only my gratitude but also my admiration. This Thesis would not have an endpoint without you.

Summary

The building sector, excluding its industry, is one of the world's largest energy consumers. 2019 accounted for around 30% of the total final energy consumed worldwide. In addition, its CO_2 emissions accounted for 28% of the total, as much of the fuel used to generate this final energy is still of non-renewable origin.

Currently, there is an extreme need to reduce these pollutant emissions over the next few years due to the global warming problems we are experiencing. In addition, the peak of fossil fuel production is either near or has already been exceeded during the last decade. This will lead to the end of affordable fossil fuels. Therefore, the world must move towards an energy strategy aimed at increasing demand-side efficiency and consuming energy produced from renewable fuels. To this end, implementing mathematical models to help characterise, simulate and predict energy consumption in the building sector is a key step in this energy transition process.

Within the framework of this Thesis, a platform for storing and massively analysing energy data has been implemented. Additionally, three more specific use cases have been proposed that refer to some of the most recurrent problems at each of the main geographical levels in the building sector (dwelling, building or district level). The objectives of these use cases are to inform and alert end-users about their energy consumption, optimising energy demand or cost, maximising energy consumption from renewable generation, or inferring apparently unknown energy characteristics of buildings and their occupants.

This Thesis presents the data analytics platform designed and developed to deal with the massive analysis of a vast amount of data coming from electricity smart meters. Furthermore, the implemented energy information services for end-users are presented, and the estimated energy savings generated by those services, quantified within the IEE Empowering project, are presented (3 to 22%).

Subsequently, three applications are introduced, each one dealing with a specific geographical level. In the first one, a novel methodology to virtually replicate the control of thermostatically-controlled systems is presented. It is applied over a set of residential dwellings and it is based on data-driven models. Some promising outcomes showed during warm conditions (7-15°C), for example, reducing the usual set-point temperature of the thermostat by 1°C or 2°C would lead to energy savings of 18.1% and 36.5% on average, respectively.

In the second application, three Model Predictive Control (MPC) strategies have been implemented in different locations in Europe to assess the energy flexibility that can be achieved when a smarter control is applied to existing electricity driven heating or cooling systems in several building typologies and electricity markets. The results showed that electric heat pumps can provide significant demand response flexibility in the respective analysed electricity markets. However, they sometimes have problems regarding response time and reliability, which can affect their availability for the standby electricity market.

Finally, in the third and last case study, a methodology for characterising the electricity consumption of large sets of buildings, e.g. entire districts or postal codes, is presented. The methodology is based on statistical analysis of the aggregated hourly energy consumption of the whole area of interest, as well as its correlation against meteorological information, cadastral data and socio-economic characteristics. This methodology has been validated to interpret the main drivers of electricity consumption along the whole province of Lleida (Spain).

Resumen

El sector de la edificación, sin incluir la industria, es uno de los principales focos de consumo energético del mundo. Supone alrededor de un 30% del total de energía final consumida mundialmente. Además, sus emisiones de CO_2 suponen un 28% respecto al total, ya que todavía buena parte del combustible utilizado para generar esta energía final es de origen no renovable.

Actualmente, existe la extrema necesidad de reducir estas emisiones contaminantes durante los siguientes años debido a los problemas de calentamiento global que estamos viviendo. Además, el pico de producción de los combustibles fósiles, o es cercano o ya lo hemos sobrepasado durante la última década. Este hecho conllevará el fin de los combustibles fósiles a precio asequible. Por lo tanto, el mundo debe dirigirse hacia una estrategia energética encaminada a incrementar la eficiencia en la demanda y a consumir energía producida mediante combustibles renovables. Con este fin, la implementación de modelos matemáticos que ayuden a caracterizar, simular y a predecir el consumo energético en el sector de la edificación supone un paso clave en este proceso de transición energética.

En el marco de esta Tesis se ha implementado una plataforma para almacenar y analizar masivamente datos energéticos, y se han planteado tres casos de uso más concretos que hacen referencia a algunas de las problemáticas más recurrentes en cada uno de los principales niveles geográficos en el sector edificación (nivel vivienda, edificio, o distrito). Los objetivos de estas analíticas son informar y alertar a usuarios finales sobre su consumo energético, optimizar la demanda o el coste energético, maximizar el consumo procedente de producción renovable, o inferir características energéticas aparentemente desconocidas.

Inicialmente, esta Tesis presenta la plataforma de analítica diseñada para el análisis masivo de contadores inteligentes de electricidad. Aparte, se detallan los servicios de información energética para usuarios finales implementados, y se presentan los resultados de ahorro estimado producido (3% a 22%) a lo largo del proyecto IEE Empowering para tres comercializadoras de electricidad.

Posteriormente, se presentan tres aplicaciones específicas tratando distintos niveles de agregación. En la primera de ellas, se presenta una metodología novedosa para replicar virtualmente el control de los sistemas comandados por termostato en el sector residencial utilizando modelos basados en datos. Los resultados de esta investigación muestran que se puede conseguir un ahorro energético del 18,1% y del

36,5% de media, si se reduce la temperatura de consigna habitual en 1°C y 2°C, respectivamente.

En la segunda aplicación se han implementado tres estrategias de Control Predictivo mediante Modelos (MPC, en inglés) en tres lugares distintos de Europa, con el objetivo de evaluar la flexibilidad energética que puede lograrse cuando se aplica un control más inteligente a sistemas de calefacción eléctricos existentes en un edificio o un conjunto muy pequeño de edificios. Los resultados del método muestran que las bombas de calor tienen el potencial de proporcionar una importante flexibilidad de respuesta a la demanda en los países analizados. Sin embargo, en ocasiones tienen problemas en cuanto a su tiempo de respuesta y fiabilidad, lo que puede afectar a su disponibilidad para el mercado de reserva de electricidad.

En la tercera y última aplicación, se presenta una metodología de caracterización del consumo eléctrico sobre grandes conjuntos de edificios, por ejemplo distritos enteros o códigos postales. Se basa en el análisis estadístico de los consumos energéticos horarios agregados a cada una de las áreas de interés, y su correlación con la información meteorológica, catastral y las características socioeconómicas. Este método se ha validado para interpretar los factores de cambio en el consumo eléctrico de la provincia de Lleida (España).

Resum

El sector de l'edificació, sense incloure la seva indústria, és un dels principals focus de consum energètic del món. Suposa al voltant d'un 30% del total d'energia final consumida mundialment. A més, les seves emissions de CO_2 suposen un 28% respecte al total, ja que encara bona part del combustible utilitzat per a generar aquesta energia final és d'origen no renovable.

Actualment, existeix l'extrema necessitat de reduir aquestes emissions contaminants durant els propers anys a causa dels problemes d'escalfament global que estem vivint. A més, el pic de producció dels combustibles fòssils, o és pròxim o ja l'hem sobrepassat durant l'última dècada. Aquest fet comportarà la fi dels combustibles fòssils a preu assequible. Per tant, mundialment ens hem de dirigir cap a una estratègia energètica encaminada a incrementar l'eficiència en la demanda i a consumir energia produïda mitjançant combustibles renovables. A aquest efecte, la implementació de models matemàtics que ajudin a caracteritzar, simular i a predir el consum energètic en el sector de l'edificació suposa un pas clau en aquest procés de transició energètica.

En el marc d'aquesta Tesi s'ha implementat una plataforma per emmagatzemar i analitzar massivament dades energètiques, i s'han plantejat tres casos d'ús més concrets que fan referència a algunes de les problemàtiques més recurrents en cadascun dels principals nivells geogràfics en el sector edificació (nivell habitatge, edifici, o districte). Els objectius d'aquestes analítiques són informar i alertar a usuaris finals sobre el seu consum, optimitzar la demanda o el cost energètic, maximitzar el consum procedent de producció energètica renovable, o inferir característiques energètiques.

Primerament, aquesta Tesi presenta la plataforma d'analítica dissenyada per a l'anàlisi massiva de comptadors intel·ligents d'electricitat. A part, es detallen els serveis d'informació energètica per a usuaris finals que s'han implementat, i es presenten els resultats d'estalvi estimat produït (del 3% a 22%) al llarg d'un projecte amb tres comercialitzadores d'electricitat europees.

Posteriorment, es presenten les tres aplicacions específiques tractant diferents nivells geogràfics. En la primera d'elles, es presenta una novedosa metodologia per tal de replicar virtualment el control dels sistemes comandats per termòstat en el sector residencial utilitzant models basats en dades. Els resultats d'aquesta recerca mostren que es pot aconseguir un estalvi energètic del 18,1% i del 36,5% de mitjana, si es redueix la temperatura de consigna habitual en 1°C i 2°C, respectivament.

En la segona aplicació, tres estratègies de Control Predictiu mitjançant Models (MPC, en anglès) s'han implementat en tres llocs diferents d'Europa, amb l'objectiu d'avaluar la flexibilitat energètica que pot aconseguir-se quan s'aplica un control més intel·ligent a sistemes de calefacció existents d'un edifici o d'un conjunt molt petit d'edificis. Els resultats del mètode mostren que les bombes de calor tenen el potencial de proporcionar una important flexibilitat de resposta a la demanda als països analitzats. No obstant això, a vegades tenen problemes quant al seu temps de resposta i fiabilitat, la qual cosa pot afectar la seva disponibilitat per al mercat de reserva d'electricitat.

En la tercera i última aplicació, es presenta una metodologia de caracterització del consum elèctric de grans conjunts d'edificis, per exemple districtes sencers o codis postals. Es basa en l'anàlisi estadística dels consums energètics horaris agregats a les diferents arees d'interès, i la seva correlació respecte informació meteorològica, cadastral o característiques socioeconòmiques. Aquest mètode s'ha validat per a interpretar els factors de canvi en el consum elèctric de la província de Lleida (Espanya).

Table of contents

Nomenclature

Acronyms

aFRR	automatic Frequency Restoration Reserve Market
AMI	Advanced Metering Infrastructure
ANN	Artificial Neural Network
AR	AutoRegressive
ARMAX	AutoRegressive Moving Average with eXogenous
ARX	AutoRegressive with eXogenous
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
BaU	Business as Usual
BES	Building Energy Simulation
CM	Cluster Manager
CVRMSE	Coefficient of Variation of the Root Mean Squared Error
DA	Day-Ahead Electricity Price
DER	Distributed Energy Resource
DHW	Domestic Hot Water

DR	Demand Response
DSO	Distribution System Operator
ECM	Energy Conservation Measures
EU	European Union
FCR	Frequency Containment Reserve
FF	Flexibility Function
HP	Heat Pump
HVAC	Heating, ventilation and air conditioning
ICT	Information and Communication Technology
IoT	Internet of Things
LPF	Low-Pass Filter function
MAPE	Mean Absolute Percentage Error
mFRR	Manual Frequency Restoration Reserve
MPC	Model Predictive Control
NEMO	Nominated Electricity Market Operator
NLME	Non Linear Mixed Effect
RMSE	Root Mean Squared Error
RR	Reserve Replacement
SAR	Seasonal Auto Regressive
SH	Space Heating

SVM	Support Vector Machines
TS	Time Series
TSO	Transmission System Operator

Subscripts and superscripts

b	baseline
bd	building number
e	active
f	trace to be tracked
fs	fourier series components
i	indoor
lp	low-pass filtered
o	outdoor
opt	optimized
s	set-point
s, sim	simulated set-point
t	time t

Variables

Φ^h	Heat consumption
A	Percentage of activation time within a time step
B	Backward shift operator

i	flexibility evaluation period
I^{sol}	Solar irradiance
n	number of time steps
P	Power
RC	Resistor and Capacitator
S^{az}	Solar azimuth
S^{el}	Solar elevation (Solar zenith - 90°)
T	Temperature
W^{d}	Wind direction
W^{s}	Wind speed

Chapter 1

Introduction

1.1 Actual and future perspectives of energy consumption in buildings

It has been observed that the worldwide final energy consumption of the building sector remained at the same level in 2019 compared to the previous year, which supposed around 130 EJ of final energy consumption or 30% of total share [1] (see Fig. 1.1). Moreover, if the buildings construction industry is included, this share increases up to 35%. Thus, although the build-up and the population have been increasing for the last years, the final energy consumption share of the building sector has remained stable for the first time since 2012. This is caused by an improvement in energy use intensity indicators and the switch from traditional biomass, oil, and coal to electricity and gas, which are assumed to be more efficient (see Fig. 1.2). Additionally, since 2018, renewable energy use (including modern biomass) has grown by around 6%, representing a strong shift from previous years and marking a return to fast growth similar to the one made in 2013.

However, looking at the fuel type share within the building sector, in 2019, the use of fossil fuels in buildings remained highly significant, summing up to around 38% (natural gas, coal and oil) without accounting for the fossil primary sources used for power production in the case of electricity and commercial heat (see Fig. 1.3). Therefore, it becomes crucial that the renewable energy share, which is presently around 5.9%, should continue the fast growth of the last years to accomplish the buildings decarbonisation.

As shown in Figure 1.4, the amount of CO_2 generated by the building's operation has reached its highest level ever, around 10 Gt CO_2 in 2019 [2]. This

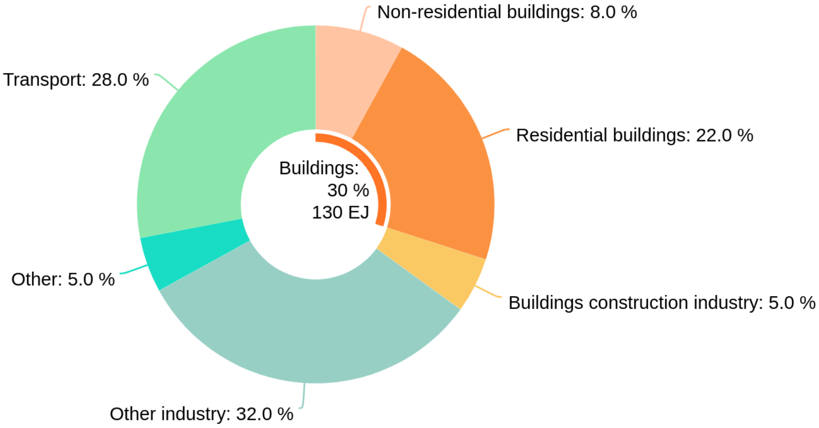


Figure 1.1: Global share of energy consumption by sectors, 2019.

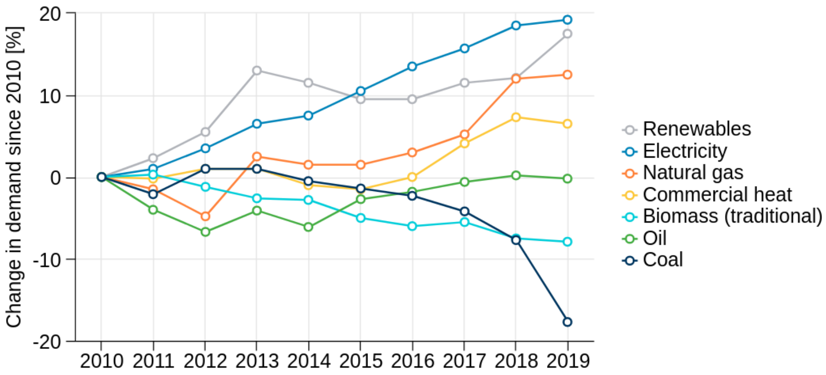


Figure 1.2: Evolution in global buildings sector final energy use by fuel type.

supposes 28% of global energy-related CO_2 emissions. Moreover, if the building construction industry is included, the share of CO_2 emissions increases to 38%, higher than the combined shares of all the other industries (32%) or the transport sector (23%) [1]. Finally, the split made between direct emissions from the building operation itself and indirect emissions from power generation for electricity and commercial heat is notable.

As a result of continued use of coal, oil, and natural gas for heating and cooking, as well as higher activity levels in regions with carbon-intensive power lines, buildings' CO_2 emissions have been increasing in absolute terms, resulting in steady levels of direct emissions but growing indirect emissions (i.e. electricity). Nowadays, around 55% of global electricity consumption is associated with buildings

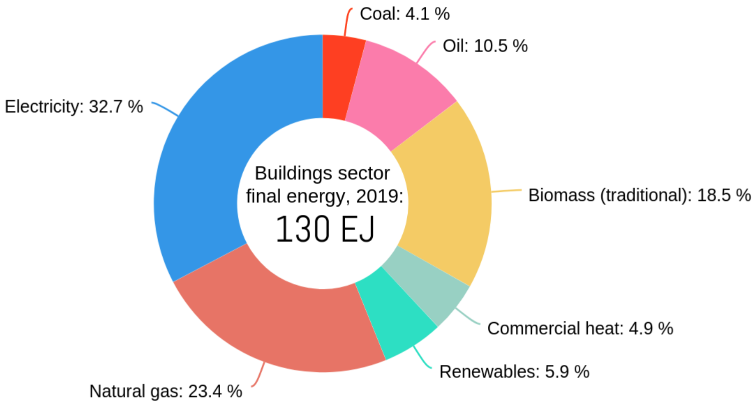


Figure 1.3: Global buildings sector energy use by fuel type, 2019

operation [2]. Nonetheless, a slight decrease of CO_2 emissions in the buildings' sector is appreciated compared to the 39% achieved in 2018. This is due to an increase in emissions from transport and other industries.[1]

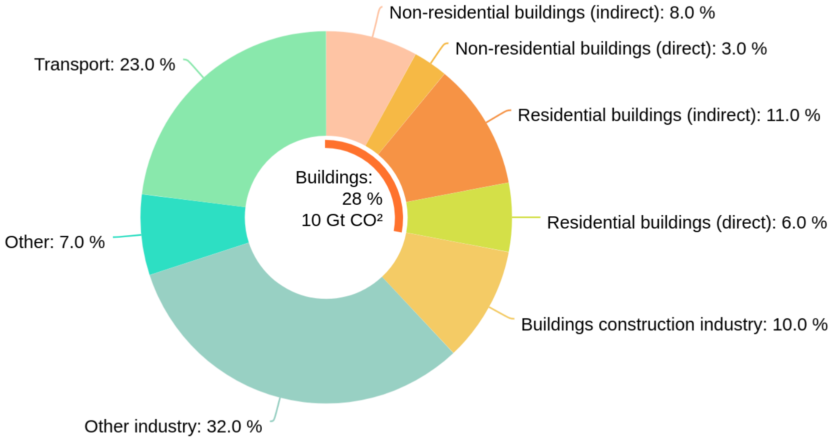


Figure 1.4: Global share of CO_2 emissions by sectors, 2019.

According to IEA [3], the vast majority of CO_2 emissions due to global energy consumption are due to fossil sources (see Fig. 1.5). Since 2015, the main growth factor of CO_2 emissions is the natural gas, principally because the increment of emissions due to coal, and oil with less emphasis, has been significantly lower in this period. As analysed by [4], the major cause was that the peak oil had been reached during the last years. The same study predicted that the natural gas peak of production will be achieved in 2030, at most.

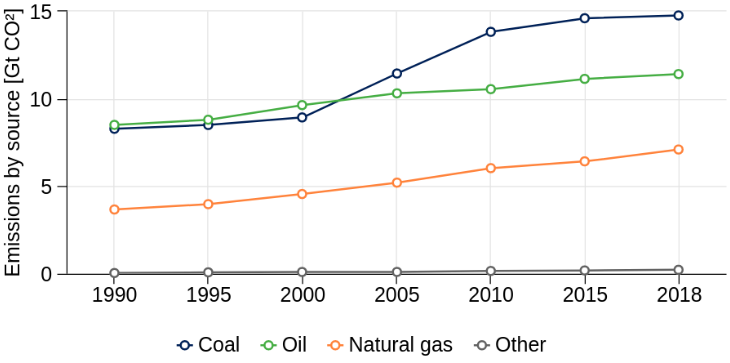


Figure 1.5: Sources of CO_2 emissions, 1990-2018.

Generally speaking, the peak of production of fossil fuels is not mainly caused by the nonexistence of new fields, but by the very limited, even negative, Energy Return on Investment (EROI) of these new resources. In addition, the associated harmful emissions that they generate [5] [6] are considered as the second main cause.

Although the current demand for fossil fuels as primary energy is enormous, it would remain very significant considering governmental policies and the current demand characteristics of the main sectors (transport, industry and buildings). The so-called Stated Policies Scenario (STEPS) [7] proposes the least change concerning the current status. Basically, it assumes an increase of the energy demand and fulfils this demand with the extra supplies coming from new fields and the availability of disruptive future technologies, making it technically and economically feasible to exploit these new reserves.

In the mid-long term, multiple sustainable scenarios need to be created. This is mainly due to the urgent need to decrease the use of traditional fuels in favour of renewable resources that do not generate polluting emissions and considering the more realistic decline in the production of fossil fuels. The IEA calls them Sustainable Development scenarios [8] and Net-zero 2050 scenario [9]. Both of them propose a significant reduction in fossil fuel demand, except for the low-carbon fossil fuels (e.g. biodiesel, bioethanol, compressed natural gas), which brings both predictions (supply and demand) closer together (see Fig. 1.6). Especially interesting is the Net-zero 2050 scenario, which is largely based on renewables, with solar being the largest source of supply. In this scenario, businesses, investors and citizens should cooperate closely with countries to develop sustainable economies

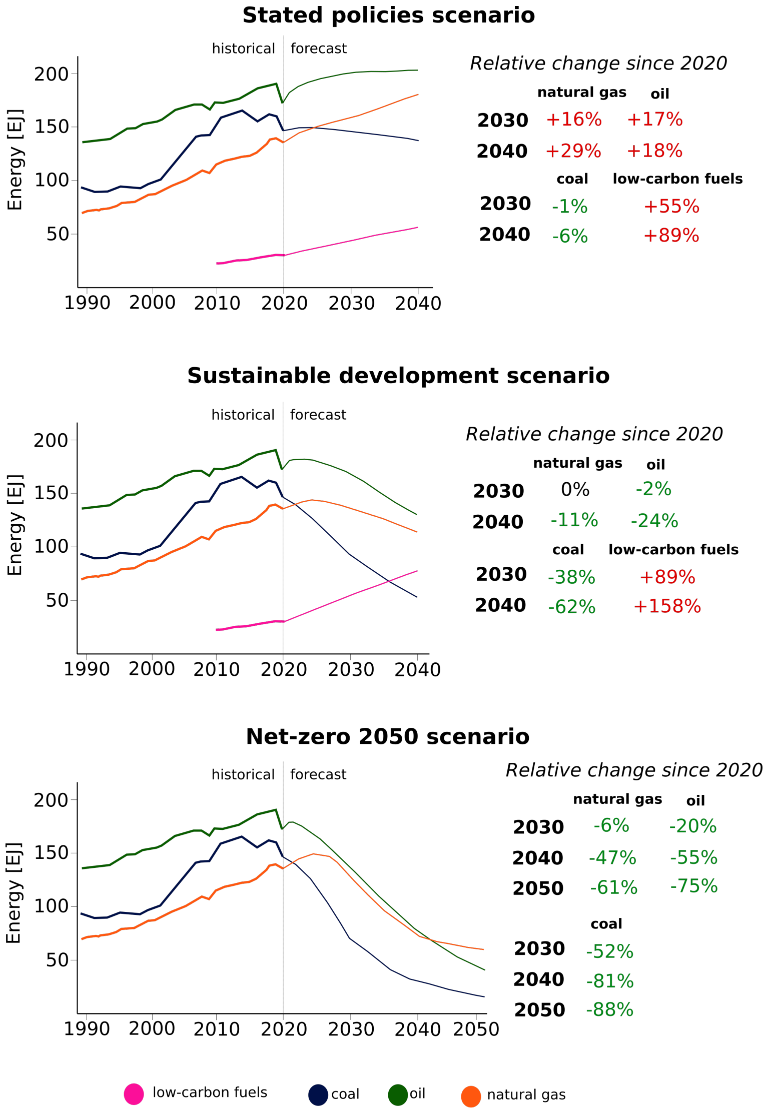


Figure 1.6: Historical and forecast demand of the main fossil fuel sources depending on future scenarios

that have the financing and technologies available to reach net-zero emissions in time.

Figure 1.7 depicts the oil supply forecasts made by IEA [3] and the demand scenarios already shown in Figure 1.6. It can be seen that the Net-zero 2050 scenario is the only one where the future is not overly committed to the discovery of new fields and future technologies that improve the EROI indicator.

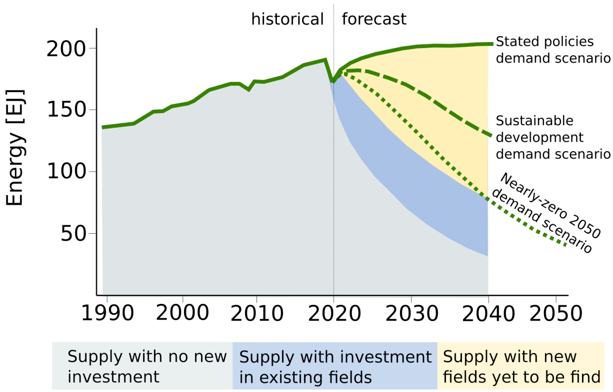


Figure 1.7: Historical and forecasted oil supply vs oil demand depending on future scenarios

Returning to the buildings sector, the Buildings Climate Tracker (BCT), promoted by the Global Alliance for Buildings and Construction, tracks the buildings sector progress in decarbonisation worldwide [10]. This index combines data from seven global indicators to demonstrate progress since 2015 in an action and impact index. The objective is to reach an index of 100 in 2050, which means an average improvement of 2.6 per year. It includes incremental energy efficiency investment in buildings and Nationally Determined Contributions (NDCs) with actions taken in the building sector. The contributions of the indicators are weighted individually to ensure they address the tracker objective (decarbonisation index) adequately and do not over- or under-represent certain aspects when aggregated. The indicators of the analysis could be categorized into two groups: impact and action. The former contains the results of the actions that determine CO_2 emissions, final energy demand, the share of renewable energy sources in buildings, or the energy use intensities. The latter represents those indicators related to initiatives that aim to reduce CO_2 emissions, such as environmental policies, green building certifications, energy-efficiency efforts, industry actions, ...

The last update of the BCT [1] finds that annual decarbonisation progress has slowed and almost halved since 2016 (see Fig. 1.8). Although the number of actions to reduce CO_2 emissions in the building sector is growing, the rate of improvement year over year is declining. Buildings sector participants across the value chain need to make joint efforts to reverse this trend to reach net-zero carbon by 2050 and increase energy efficiency and decarbonisation actions by a factor of five.

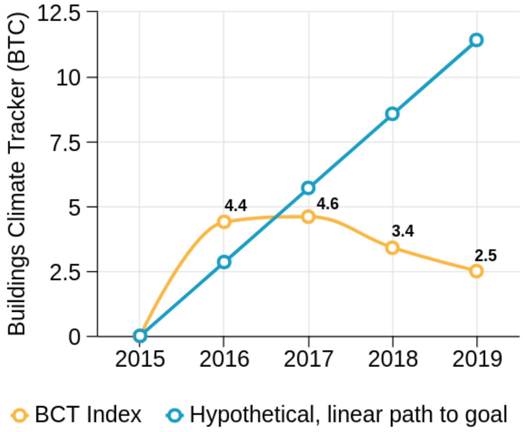


Figure 1.8: Buildings Climate Tracker, 2015-2019.

Moreover, by 2050, the European Union (EU) is committed to becoming the first climate-neutral continent. The European countries pledge to reduce greenhouse gas emissions by at least 55% by 2030, compared to 1990 levels [11]. Additionally, the EU agreed in 2018 to adopt the amending Directive on Energy Efficiency (2018/2002), providing policy frameworks for 2030 and beyond, in conjunction with the "Clean Energy for All Europe package". This amendment includes an ambitious headline energy efficiency target of 32.5% for 2030. To reach the goal, the EU must achieve the projections that were made in 2007 for 2030. Accordingly, EU energy consumption must not exceed 1128 Mtoe of primary energy and/or 846 Mtoe of final energy (following the withdrawal of the United Kingdom)[12].

In sum, a triple strategy is essential to reduce energy demand and emissions in the building sector. First, decarbonising the electricity power sector while implementing materials strategies that decrease lifecycle carbon emissions strengthen the case for reducing energy demand and emissions. According to the IEA, direct building CO_2 emissions will have to decline by 50% by 2050, while indirect building sector emissions will have to decline through a 60% reduction in electricity power generation emissions by 2030 if a zero-carbon building stock has to be achieved by 2050 [1]. Efforts like these would need to reduce around 6% of emissions per year from 2020 to 2030. As a comparison, according to IEA predictions [3], global CO_2 emissions dropped by 7% during the COVID-19 epidemic.

Therefore, based on the current status and keeping in mind the objectives presented above, a huge improvement is needed during the upcoming years. Solving this problem will require applying multiple approaches, from demand reduction to energy generation based on renewable resources. Furthermore, it is well known that the most cost-effective and environmentally friendly energy is the one that is not consumed. Therefore, increasing energy efficiency is one of the key aspects to be addressed in the current and future energy transition.

To this end, applying statistics and machine learning techniques to energy-related problems can help achieve this important transformation to boost energy efficiency and meet future decarbonisation targets, both at the global and EU level. For instance, plenty of data-driven methodologies have demonstrated they can assess and predict the energy performance of buildings, provide optimised controls for Heating, Ventilation and Air Conditioning (HVAC) systems, or increase users' awareness towards buildings' energy consumption.

1.1.1 From dwelling to district level

In the framework of this Thesis, multiple geographical levels related to buildings are considered. In the subsequent paragraphs, these levels are explained to align the meaning of these concepts between the author and the readers.

The formal definition of a dwelling is a place where someone lives. According to the EPBD (Directive 2010/31/EU), the closest definition to dwelling refers to a building unit, which means a section, floor or apartment within a building designed or altered to be used separately. Besides, the formal definition of a building is a structure with a roof and walls, for example, a house or a factory. The same EPBD (Directive 2010/31/EU) defines a building as a roofed construction having walls, for which energy is used to condition the indoor climate. At this point, the energy consumption made by the users to fulfil their necessities (e.g. cooking, cleaning, lighting, entertainment,...) could be added.

Depending on the building typology and the specific characteristics, a single building could be constituted by one or multiple dwellings or building parts. Furthermore, each of these parts could be related to singular energy consumption characteristics due to differences in user behavioural patterns, occupancy, domestic appliances or building characteristics.

Finally, the formal definition of a district is an area of a town or a city that has been given official boundaries for administration. Regarding the energy context of this Thesis, the definition of district refers to a geographical zone in a city or a region with a particular characteristic or condition. For instance, in district heating installations, the whole group of buildings supplied by the system forms a district because the heating system is the particular characteristic in common. Or, in the case of electricity consumption, all the consumers who are located in a close area given a certain economic sector (residential, industrial or services) and tariff.

1.2 Applied statistical learning techniques to buildings energy-related data

Statistical learning uses a wide range of tools to understand data. There are two categories of tools: supervised and unsupervised. The broad definition of supervised statistical learning is the construction of a statistical model to predict or estimate output based on one or more inputs. This problem is prevalent in various fields, including business, medicine, astronomy and public policy. As opposite, unsupervised statistical learning involves an output without supervision. However, it is still possible to learn relationships and structure from these data [13].

More precisely, statistical learning is a field between mathematics and machine learning, composed of a toolbox for modelling and understanding complex data to obtain statistical models for prediction and characterisation purposes. Compared to machine learning models, the training process usually needs major human interaction during the definition of the models and the transformation of the inputs to boost and optimise them. However, statistical learning techniques tend to need less data to be reliable, as certain relationships between variables are considered priorly during the definition of the models. Several statistical learning tools are nowadays applied to buildings' energy consumption data. They range from unsupervised techniques, such as clustering of consumers, to supervised techniques, like regression or classification models.

In this Thesis, multiple types of statistical learning techniques are used to provide a solution to each of the applications described, mainly concerning forecast, characterisation, and simulation scenarios. Thus, this document offers a powerful example that data-driven methodologies should be used from now on as a method to massively model complex phenomena related to energy consumption in the buildings sector, based on data gathered from IoT and metering devices.

1.3 Outline of the thesis

This Thesis applies several statistical learning techniques to real building-energy-related datasets obtained from Internet of Things (IoT) monitoring devices, Advanced Metering Infrastructure (AMI), or online services. The main objective is to validate that these techniques are suitable and convenient for assessing energy performance across multiple geographical levels within this sector.

First, this document describes the design and implementation of an ICT infrastructure for the statistical analysis of high-frequency data gathered and communicated by electricity smart meters, focusing on user awareness applications for domestic electricity consumption. Then, this IT platform is used to implement three different case studies involving analysis at multiple geographical levels (see Fig. 1.9). The commonality between these chapters is the usage of data-driven models to characterise, predict or optimise energy consumption in buildings.

1.3.1 IT infrastructure

The amount of information available and suitable for the energy assessment of buildings is increasing year by year. However, the heterogeneity of these data increases in parallel, which makes its appropriate usage more difficult. Nowadays, the information gathered from buildings monitoring systems contains a wide combination of consumption data, sensors data, or information from controllers or IoT devices. Therefore, this data should be stored in a platform where it can be managed and analysed adequately.

In chapter 2, the results achieved during the EMPOWERING project are presented. From 2013 to 2017, a Big Data platform for the assessment of energy-related data was developed. Initially, it aimed at helping domestic customers to save electricity by managing their consumption positively. This is achieved by improving the information received about energy bills and offering online tools to the end-users. The main contributions of EMPOWERING were creating a novel workflow in the electric utility sector regarding the implementation of data analytics for their customers and the fast implementation of data-mining techniques over massive data sets within a Big Data platform. The results obtained showed that EMPOWERING can be used for customers of electricity suppliers to change

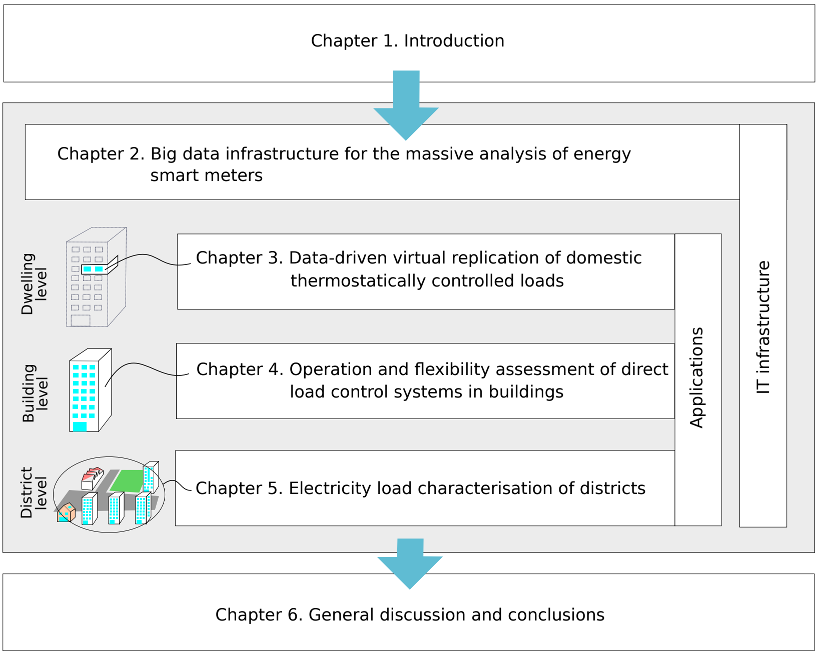

Chapter 1. Introduction
Chapter 2. Big data infrastructure for the massive analysis of energy smart meters
Dwelling level
Chapter 3. Data-driven virtual replication of domestic thermostatically controlled loads
Building level
Chapter 4. Operation and flexibility assessment of direct load control systems in buildings
District level
Chapter 5. Electricity load characterisation of districts
Applications
IT infrastructure
Chapter 6. General discussion and conclusions

Figure 1.9: Flow chart of the Thesis chapters

their energy habits to decrease energy consumption and increase environmental sustainability.

1.3.2 Application at dwelling level

Chapter 3 presents a data-driven method to model the energy performance of thermostatically-controlled heating systems. These systems are widely spread in the residential sector since they control the heat consumption provided for domestic hot water and space heating. Therefore, assessing the energy performance at the thermostat level and the effect of different control strategies requires simplified modelling techniques demanding few inputs and low computational resources. Data-driven techniques are envisaged as one of the best options to meet these constraints.

This chapter presents a novel methodology consisting of an optimization algorithm, two auto-regressive models and a control loop algorithm able to virtually replicate the control of thermostatically driven systems. This combined strategy includes all the modes governed by the setpoint temperature and enables automatic assessment of the energy consumption impact of multiple scenarios. The required inputs are limited to available historical readings from smart thermostats and external climate data sources. The methodology has been trained and validated with datasets coming from 11 smart thermostats connected to gas boilers and placed in several households in northeastern Spain. Important conclusions of the research are that these techniques can estimate the temperature decay of households when the space heating is off, and the energy consumption needed to reach the comfort conditions. Furthermore, this research shows that energy savings of 18.1% and 36.5% can be achieved on average if the usual setpoint temperature schedule is lowered by 1°C and 2°C, respectively.

1.3.3 Application at the building level

In chapter 4, three different MPC strategies were implemented in three different European locations to evaluate the energy flexibility achievable when a smarter control is applied to legacy HVAC systems at the building level. To date, the assessment of the energy flexibility to be delivered by existing buildings and by their legacy HVAC systems is hindered by a lack of commonly agreed-upon methodologies. There are many research works in the field; however, many of them are focused on the design stage or, in case of addressing building operation, they are based on controlled experimental set-ups.

The novelty of this chapter lies in the fact that it develops and validates an original methodology for the Flexibility Function estimation to evaluate the delivered energy flexibility of several Automated Demand Response services applied on different heat pump systems working under real operations. Furthermore, the active interaction with several electricity markets, ranging from the Spanish day-ahead market to the German and Swiss ancillary services markets, have also been evaluated during the winter and spring seasons. The method results showed that heat pumps could offer a significant potential of flexibility in the analysed countries. Nevertheless, it has also been envisaged that some restrictions concerning reaction times and reliability may affect its readiness for certain ancillary services markets.

1.3.4 Application at the district level

In chapter 5, a bottom-up electricity load characterisation methodology of the building stock is presented. It is based on the statistical analysis of aggregated hourly energy consumption, weather data, and cadastral and socioeconomic information. To demonstrate the validity of this methodology, the whole province of Lleida, located in northeast Spain, was used as a case study. The geographical aggregation level considered is the postal code level since it is the highest data resolution available through the open data sources used in the research work. The development and the experimental tests are supported by a web application environment formed by interactive user interfaces specifically developed for this purpose.

The major novelty of this chapter relies on the application of statistical data methods able to infer the energy performance characteristics by principal components without prior knowledge of its specific building characteristics. First, a multi-step data-driven technique is used to disaggregate the electricity consumption in multiple uses (space heating, cooling, holidays and baseload). Afterwards, multiple Key Performance Indicators (KPIs) are derived from this decomposition to obtain the energy characterisation over a certain area. The potential reuse of this methodology allows for a better understanding of the drivers of electricity use, with multiple applications for the public and private sectors.

1.3.5 Projects and publications related

This Thesis has been elaborated in the framework of multiple projects and publications developed in CIMNE - BEE Group from 2016 to 2021, collaborating with other researchers, companies and entities.

Chapter 2: Big data infrastructure for the massive analysis of energy smart meters

The development and implementation of the IT infrastructure were mainly supported by the Intelligent Energy for Europe (IEE) programme in a project called EMPOWERING and partially by the Ministerio de Economía y Competitividad under contract TIN2017-84553-C2-2-R, and the European Union FEDER (CAPAP-H6 network TIN2016-81840-REDT).

Additionally, this chapter was published as a journal article: Mor, G.; Vilaplana, J.; Danov, S.; Cipriano, J.; Solsona, F.; Chemisana, D. EMPOWERING, a smart Big Data framework for sustainable electricity suppliers, IEEE Access 2018, https://doi.org/10.1109/ACCESS.2018.2881413

Lastly, the open-source code of the ENMA architecture, which is the last production version of the data analytics platform presented in this chapter, is accessible in https://github.com/BeeGroup-cimne/ENMA.

Chapter 3: Data-driven virtual replication of domestic thermostatically controlled loads

The methodology and case study to simulate thermostatically controlled domestic systems using data-driven models was funded by the project COMRDI15-1-0036, so-called REFER, within the RIS3CAT community of the Catalan government.

In addition, this research was published in:

Mor, G.; Cipriano, J.; Gabaldon, E.; Grillone, B.; Tur, M.; Chemisana, D. Data-Driven Virtual Replication of Thermostatically Controlled Domestic Heating Systems. Energies 2021, 14, 5430. https://doi.org/10.3390/en14175430

Chapter 4: Operation and flexibility assessment of direct load control systems in buildings

The work regarding the assessment of flexibility demand in buildings emanated from collaborative research conducted with the financial support of the European Commission through the H2020 project Sim4Blocks, grant agreement 695965.

Furthermore, this chapter is already a published paper:

Mor, G.; Cipriano, J.; Grillone, B.; Amblard, F.; Menon, R.P.; Page, J.; Brennenstuhl, M.; Pietruschka, D.; Baumer, R.; Eicker, U. Operation and energy flexibility evaluation of direct load controlled buildings equipped with heat pumps, Energy and Buildings 2021, https://doi.org/10.1016/j.enbuild.2021.111484

Chapter 5: Electricity load characterization of districts

The work related to data-driven energy characterisation of districts was executed in several projects and contracts. Mainly, the work emanated from the collaboration between JRC and CIMNE, through the Energy & Location Applications of the ELISE (European Location Interoperability Solutions for e-Government) action of the ISA^2 (Interoperability solutions for public administrations, businesses and citizens) programme. In particular, this collaboration has been materialised through a personal JRC Expert Contract, grant agreement CT-EX2017D306558-102. Additionally, the European Commission has also been financing this research through the BIGG H2020 project, grant agreement 957047.

This research has been accepted as a journal paper in:

Mor, G.; Cipriano, J.; Martirano, G.; Pignatelli, F.; Lodi, C.; Lazzari, F.; Grillone, B.; Chemisana, D. A data-driven method for unsupervised electricity consumption characterisation at the district level and beyond, Energy Reports 2021. https://doi.org/10.1016/j.egyr.2021.08.195b

Chapter 2

Big data infrastructure for the massive analysis of energy smart meters

This chapter has been published as a paper:

Mor, G.; Vilaplana, J.; Danov, S.; Cipriano, J.; Solsona, F.; Chemisana, D. EMPOWERING, a smart Big Data framework for sustainable electricity suppliers, IEEE Access 2018, https://doi.org/10.1109/ACCESS.2018.2881413

2.1 Introduction

The built environment sector is becoming the leading consumer of energy in the world, accounting for 40% of global energy use and one third of overall greenhouse gas emissions [14]. Within the built environment, in 2015, residential energy consumption amounted to around 25.4% of total final energy use in the European Union [15]. Therefore, to achieve the European 2020 targets, changes in the consumption patterns of EU households are urgent and necessary. To mitigate the energy and environmental pressures caused by household energy use, substantial research and development efforts have been made into energy-efficient technologies [16]. In recent years, improving energy efficiency and reducing energy demand have been widely regarded as the most promising, fastest, cheapest and safest ways to mitigate environmental pressures and climate change [17]. As a result, heating and cooling systems now use less energy than ever. However, final energy consumption has not decreased as expected. On the contrary, energy consumption has tended to increase. An analysis carried out within the EU-funded ODYSEE and MURE projects [18] quantified the increase in the energy efficiency of domestic appliances in Europe over the 2000 to 2012 period at 21% while the increase in final energy consumption was 75 Mtoe for the same period. One reason appears to be that much technology is made available to the public without adequate instruction and support. Although technological advances are significant for promoting energy conservation and improving energy efficiency [19], it is increasingly recognized that behavioral factors are of greater significance for energy conservation [20]. It has been suggested that behavioral changes can be just as effective as technological changes [21]. In [18], it was stated that changes in heating behavior had an impact on energy consumption by reducing it by 20 Mtoe over the over the period from 2000 to 2012. Since 2008, the level of this behavioural effect has doubled to 2.6 Mtoe/year, compared with 1.2 Mtoe before. Effective long-term strategies should engage people directly in efforts to reduce their energy consumption. This should be achieved through the implementation of environmental policies aiming at changing energy use behavior, as highlighted in [22]. Acknowledging people as an active element in the energy system should lead to efforts to better understand how people interact with energy and to stimulate the development of Energy Awareness services that attempt to change how and when people use energy.

Regarding the change of the energy behavior of consumers, in recent decades, many psychological models have been developed and adopted to explore how householders consume energy and the factors that influence this [23]. Different types of intervention strategies have been developed with the aim of stimulating changes in people's energy use behavior and thus achieving energy savings [24].

The overall aim of the EMPOWERING project is to empower consumers by involving, informing and helping them to take measures to save energy on the basis of the information they receive from their utility company. More specifically, the consumers' aim consists of achieving measurable energy savings.

The main contribution of EMPOWERING consists of a novel dataflow procedure for electric utility companies to standardize data communication, cleaning, storage and analysis. This workflow is based on secure API REST [25] communication, a set of ETL (Extract, Transform and Load) modules to clean and store the data in the EMPOWERING databases and a set of analytical modules to infer information from the energy consumption. EMPOWERING analyses data across the database of clients by making unsupervised learning searches and inferring clusters of similar types of domestic customers according to different information fields by means of data-mining techniques. This procedure can account for similarity between neighborhoods, size of building, number of occupants, climatic zone, etc. It provides a means to make comparisons of energy consumption with similar customers, namely between members of the same cluster. EMPOWERING offers specific, personalized, targeted information about whether one's consumption is above or below a cluster average over a season. This can show a need for space heating systems to be checked, or the building envelope to be improved. The large amount of data handled cannot be processed efficiently using traditional databases. These are the foundations of the smart Big Data framework developed within the EMPOWERING project.

The EMPOWERING services can deal with different data granularity, from monthly-based data coming from standard meters, to hourly-based data from smart meters. However, notable benefits are reached when hourly metering is used. For instance, alarms can be set up that detect abnormally high consumption levels for base-load appliances such as refrigerators or freezers. Some of these possibilities have already been developed within the EMPOWERING project with the collaboration of four electric utility companies in Europe, but the potential is far from the mainstream. The EMPOWERING project aims to accelerate the transition of the use of this type of service from pioneering companies to mainstream best practice.

2.2 Related Work

Many data-mining techniques have been used to predict electricity consumption [26, 27]. These include neural networks (NN) [28], support vector machines (SVM) [29], support vector regression (SVR) [30], decision trees [31], auto regressive integrated moving average (ARIMA) models [32], clustering models [33], decomposition models, grey box models [34], and regression models [35]. The authors in [36] noted that NN and SVR have been used extensively for forecasting residential electricity consumption. In [30], the authors considered NN and SVR suitable for predicting industrial energy demand. They concluded that the two models have advantages and disadvantages and that it is inconclusive which is the best for energy forecasting. In [31], the performance of regression analysis models, decision trees, and NN for energy forecasting were compared. In the winter period, NN performed slightly better, whereas in the summer period, the decision tree model performed somewhat better than the other two. The authors in [37] presented a multidimensional hybrid architecture to make energy consumption predictions based on energy data-mining techniques that additionally makes use of current energy data enriched by external unstructured Big Data information. Predictive data-mining has been also applied to the building operation stage to predict its overall energy consumption [27]. Data-mining can be also used to obtain deeper insights into the data, to try to discover associations, correlations, and intrinsic data structures in Big-data. This is called descriptive data-mining. Compared with predictive data-mining, the descriptive version is more flexible in application, as it does not involve a training process and the knowledge of the discovery process is not guided by predefined targets. Descriptive data-mining has mainly been applied at the building operation stage for fault detection and diagnostics [38, 39]. Popular techniques include association rule mining, anomaly detection and clustering analysis.

Quilumba et al. [40] proposed a combination of predictive and descriptive data-mining procedures, recognizing the importance of differences in energy consumption patterns. They proposed a prediction approach based on clustering customers according to their consumption behavior and then predicting the energy consumption of the whole population by aggregating the forecasting of each single cluster. They applied this strategy to predict electricity consumption and demand for event-organising venues in the residential and commercial sectors. Clustering has also been used in the literature to group energy consumers with similar characteristics [41, 42] and to detect atypical, usually undesired, user behavior [43, 44].

The results of the studies [45, 46] show that a combination of statistical analysis with prediction models (holistic, simulation and inverse models), complemented in some cases with monitoring data analysis, can be a powerful tool for developing urban energy action aimed at reducing the energy consumption not only of existing buildings but also in higher geographical areas, such as neighborhoods or districts.

Big Data technology gives insights into how we think about a certain topic [47]. Big Data tools can manage structured, unstructured and semi-structured data [48]. Various data-acquisition Internet-of-Things (IoT) devices are penetrating into the wider world and are able to collect information spanning different areas [49]. The estimated installed base of smart meters worldwide will surpass 1.1 billion by 2022 [50], and will collect electricity usage data in the range of 15 minutes each. This is up to a three thousand-fold increase in the amount of data utilities processed in the past. It means that by 2022 the electric utility industry will be swamped by more than 2 petabytes of data annually from smart meters alone. Cisco [51] estimated that the data generated by devices would reach 507.5 zettabytes (ZB) per year (42.3 ZB per month) by 2019. This immense growth of data cannot be processed efficiently using relational databases.

2.3 The EMPOWERING platform

2.3.1 Architecture

Fig. 2.1 shows the general architecture of the EMPOWERING system, which is designed to tackle the following IT challenges: (i) to provide a means to link the local utility database to the Big Data analysis environment, (ii) to offer high quality in the delivered services, (iii) to provide batch-processing data analytics services and (iv) to ensure data privacy and security.

The Big Data architecture developed within EMPOWERING is a Representational State Transfer (REST) framework which provides an Engine with a technology aware interface.

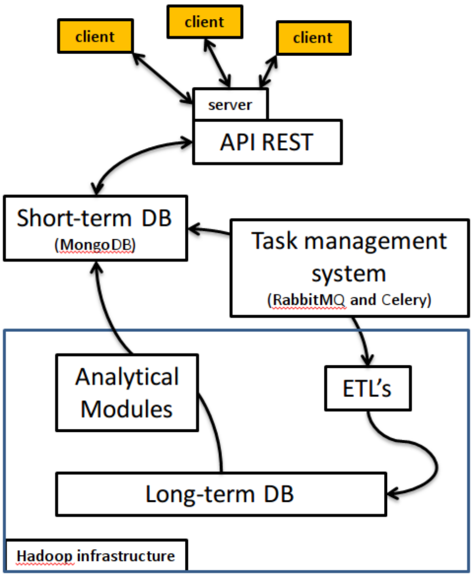


Figure 2.1: EMPOWERING Architecture.

A REST style architecture conventionally consist of a client-server paradigm. REST's client-server separation simplifies component implementation and allows intermediary modules, like proxies, gateways, caching systems and firewalls to be inserted into middle levels without changing the interface between the main components.

This architecture allows the storage and wrangling of large amounts of data. This is made up by a combination of low-cost hardware and database technologies that allows the acquisition, allocation and extraction of data to be processed in a distributed cluster. Essentially, the storage is split into **Short-term** and **Long-term** databases (DB), which have different characteristics according to the quantity, type and usage of data stored in them.

The EMPOWERING Big Data framework is entirely developed using open-source software. It is mainly composed of 3 components: *API REST*, *Task Management System* and the *Hadoop infrastructure.*

API REST

This is the communication interface between the server and the Client REST, and thus also with the utilities. The Application Program Interface (API) is fully developed following the REST standard. This component is the utilities' gateway to communicate and configure the Engine. The aim is to enable a Service-Oriented Architecture (SOA), offering specialized energy services to the customer and the utility system administrators. This is not a simple issue. The main objectives of the API are (i) to set and configure the services (see section 2.3.3), (ii) to import data into the Engine and (iii) to export data from the Engine. These objectives are addressed using different technologies. Data import and export are enabled using the *Eve* framework to implement the Web service. *MongoDB* is the technology used for the short-term DB. It is buffer storage for data reception and sending in fast environments. It is the data storage directly connected to the API and provides high communication bandwidth. It supplies temporary storage, acting as a cache memory, prior to permanent storage in the long-term database. *ExtJS* technology was used to implement User Interface (UI) for setting and configuring the services. *OpenAM* provides open source Authentication, Authorization, Entitlement and Federation software. The *Flask* and *Python* modules implement all the server functionalities in order to deploy a web API server. Flask allows customizable, fully featured REST Web Services to be built and deployed effortlessly, which greatly simplifies the configuration of the API.

Task Management System

This level is in charge of scheduling and synchronizing the tasks in the engine by means of *RabbitMQ* and *Celery*. In essence, the scheduler picks up the new task to be executed in EMPOWERING according to a scheduling policy. The FIFO policy was chosen because the batch operation of the tasks made other variants (like Round Robin), frequently applied in time-sharing environments, inefficient. Celery is the scheduler itself. RabbitMQ is a fast internal message-queuing system used to interchange information between tasks with different paradigm technologies.

Hadoop infrastructure

Apache Hadoop is an open-source framework that provides tools for distributed storage and processing. It allows organizations to process and analyze large volumes of unstructured and semi-structured data, heretofore inaccessible, in a cost- and time-effective way.

Apache Ambari is used in order to manage the Hadoop cluster. It allows nodes to be added and removed, new components to be installed in existing working nodes, the cluster monitored, etc.

The two main Hadoop components are YARN and HDFS:

- *HDFS (Hadoop Distributed File System)* consists of slave components called DataNodes where data is physically saved and a master process called NameNode that is responsible for mantaining the file system directory tree and has the information of where data effectively is (i.e. which blocks are available in every DataNode). All HDFS reads and writes are managed by DataNode.
- *YARN (Yet Another Resource Negotiator)* is responsible for processing Map-Reduce tasks using the master-slave paradigm. It consists of the Resource-Manager (master similar to NameNode). It is in charge of managing the launched tasks. The NodeManager resides in the slave nodes. It receives Map or Reduce orders from the ResourceManager and executes those tasks in YARN containers.

There are many high level applications running on the main components. Two of them were used in this project:

- *Hbase*: Distributed key-value database. Provides real-time read/write access and is built on top of HDFS. Hbase is used as the long-term big-data DB. It is formed by hundreds of thousands of AMI (Advanced Metering Infrastructure) devices used in EMPOWERING.

- *Hive*: Data warehouse on top of HDFS which provides SQL-like querying which are translated into MapReduce functions.

The YARN component is recursively used when Extract, Transforms and Load (ETL), and analytical modules are running. Initially, multiple asynchronous ETL functionalities aggregate, clean and transfer the data from the short-term to the long-term DB. These functions pre-process the input data to ensure the quality and format of the long-term DB.

Once the information is stored in the long-term DB, asynchronous analytical modules are implemented to generate the needed results for the services offered to the utility. The technologies used for the algorithms are a combination of *R*, *Hive*, and *Python software libraries* using the Map Reduce [52] paradigm to allow complex calculations over large sets of data. R is an open-source programming language for statistical computing. In order to use R in the Hadoop environment, the Rhipe and Rhadoop packages were used. These packages offer access to the long-term DB and facilitate the implementation of Map Reduce algorithms using common R functionalities. Python can also be used in the same manner with the MRjob, Happybase and Snakebite libraries. Python scientific libraries, such as Pandas, SciPy or NumPy, enable other advanced means for data analysis as an alternative to R. Hive is a data warehouse system for Hadoop. It provides functionality for data summarization, querying, and analysis of data. Hive queries are written in HiveQL, an SQL-like language.

The combination of these languages allows the use of the most highly optimized implementations according to the requirements of the algorithm and this generates less development effort and a shorter data processing time when the code is executed.

2.3.2 Data

EMPOWERING services mainly rely on three categories of data: (1) energy consumption and contract, (2) end-user's and (3) third-party data.

- **Energy consumption and contract data** is the information used for billing (e.g. consumption data, contract details). This encompasses consumption data, either read at a low frequency manually, or by analogue meters, or estimated (quarterly, bi-annually, annually, etc.), as well as fine consumption data from smart meters (sub-hourly, hourly, daily, monthly, etc.). A certain type of consumption data may require clients' consent to collect or display. Thus, this type of data may not be available for all customers.

- **End-user data** is not directly accessible by the customers because it does not serve for billing purposes. It is usually collected via online forms or surveys. Services relying on this type of data depend on the willingness of customers to fill in information about their dwellings and equipment. It can be erroneous or incoherent, so services based on this information have to consider data inaccuracies.

- **Third-party data** these data is obtained from remote databases or provided by third parties and do not concern the user directly. This can be meteorological, statistical, etc.

EMPOWERING was conceived as a Big Data ICT architecture because of the large amount of data to be managed. More specifically, in the first services implemented from 2013 to 2016, the EMPOWERING architecture was managing 3 years of historical data from 70,000 contracts with the end users of two European electricity trading companies on an hourly basis and 30,000 contracts on a monthly basis, altogether corresponding to 1,831 million measurements of electricity consumption.

2.3.3 Services

This section describes the EMPOWERING services. These constitute the main outputs of the analytic modules and are delivered to the final user in multiple formats and timescales (i.e. web, paper reports).

In addition to the usual services currently provided by electric utilities, such as consumption billing or historical monthly consumption, others seek to increase the benefits and volume of useful information that reaches the end users. These innovative services focus on the following topics: weather-normalized consumption comparisons compared with similar consumers, personalized energy-saving tips, tariff comparisons, consumption prediction and consumption alerts. These are the most relevant services currently developed in EMPOWERING. Most of them are based on one or multiple data-mining techniques to detect the weather-dependent share of consumption, clustering similar neighbors or forecasting the energy consumption.

Weather-dependence analysis

This can be understood as a pre-treatment service. It is widely used in many services, e.g. normalized benchmarks, clustering of similar neighbors and consumption prediction or alerts. It estimates customers' energy consumption with respect to the weather at their locations. These services use several linear-regression techniques to correlate the energy consumption for space heating or cooling with the outdoor temperature. Households with strong weather dependence are associated with higher consumption levels in winter and higher outdoor temperatures in summer. Fig. 2.2 depicts the information provided to the customers: monthly consumption and the average monthly evolution of temperature over the preceding 12-month period. An explanatory text is attached so that the customers can understand the correlations between their energy consumption and the outdoor temperature better. In this case, it seems that electricity consumption is not weather dependent. Thus, this customer's consumption was similar throughout the year.

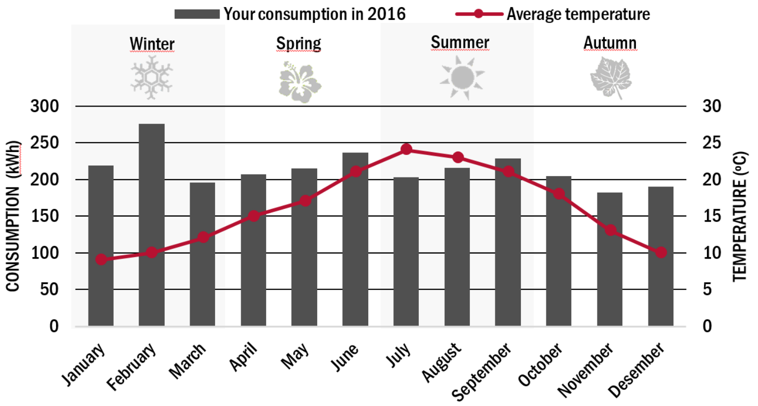


Figure 2.2: Monthly consumption and average temperature over the last 12 months.

The models used to determine weather dependence differ depending on the data granularity of the consumption data:

- **Monthly data:** A linear regression is used to fit the monthly energy consumption based on heating or cooling degree days. The regression during heating and cooling periods are expressed in equation 2.1.

$$E_t = \alpha + H_c * CD_t + H_h * HD_t + \varepsilon_t, \tag{2.1}$$

where E_t is the electricity consumption at month t (Wh), α is the estimated baseload consumption (Wh), H_h is the estimated Heat Transfer Coefficient of the dwelling (not considering the performance of the systems) (Wh/K), H_c is the estimated Cool Transfer Coefficient of the dwelling during the cooling period (not considering the performance of the systems) (Wh/K), HD_t and CD_t are respectively the heating and cooling days in month t (K), and $\varepsilon_t s \sim N(0, \sigma)$ and i.i.d.

The model parameters are estimated using the least-squares minimization approach. Customers with an estimated Heat Transfer Coefficient (HTC) higher than 100 Wh/K are assumed to have weather dependence during the cooling or heating periods. The adjusted coefficients α and HTC are also used to weather-normalize the monthly consumption when this information is used to compare historical consumption. In this case, the HD_t and CD_t considered correspond to the values of the last 12 months.

- **Hourly data:** Three-parameter (3P) and five-parameter (5P) models are used to estimate the relation between the daily aggregated electricity consumption and the average daily outdoor temperature. The 5P model is appropriate for modeling energy consumption data that include both heating and cooling, e.g. dwellings with a heat pump installed. 3P models are appropriate for modeling the electricity use in residences with a weather dependence during one of the periods (cooling or heating), e.g. dwellings with an electric chiller or boiler installed. The 5P model is presented in Equation 2.2.

$$E_t = \alpha_s + H_{c,s} * (T_t - T_c)^+ + H_{h,s} * (T_t - T_h)^- + \varepsilon_t, \qquad (2.2)$$

where E_t is the electricity consumption at day t in Wh, s corresponds to a certain daily load curve pattern (each day is clustered in a specific pattern), T_t is the average daily temperature at day t (K), T_c is the cooling change point temperature (K), T_h is the heating change point temperature (K). These temperatures are optimised for each customer using a BFGS method considering the Root Mean Square error (RMSE) between the predicted and actual energy consumption as the cost function. The α_s, $H_{h,s}$, and $H_{c,s}$ are adjusted to each daily load curve pattern detected, to characterise better the different cooling or heating dependencies along weekdays and weekends, holidays, or specific parts of the year related with extremely particular intradaily seasonality.

For the time-dependent consumption comparison modules, a weather normalization analysis is applied. It consists of estimating the actual consumption by considering the ratios between the HD_t or CD_t from the previous and current periods. This estimation allows the comparison of the electricity consumption for different periods on a basis of similar weather. Once weather normalization has been applied, the differences in energy consumption between both periods are considered to be due to other factors (user behavior, new appliances, etc.).

Clustering of similar customers

Energy consumption can be compared either against historical customer consumption or with other customers with similar characteristics. These consumption comparisons also take into account different time periods: daily, monthly, quarterly, semi-annual, yearly, bi-annual and triennial. In order to obtain similar customers, a clustering procedure is performed.

Several data-mining techniques are used, ranging from supervised learning approaches, based on similar contract information (contracted power, tariff) or geographical information (municipality, postal code, region), as K-nearest neighbours, to unsupervised learning algorithms such as Self-Organizing Maps (SOM) and K-means. The selection of the best grouping criteria for each customer is based on an optimization procedure which is aimed at minimizing a cost function (Equation 2.3) that consists of the difference between the monthly electricity consumption of a customer and the average of their peers and the dispersion of this monthly consumption within this group. To increase the robustness of the grouping criteria selection, the optimization procedure considers the last 12 months available.

$$\min_{X} f(X,c) = \frac{1}{n}\sum_{i=1}^{n} |E_{ic} - \overline{E_{iX}}| + \frac{Q_3(E_{iX})) - Q_1(E_{iX}))}{Q_3(E_{iX})) + Q_1(E_{iX}))}, \tag{2.3}$$

Where X = similar customers, c = customer, n = number of months, $Q_3(E_{iX})$ and $Q_1(E_{iX})$ are the ith monthly consumption 75% and 25% percentiles of the similar users.

In general, the meaningfulness of the grouping criteria is related to the availability of input data and their characteristics. For instance, in the case of customers with only consumption data and no contract or survey information available, the meaning of the chosen grouping criteria is only related to the range of yearly consumption or the shape of the yearly profile. In other cases, when more contract information is available, the meaning of the best grouping criteria could be related to similar contracted power, heating or cooling dependencies, weather severity and also consumption indicators.

In the case of the unsupervised approach, a clustering algorithm is used to group the different kinds of customer. The first step is to train an SOM, a Neural Network (NN) that makes up the low-dimensional representation of the overall set of customers. When some customer features are clustered in a specific neuron, that represents a similar group of customers. The next step consists of a second clustering procedure, using the K-means technique, to find the emergent structures. The inputs used in this second clustering are the centroids from the SOM neurons and their mapping position. The emergent structure offers a more abstract description of a complex system consisting of low-level individuals. The number of groups for the K-means algorithm is optimized using the Gap Statistic index [53].

Two types of features are used in the training phase of the clustering algorithm. The first considers static customer features, such as contract information (contracted power, tariff, heating or cooling resources, location or yearly consumption), weather dependence indicators (explained in section 2.3.3) and daily or weekly consumption profiles averaged over a long period. The second one considers customer dynamic features. It consists of determining likely cyclical consumption patterns. Daylight imposes a natural rhythm on human behavior, making daily series an obvious choice. Inferring how consumers use electricity during different periods of the day on different days of the week and seasons of the year, is considered the most relevant information to be found. In order to obtain this, the SOM + K-means algorithm is used to detect patterns of daily consumption series over the whole set of customers. Fig. 2.3 depicts a subset of those detected patterns for a utility of 6,500 customers. This information is the used by a classification algorithm (i.e. SVM) that detects the pattern closest to every real daily consumption series of each customer. Thus, the results of this classification are a discrete time series of the closest consumption pattern over time. With this discrete time series, the signature of daily consumption series over a limited period (3-4 months, at least) is calculated. Finally, the signature and a K-means algorithm are used to detect the groups of similar customers in terms of energy behavior, because users with similar user behavior over time seem to have a similar signature for daily consumption over a time period.

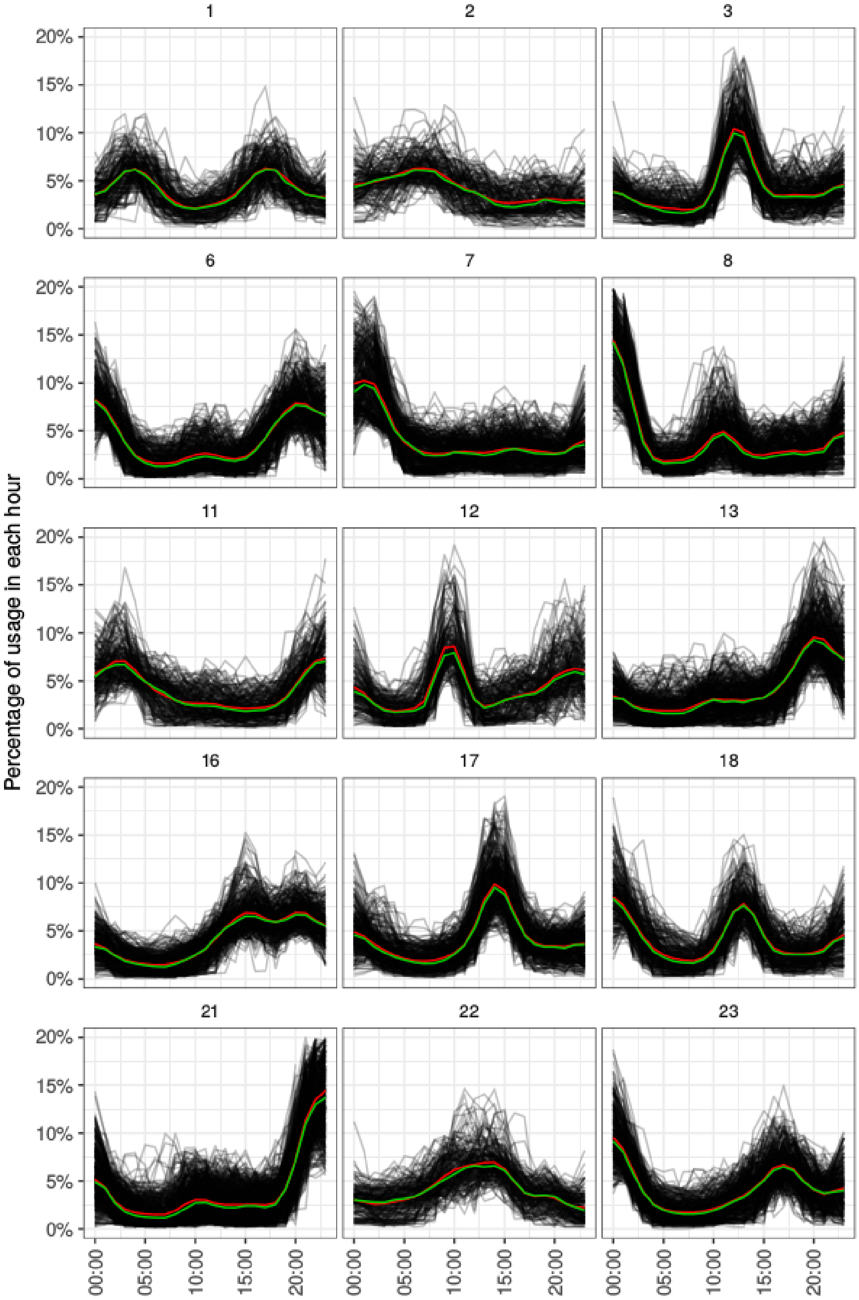


Figure 2.3: Subset of representative daily consumption series of all the customers

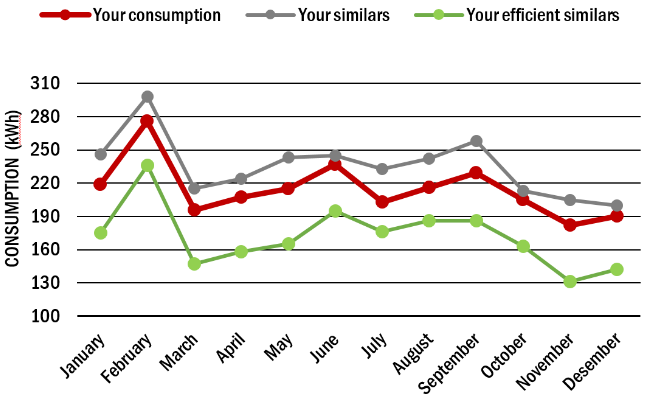


Figure 2.4: Monthly consumption over the last 12 months compared with similar users.

Fig. 2.4 shows a comparison of consumption over the previous 12 months between a customer, similar customers and the most efficient customers within the corresponding cluster.

Forecasting

Electricity forecasting is widely used in EMPOWERING to give consumption prediction information to customers and the utility. The techniques used for forecasting depend highly on the customer characteristics and their energy usage. The AutoRegresive Integrated Moving Average with eXogeneous variables (ARIMAX) is used for those contracts that are weather dependent, because multiple independent variables could be considered in addition to the lagged consumption time series, e.g. the outdoor temperature, solar radiation or wind speed. In the case of contracts without weather dependence, Generalized Additive Models with Autoregressive fitting of the Residuals (GAMAR) are used. Alternatively, the consumption of this type of customers could be forecast using a decision tree which is trained by the information inferred in the clustering of similar customers in order to make the day-ahead predictions.

Tariff comparison

This set of services summarizes high-frequency energy consumption measures and integrates them into the tariff information of the utility. Thus, the customer can visualize which tariff is the most cost-effective considering their real energy usage. Fig. 2.5 depicts the result of the daily consumption of a customer considering a time-of-use tariff in a single month.

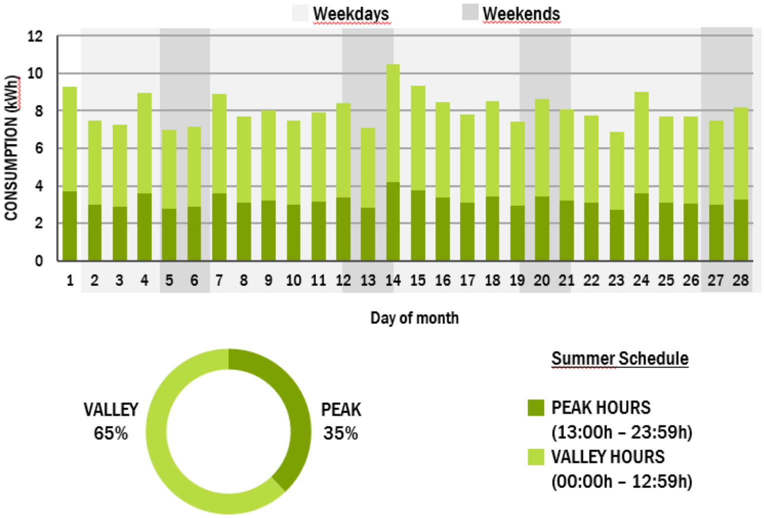


Figure 2.5: Daily consumption for each of the two tariff periods contracted by a customer in one month.

Personalized energy-saving tips

These are the most important services for energy awareness. The energy-saving tips are delivered once a month and the methodology used to select them differs depending on the data frequency. For the monthly data, the energy-saving tips are related to each customer's weather dependence (as defined in 2.3.3) and the season of the year. When a customer has strong weather dependence, they will receive tips related to space heating or cooling systems. In the case of customers with smart meters, the selection of tips is done following the procedure explained in Algorithm 2.1. This procedure is performed once a month. To avoid repetition, the selected tips are excluded from the procedure for a period of four months.

Alerts and alarms

In order to avoid over-consumption in upcoming bills, these services calculate the bias between the actual consumption of each customer and their historical

Algorithm 2.1: Energy Saving Tips

```
1: Clustering of users into similar groups based on each customer's daily consumption pattern
   and following the same techniques as in Section 2.3.3.
2: Evaluation of the average daily pattern of each cluster and the hourly percentage difference
   between this averaged pattern and each customer's pattern.
3: Definition of a set of around 100 energy-saving tips and weighting of each tip every hour of
   the day. For example, tips linked to cooking have a higher weight at midday and in the
   evening.
4: Calculation of the accumulative product between the hourly weight of each tip and the
   hourly percentage difference of each customer's pattern.
5: Classification of the tips to be delivered to each customer based on the score obtained.
6: Delivery of the three tips with the highest scores.
```

consumption and set up an alarm. This allows customers to react within the period between two consecutive energy bills. The frequency of the alarms is directly dependent on the data granularity. For instance, daily or weekly consumption data is needed for monthly alarms. The platform delivers the alerts to the customers through visual interfaces and direct messages.

2.4 Results

2.4.1 Evaluation Metrics

The evaluation of energy savings was based on the difference-in-difference multi-parameter linear regression method according to [54] and re-arranged in Equation (2.4).

$$ADC_m = \alpha + \beta * G_{E,m} + \gamma * t_m + \delta * (G_{E,m} * t_m) + \varepsilon_m, \tag{2.4}$$

This method is widely adopted to evaluate the behavior of energy-efficiency based programs. It only evaluates the differences caused by the delivering of the EMPOWERING services, avoiding the rest of the external factors that affect the customers. It was implemented as a service within the analytical tool with access to the long-term databases. The EMPOWERING databases contain consumption data for all customers, classified as customers who receive the EMPOWERING services, EMPOWERING Group (*EG*), and the remaining ones, making up the Non-EMPOWERING Group (*NG*). An extension of the EMPOWERING data model, within the API Restful, was implemented to include each customer evaluated in the corresponding group and the date when they started using the EMPOWERING services. The energy savings analysis is calculated for each group according to the Averaged Daily Consumption (ADC). The ADC is obtained as the ratio of the aggregated monthly consumption. Once the ADC of each customer and month (m) has been determined, the customer is inserted into a group. The relationship between these variables can be found as a multiple linear regression model and are used to find the ADC of each month (Equation (2.4)).

where:

α Independent parameter. It could be assumed to be the theoretical base-load average daily energy consumption of the total number of customers.

β Parameter related to the difference in energy consumption caused by the effect of belonging to the (EG) or (NG) groups.

G_E Treatment variable. $G_E = 1$ if the customer belongs to EG and $G_E = 0$ if the customer belongs to NG.

t Time period variable. $t = 1$ if the month falls within the evaluation period and $t = 0$ if it is outside the evaluation period.

γ Parameter related to the time trend effect.

δ Parameter related to the combined effect of the customer group and the time trend.

ε Uncertainty error accounting for all effects not considered in the model.

m It corresponds to the time index of all the variables (month).

The parameters are determined through a least square minimisation of the residuals. Once all the parameters have been determined, the expected energy savings, (E_S), achieved by the customers belonging to EG is determined with Equation 2.5.

$$E_S = \frac{\delta}{(\alpha + \beta + \gamma)} * 100\% \tag{2.5}$$

2.4.2 Energy Savings

The EMPOWERING architecture and services were applied in three pilot experiments in France, Spain and Austria for slightly over 2 years, from November 2013 to December 2015. In each country, a local electricity-supplier was responsible for gathering data from customers, putting this into the analytical platform presented in section 2.3.1 to obtain data analysis and deliver them to the customers through several user interfaces. The details of the communication channel to deliver the services and the number of users included in the *EG* and *NG* groups as follows:

- **Spain:** The services were provided to the customers in two ways: (group 1) through an on-line portal and (group 2) as a monthly energy report. Meter readings were taken daily. The energy reports were sent together with the energy bill every 2 months. The *NG* and *EG* groups consisted of 3,129 and 1,582 customers respectively.

- **France:** The services were also offered to customers as an on-line tool within the utility web portal. Meters were read at a frequency of 6 months, but the services used an estimated 3-month consumption. To evaluate the energy savings of similar users, a clustering of the customers belonging to the *NG* was performed based on the contracted power: (group 1) low contracted power and electricity use limited to home appliances; (group 2) high contracted power and electricity used mainly in space heating systems; (group 3) low contracted power with occasional use of electricity for domestic hot water and space heating. The *NG* and *EG* were formed of 4,632 and 60 customers respectively.

- **Austria:** The services were offered to the customers as an on-line tool within the utility web portal. Meters were read every 15 minutes. The *NG* and *EG* groups were made up of 45,423 and 115 customers respectively.

After the test, the evaluation of the energy savings achieved by the group of customers belonging to *EG* was performed following the methodology defined in Eq. 2.5. Fig. 2.6 shows the percentage of energy savings achieved in the Spanish, French and the Austrian pilot projects. As can be seen, the customers who used electricity mainly for space heating and domestic hot water systems or those who received both energy reports and access to on-line tools achieved greater savings in electricity.

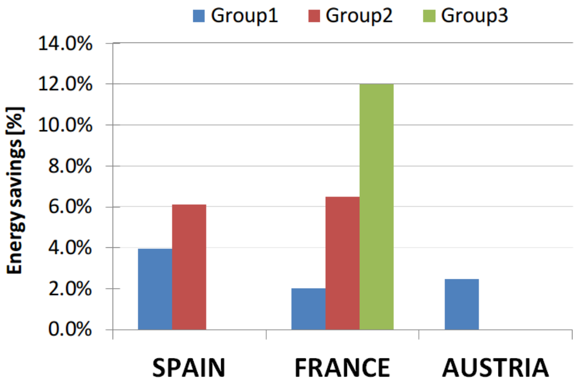


Figure 2.6: Average energy savings achieved in the Spanish, French and Austrian pilot projects.

Figure 2.7 shows the energy savings of the two groups of customers in the Spanish pilot project segmented into percentiles of electricity consumption. A similar pattern can be appreciated for both groups. In general, higher savings were achieved in the higher energy consumption segments. Both groups reduced consumption significantly. The savings achieved were considerably more higher for the customers that receive billing tools (6%), compared to the users that only used the online tool (4%), which could be related to the low user acceptance of utility web dashboards. Thus, offering the services through the proper channels could improve the savings significantly.

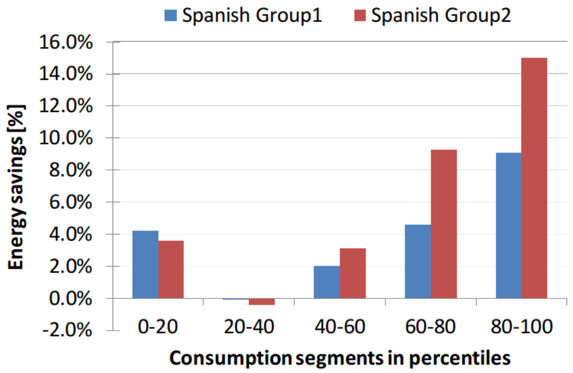


Figure 2.7: Energy savings per consumption distribution percentile range for the Spanish pilot project.

Fig. 2.8 shows the energy savings achieved by the three groups of customers in the French pilot project, segmented into the percentiles of electricity consumption. It indicates higher energy savings for the middle-high percentile range customers in groups 1 and 2. In group 3, the savings are present over the whole range of consumption, with higher (up to 22%) savings for the largest consumers. It can also be seen that electricity savings were achieved in all three groups. Savings were higher in the groups where the electricity was used for both space heating and hot water (Groups 2 and 3). Considerable savings, above 20%, were achieved in Group 3, this being the group with more opportunities to modify their energy usage habits. The number of customers in the *EG* was relatively small for the three evaluation groups, allowing room for large uncertainty in the evaluated results.

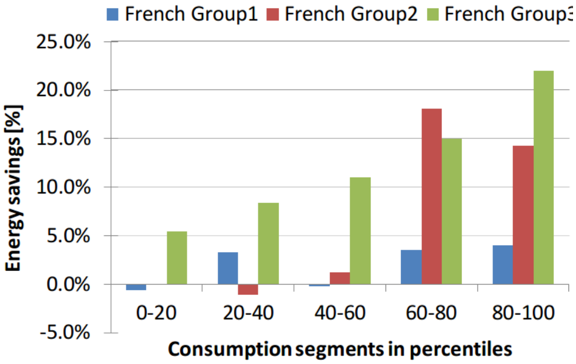


Figure 2.8: Energy savings per consumption distribution percentile range for the French pilot project.

Fig. 2.9 shows the energy savings achieved among the customers in the Austrian pilot project, segmented into the percentiles of electricity consumption. Higher energy savings were achieved by customers in the upper consumption segments, while the customers with lower consumption barely increased their energy consumption. The opt-in strategy also meant that only very motivated customers entered the portal. The baseline consumption before the services was 8.15 KWh/day and the average savings per user were 0.5 KWh/day.

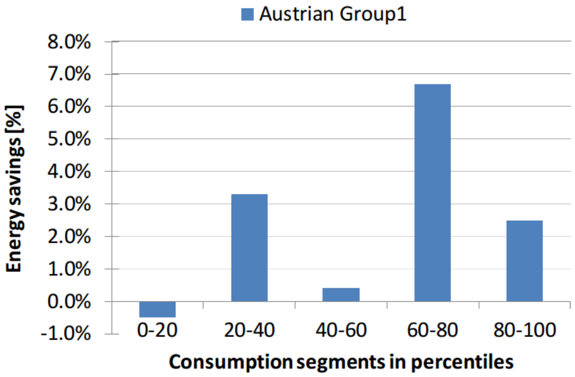


Figure 2.9: Energy savings per consumption distribution percentile range for the Austrian pilot project.

2.5 Conclusions and future work

In this chapter, we present an efficient and scalable platform aimed at helping domestic customers to save energy by managing their energy consumption positively. The electricity savings achieved by using EMPOWERING ranged from an average of 2 to 12% among the different pilot and user groups. Improvements in the behavioral aspects in energy use have considerable potential. The users' own motivation also seems to play an important role and thus, better results were achieved with customer involvement. The personal motivation for energy savings is based on different reasons and money saving is only one of them. Environmental concerns, governmental laws, social policies and technological restrictions are other powerful reasons where the future services should diversify in order to have greater impact. In addition, more encouraging and ad-hoc services must be provided to the final customers. Future work will analyze energy awareness depending on the nationality of the customers. We leave this for the future work due to the complexity of the diversity of features as well as clustering groups to be analyzed.

Chapter 3

Data-driven virtual replication of domestic thermostatically controlled loads

This chapter has been published as a paper:

Mor, G.; Cipriano, J.; Gabaldon, E.; Grillone, B.; Tur, M.; Chemisana, D. Data-Driven Virtual Replication of Thermostatically Controlled Domestic Heating Systems. Energies 2021, 14, 5430. https://doi.org/10.3390/en14175430

3.1 Introduction

In 2019, the final energy consumption of the residential sector accounted for 26% of the overall final energy consumption in the EU [55]. The main use of this final energy was for space heating, representing around 64% [55]. Most EU Member States rely mainly on natural gas and electricity for meeting these needs, followed by renewable energies, mostly solid bio fuels. This high dependence on natural gas clearly determines any achievable strategy to reach the binding carbon targets. As stated in [56], energy saving is one of the easiest ways to save money for consumers and to reduce greenhouse gas emissions. The EU has set binding targets of at least 32.5% improvement in energy efficiency by 2030. To achieve this increase in energy efficiency on the global scale, more effort in energy conservation strategies or in electrification of buildings' technical systems should be dedicated to this endeavor. The electrification can be based on several mature technologies, such as electricity driven heat pumps, hybrid heat pumps, or district heating networks. Many research studies have focused on demonstrating their cost effectiveness and how these technologies can increase the energy efficiency in several

European countries [57, 58, 59, 60, 61]. This strategy is the best option in the mid-long term. However, in the short term, cost-efficient strategies, able to drastically reduce the energy consumption of legacy space heating systems and, in particular, thermostatically driven systems (fed with gas), should be also accelerated.

Another challenge to address is related to the users' involvement in the energy transition. Although the technologies are readily available, the control strategy, as well as the involvement of end users in their management is not fully clarified yet. End users must be part of the solution, and this can only be achieved if manufacturers of home space heating/cooling systems, which should be one of the drivers of the low-carbon transition, can find new and more interactive ways to support end customers. The unfolding of these user driven energy control strategies requires higher digitization of the existing systems. Manufacturers should accelerate the virtualization (digital twins) of the operation of their systems to drastically improve the user interaction and the automatic demand response. This process needs some kind of Advanced Metering Infrastructure (AMI) or a massive adoption of smart home devices. To date, Member States committed to rolling out close to 200 million smart meters for electricity and 45 million for gas by the end of 2020 at a total potential investment of EUR 45 billion [62]. By the end of 2021, it is expected that almost 72% of European consumers should have a smart meter for electricity, while 40% should have one for gas.

On the other hand, for the few last years we have seen a fast penetration of the emerging Internet of Things (IoT) technologies into residential homes. Nowadays, smart devices are inevitable in our lives [63, 64]. Smart thermostats are one of them. These smart thermostats allow remote control of the home climate, display of the temperature and energy consumption in real time or communication with intelligent cloud-based IT systems to incorporate self-learning capabilities. These are crucial features to accommodate efficient techniques to increase the energy efficiency of space conditioning systems and decrease energy costs. However, some studies [65] showed that 40% of programmable thermostats are used in manual modes, mainly due to confusing user interfaces. Peffer et al. [66] stated significant failures in people–technology interactions when they set their programmable thermostats. They also pointed out some of the needed characteristics to overcome the misconceptions about thermostat operation. For instance, to provide accessible web portals or mobile applications or to add voice recognition features, or indicators of how much time the heating system needs to achieve a desired temperature. Although smart thermostats include some of these features, which help increase the user's satisfaction, some studies [67] reflect that the end users are still reluctant to rely on the smart thermostat to control their boiler or heat pump. In [68], product reviews of five smart thermostats were collected and analyzed. When comparing the most commonly discussed topics, generally they were not related to energy and cost saving. The most discussed topics were control, ease of use, and installation. In [67] a comparison of two different smart thermostats included an evaluation of the achieved gas savings. The main conclusion was that there appeared to be higher gas savings in homes where the occupancy detection features were enabled. Data gathered by connected thermostats are also useful in understanding the operational and occupancy patterns of users. A longitudinal analysis [69] was conducted in relation to thermostat operation behavior due to the climate, season, and price and to the thermal preferences. It was used to categorize users based on operation. Furthermore, a study [70] on residential households located in high-rise buildings, using complementary survey data, demonstrated the potential benefit of using connected thermostat data as a diagnostic tool to identify opportunities for energy savings in this type of building. In [71], various models designed to predict the user occupancy, based on machine learning and deep learning methods, are compared. Optimal set point temperature scenarios can be also estimated using these occupancy prediction models.

Therefore, while thermostats' capabilities to control the indoor temperature, mainly based on occupancy detection, are well understood, less is known about their effectiveness to enable energy savings. The uncertainty in relation to the potential energy savings is increasingly important because manufacturers are adding many new features and functions to the thermostats without detailed assessment of their impact on the gas or electricity consumption. Previous research studies demonstrated a high variation in the achieved energy savings due to the substitution of conventional thermostats with smart thermostats. In [72], an assessment of two smart thermostat models is performed, and a high variation of the achieved energy savings, among users with the same smart thermostat, is documented. Moreover, although these smart thermostats were focused on occupancy-responsive control, the specific actions which led to the energy savings as well as the reasons of these high variations are not clearly determined. In [73, 74], more detailed assessments of the energy savings achieved by occupancy responsive thermostat control are performed. A clear relationship between this occupancy-based control and the achieved energy savings, supported by supervised learning data-driven models, can be found. Nonetheless, the effect of other control variables such as variations in the set point temperature are not analyzed in detail. Some studies, performed by the National Research Council Canada in their experimental set up (CCHT twin houses), analyzed the effect of thermostat setback strategies over the energy consumption [75]. They tested three setback strategies for the winter season and two more for the summer season. Their research conclusion was that these strategies can be very effective in winter but not in the summer. The research was very accurate in evaluating setback strategies; however, they were tested in non-occupied and highly controlled home environments and they were limited to the applied setback schedules. They did not include dynamic modeling calibration or advanced thermostatically controlled strategies. More research in prediction and control optimization techniques, addressing the uncertainty in the evaluation of the effect over the energy consumption, are certainly necessary.

The prediction and control optimization models should be able to include not only the occupancy and the weather-dependent variables but also the control variable which, in most cases, is the set point temperature. In [76], a review of the state of the art of dynamic models able to predict natural gas consumption, from 2000 to 2010, was presented. From this review, it can be ascertained that an exponential increase in papers was detected in this field, especially in the lower forecasting area level (regional, gas distribution and individual). The predominant trend of these research works was a combination of optimization tools with more classic forecasting models. After 2010, several authors continued using statistical and stochastic methodologies to predict and characterize aggregated gas consumption of residential units or groups of commercial buildings [77, 78]. At the individual level, in [79] Nonlinear Mixed-Effects models (NLME) are used for the prediction of single gas consumption at daily basis. After comparing the results among auto regressive models, such as AutoRegressive with eXogenous variables (ARX) and AutoRegressive Moving Average with eXogenous variables (ARMAX) models, the conclusion was that such models perform similarly but have both merits and problems. The NLME models are cleaner and clearer, while ARX and ARMAX are better for local adaptation to sudden and abrupt changes within a single individual. In [80], linear ARX, Artificial Neural Networks (ANN) and Support Vector Machine Models (SVM) are applied to forecast natural gas consumption on a daily basis. The solar radiation as an exogenous variable was included in the models and the accuracy improved. That research work performed a very detailed evaluation of several Time Series (TS) models in non-occupied test homes and clearly quantified the model accuracy improvement by introducing the solar radiation as an exogenous variable. The results were encouraging, however these test conditions were very far from real and occupied buildings where the heating system is thermostatically controlled by the user through the set point temperature. In [81], a step wise calibration of a dynamic thermal empirical model of a residential building was performed. The calibration included some user-dependent parameters, such as the air ventilation rates; however, the constraints derived by the set point temperature control were not included in the analysis. More recently, Wang et al. [82] developed a home thermal dynamic model built upon the standard Resistance and Capacitance (R-C) approach and tested it with data from a test home in free-floating conditions. This R-C model included the effect of most of the exogenous variables, such as the internal and external temperatures, the wind direction and the solar radiation, though it did not consider the effect of the set point temperature and of the user behavior. Alinberti et al. [83] developed a non-linear Autoregressive Neural Network model for short and medium-term predictions of the indoor temperature of a secondary school building. The accuracy of the predictions is very well evaluated; however, as in the previous literature works, the model cannot evaluate the effect of the set point temperature in the energy consumption. In [84] a machine learning model to predict residential energy consumption based on data from Wireless thermostats is developed. Although the results are very promising in relation to the energy savings evaluation, the developed technique requires many data of the building features and it is limited to monthly frequency. This could be a clear limit for wider application and for near-real time control solutions.

Recent studies moved one step beyond the prediction of the energy performance of thermostatic load control systems and assessed control-optimized techniques within Demand Response (DR) programs or in relation to the electricity network operation [85]. In [86], the set point temperature of thermostatically controlled systems is included in the evaluation of the demand response programs in 1000 households. That paper is based on synthetic data; however, it demonstrated how an accurate modeling of the thermostatic control of space heating and cooling systems enables simple and reliable evaluation of demand response and of Energy Conservative Measures (ECM) in the residential sector. These emerging applications require very fast and computation efficient data-driven models able to provide the necessary response.

From these previous research works, it can be concluded that, although the knowledge of the energy performance of thermostatic load controlled systems is growing fast, there are still some gaps in relation to the modeling of the combined effect of the thermal energy supplied by the heating system, of the user-based thermostatic control driven by the set point temperature and of the exogenous variables (external weather conditions). Furthermore it is also stated that more advanced modeling strategies, able to virtually mimic the performance of the thermostatic control, are needed if we want to increase the smartness of these systems and to enhance interactions with the customers. In our research, a new methodology to emulate the performance of thermostatic load controlled systems is developed and put in practice. The novelty relies on the fact that, unlike most literature solutions, which limit their applicability to forecasting the indoor temperature or the energy consumption separately, our approach combines several optimization techniques, with auto regressive models and a control loop, to model cross-combined effects and to mimic all the possible control modes driven by the control variable (the set point temperature). The control loop included in the methodology is based on the difference between the indoor and the set point temperatures. The mode when the indoor temperature is higher than the set point threshold is modeled by a first regression model where the indoor temperature is the dependent variable and the space heating power consumption is one of the input variables. This space heating power consumption becomes a dependent variable, fed by the indoor temperature and other exogenous variables, when the indoor temperature is lower than the set point temperature threshold. Both regression models are combined to forecast the expected energy consumption and the potential energy savings when a certain set point temperature schedule is applied. The methodology was validated in real cases within a heating season. However, a similar implementation should be applicable also to space cooling system as long as they are thermostatically-controlled systems.

The paper starts with a mathematical description of the regression models and of the input variables transformation. It follows with a description of the processes used to train both models and to optimize the regression parameters. The procedure used to combine the two regression models to predict the energy consumption, and the potential energy savings due to a certain set point temperature schedule, is then described. The paper finishes with the application of the methodology over a set of households in northeastern Spain, which are equipped with condensing gas boilers driven by smart thermostats.

3.2 Methodology

The energy performance of a household is influenced by many factors that include the dynamic indoor and outdoor conditions, the physical and geometric characteristics of the building, the type of space conditioning system and, finally, the control of this system, which in most cases is a thermostat controlled by the end-users. Therefore, when modeling the energy performance of real households using data-driven models, all these factors should be considered. In this chapter, a methodology is developed to accurately predict the energy consumption and indoor temperature of thermostatically controlled heating systems. Technically, the methodology combines two ARX models, named the demand-side and the supply-side models, in order to dynamically simulate the heat losses and gains of the building due to changes in the thermostat set point temperature. The demand-side model captures the heat dynamics affecting the indoor temperature of the household, while the supply-side model determines the heat dynamics concerning boiler energy consumption. The two models, and their control loop coupling, are trained using historical data of real systems performance during occupancy. Figure 3.1 depicts the general flow diagram of the developed methodology. The first step starts with the gathering of historical data available from smart thermostats reading and from weather forecasting web services which provide climatic data. Then, both data-driven models are trained using these data sets. Subsequently, these models are used as a simulation tool to estimate the energy consumption and indoor temperature due to changes in set point temperatures. Finally, the set of validated algorithms can be used for multiple smart-control applications, such as Model Predictive Control (MPC) or short-term forecasting. The outputs of these applications, in turn, can generate more data which can be fed into an iterative self-learning process to re-train the models.

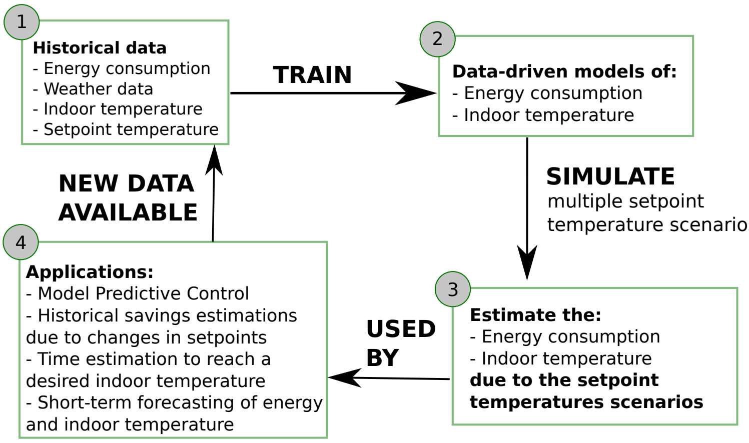


Figure 3.1: General steps and objectives of the modeling technique presented.

The use of two models is justified because the heat dynamics of the building are not affected by only the external variables and the supplied energy. They are also affected by the indoor conditions. The lower the indoor temperature, the higher the energy to be supplied to reach the comfort conditions defined by the set point temperature thresholds. One of the models, the demand-side model, is used to simulate the indoor temperature of the household in free-floating conditions, when energy delivered by the heating system is zero. The other model, the supply-side model, is used to estimate the energy needed to recover the indoor comfort conditions when the supply system is activated again. Figure 3.2 shows 3 different scenarios of simulated set point temperature schedules,$T^{s,sim}$, the corresponding simulated indoor temperature changes, T^i and the supply energy delivered by the gas boiler to recover the comfort conditions, Φ^h. As can be seen, the length of the free-floating periods determine the indoor temperature decay and the energy to be supplied by the gas boiler consumption to reach the set point temperature schedule again. ARX models were selected, because these kind of black-box models contain autoregressive impulse responses which can properly describe time-varying processes in a fast and efficient way. In addition, as can be seen in Section 3.2.5, a hybrid optimization procedure, considering least squares and a Genetic Algorithm (GA), is applied to fit the models and to identify the unknown parameters. Last but not least, in Section 3.2.3, a description of the prior transformations applied to several input variables, along the training phase, are presented.

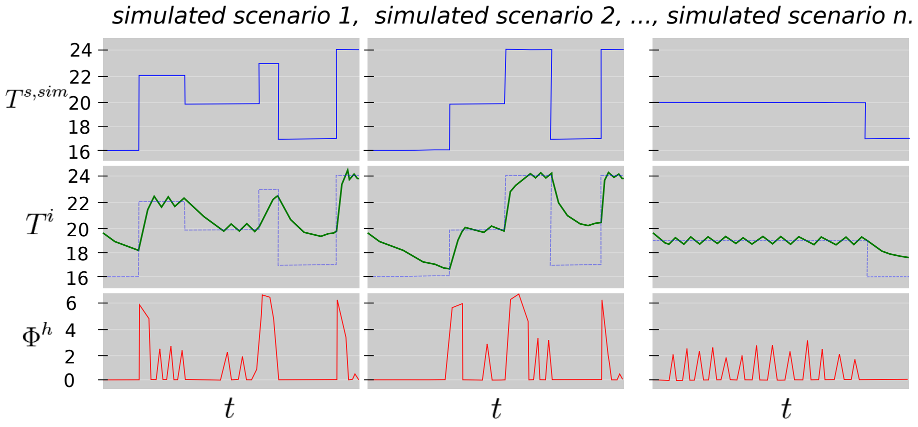


Figure 3.2: Theoretical examples of 3 different set point temperatures scenarios over the same time period and their related indoor temperature and space heating consumption.

3.2.1 Demand-Side Model

The demand-side model is defined by an ARX model represented by the indoor temperature (T^i) as the output. The external weather conditions and the space heating consumption are the input variables. This model captures how the heat flows out of the building and how the indoor temperature is affected by the space heating system. The model formula is described in Equation (3.1).

$$\begin{aligned}\phi(B)T^i_t = \omega_h(B)\Phi^h_t + \omega_e(B)T^e_t + \omega_p T^{e,lp}_t + \omega_i(W^{s,lp}_t \times W^{d,fs}_t \times \Psi_t) + \\ \omega_s(I^{sol,lp}_t \times S^{az,fs}_t \times S^{el,fs}) + \varepsilon_t\end{aligned} \tag{3.1}$$

The coefficients $\phi(B)$, $\omega_h(B)$, $\omega_e(B)$, ω_p ,ω_i and ω_s are the parameters of the model, where: B is the backward shift operator B, defined as $B^k y_t = y_{t-k}$, k is the auto-regression order, y is the considered variable, for instance, the indoor temperature in the case of $\phi(B)$ or the outdoor temperature in $\omega_e(B)$.

The independent variables considered in the model are:

- Time-lagged (n) indoor temperatures (T^i_{t-n}) to characterize the inertia of the building.

- Low-pass filtered outdoor temperature ($T^{e,lp}$) to characterize the heat loses through the envelope of the building due to changes in the outdoor temperature. Compared to the T^e, should be understood as a temperature that represents better the internal temperature of the envelope.

- Raw outdoor temperature (T^e) to properly model changes in indoor temperature due to fast changes in the outdoor temperature, specially convenient, for example, in low-inertia buildings, or buildings with large single-glazed windows.

- Heat consumption of the boiler (Φ^h) to characterize the increase in the indoor temperature due to the operation of the heating system.

- Solar direct normal irradiance ($I^{sol,lp}$), interacting with the Fourier series of the solar azimuth ($S^{az,fs}$) and of the solar elevation ($S^{el,fs}$) to characterize the solar gains of the building.

- Wind speed ($W^{s,lp}$), interacting with Fourier series of the wind direction ($W^{d,fs}$) and the temperature difference between indoors and outdoors ($\Psi =$

$T^i - T^e$) to characterize the heat losses due to air leakage and convection effects through the envelope.

3.2.2 Supply-Side Model

This dynamic model estimates the amount of energy needed to warm up the household considering the inertia of the building, the external weather conditions, the performance of the boiler and its thermostatic control.

$$\begin{aligned}\gamma(B)\Phi_t^h = \beta_t(B)T_t^i + \beta_e(B)T_t^e + \beta_p T_t^{e,lp} + \beta_i(W_t^{s,lp} \times W_t^{d,fs} \times \Psi_t)+ \\ \beta_s(I_t^{sol,lp} \times S_t^{az,fs} \times S^{el,fs}) + \varepsilon_t\end{aligned} \tag{3.2}$$

In this model, $\gamma(B)$, $\beta_t(B)$ $\beta_e(B)$, β_p, β_i, β_s are the coefficients of the model. The output is the log-transformed consumption Φ^h. The inputs of the model are:

- Time-lagged (n) heat consumption (Φ_{t-n}^h) to consider how the boiler was performing in the last time steps.

- Raw data of the outdoor temperature (T^e) to consider the variation of the coefficient of performance of the boiler due to changes in the outdoor temperature.

- $T^{e,lp}$ is the low-pass filtered version of the outdoor temperature. It represents the temperature of the building envelope.

- As in the demand-side model, the solar direct normal irradiance ($I^{sol,lp}$) interacts with the Fourier series of the solar azimuth ($S^{az,fs}$) and of the solar elevation ($S^{el,fs}$).

- Ψ as in the case of the demand side model, it is the temperature difference between indoors and outdoors.

- Wind speed ($W^{s,lp}$) interacts with Fourier series of the wind direction ($W^{d,fs}$) and the temperature difference between indoors and outdoors ($\Psi = T^i - T^e$).

Unlike the demand-side model, the datasets used to estimate the γ and β parameters only consider the periods where $\Phi_t^h > 0$. This is because no information can be extracted about the performance of the boiler in the periods when it is not operating.

3.2.3 Transformation of Input Variables

Low-Pass Filter

The application of a Low-Pass Filter (LPF) over the exogenous variables, used as inputs of the models, transforms them into variables that better represent the dynamics of the system and, therefore, the model fitting is improved. The LPF assumes that the dynamics of the buildings can be described by lumped parameter R-C models; see for example [87, 88]. This assumption means the response of the indoor temperature or the energy consumption to changes in some climate exogenous variables can be modeled as a first order LPF. Based on this assumption, it is reasonable to apply LPF to all the exogenous variables in order to eliminate the high input frequencies that might negatively affect the model training. The discrete time implementation of this first order R-C LPF is the exponentially weighted moving average of each variable with the filter parameter (α) tuned to match the response of the building to each effect separately:

$$x^{lp} = LPF(x, \alpha) \tag{3.3}$$

$$x_t^{lp} = \alpha x_t + (1 - \alpha) x_{t-1}^{lp}, \tag{3.4}$$

where x^{lp} is the filtered exogenous variable, α is the filter parameter $[0, 1]$, and x is the original time series of the exogenous variable.

As described in Equations (3.1) and (3.2), outdoor temperature $T^{e,lp} = LPF(T^e, \alpha_e)$, wind speed $W^{s,lp} = LPF(W^s, \alpha_w)$ and solar irradiance $I^{sol,lp} = LPF(I^{sol}, \alpha_s)$ are the inputs which are low-pass filtered for some of the terms used in the models.

Fourier Series

The correlation between indoor temperature (T^i), solar irradiance ($I^{sol,lp}$) and air leakage ($W^{s,lp}\Psi$) is, normally, non-linear. Multiple reasons lead to this behavior, such as: building envelope orientation and characteristics, sun position and wind direction. To solve this issue, a harmonic function, based on a Fourier series, is used to account for these non-linearities. Solar azimuth S^{az}, solar elevation S^{el} and wind direction W^d are the observations transformed using this technique, and the number of harmonics considered are, respectively, $n_{har,az}$, $n_{har,el}$ and $n_{har,wd}$.

$$Y^{fs} = \sum_{h=0}^{n_{har}} \begin{cases} \theta_0 & if\ h = 0; \\ \theta_{h,1} sin\left(2\pi hY\right) + \theta_{h,2} cos\left(2\pi hY\right) & otherwise \end{cases} \tag{3.5}$$

In Equation (3.5), Y represents the observation to be transformed, Y^{fs} is the transformed variable, n_{har} is the maximum number of harmonics included in the Fourier series $[0, \infty)$, and θ are the regressors of each component. In the demand and the supply-side models, the generic θ coefficients depicted in Equation (3.5) are identified following the same procedure as ω_i, β_i, ω_s and β_s parameters

3.2.4 Models Coupling

The supply-side and the demand-side models are coupled to allow the simulation of both the space heating energy consumption and the indoor temperature, given a certain set point temperature schedule.

Figure 3.3 accurately describes how the models are coupled (Algorithm 3.1 of the Appendix). In essence, it mimics the operation of a thermostat considering the heat transfers of a household and setting on or off the operation of the boiler according to the set point temperature. At each time step, the algorithm predicts the variation of the indoor temperature in free-floating conditions, and then, when the set point temperature is higher than the indoor temperature, it simulates the space heating operation by estimating both the energy consumption and the indoor temperature the household will reach.

3.2.5 Model Training and Parameter Optimization

The linear least squares method is used to estimate the ω, β, ϕ and γ parameters of both ARX models. However, there are more parameters to be optimized: the coefficients of the input feature transformations and the auto regressive orders of the ARX models. Those parameters cannot be estimated using the least squares method used in the regression analysis. Therefore, a Genetic Algorithm (GA) technique is used as the optimizer for those coefficients. The GA evaluates several combinations of a set of coefficients and then estimates the remaining ones (ω, β, ϕ and γ) using the least squares method. The cost function is defined in Equation (3.6). The GA is based on the R package GA, developed by Scrucca et al. [89]. The GA package provides a flexible general-purpose set of tools for implementing a genetic algorithm search in both the continuous and discrete case, whether constrained or not. In this research, a binary GA is selected within the available tools of the GA package. A binary GA is a simple and flexible optimizer able to simultaneously include multiple integer, continuous and discrete variables. More specifically, a Reflected Binary Code (RBC) representation, which is an ordering of the binary numeral system such that two successive values differ in only one bit, is used as the binary representation of each chromosome evaluated by the GA. This RBC enhances the optimization process during the recombination and mutation steps. Algorithm 3.2 describes in detail this optimization procedure. Algorithm 3.3 describes the way in which the cost of each chromosome is calculated during the evaluation steps of Algorithm 3.2. The cost function considered in this optimization is defined in Equation (3.6). It consists of a combination of the Coefficient of Variation of the Root Mean Squared Error (CVRMSE) of the indoor temperature and of the space heating energy consumption. Although the CVRMSE is not affected by zero values of the boiler energy consumption, it is only computed for households with aggregated historical energy consumption greater than zero, $\overline{\Phi^h} > 0$. As can be seen in Algorithm 3.3, the cost of each chromosome is evaluated using the cross-validation folds along a testing period.

$$C = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(\widehat{T_t^i} - T_t^i)^2}}{\overline{T^i}} + \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(\widehat{\Phi_t^h} - \Phi_t^h)^2}}{\overline{\Phi^h}} \tag{3.6}$$

Once all the parameters are optimized, the supply-side and the demand-side models are considered as correctly validated and are ready to be used for further evaluations.

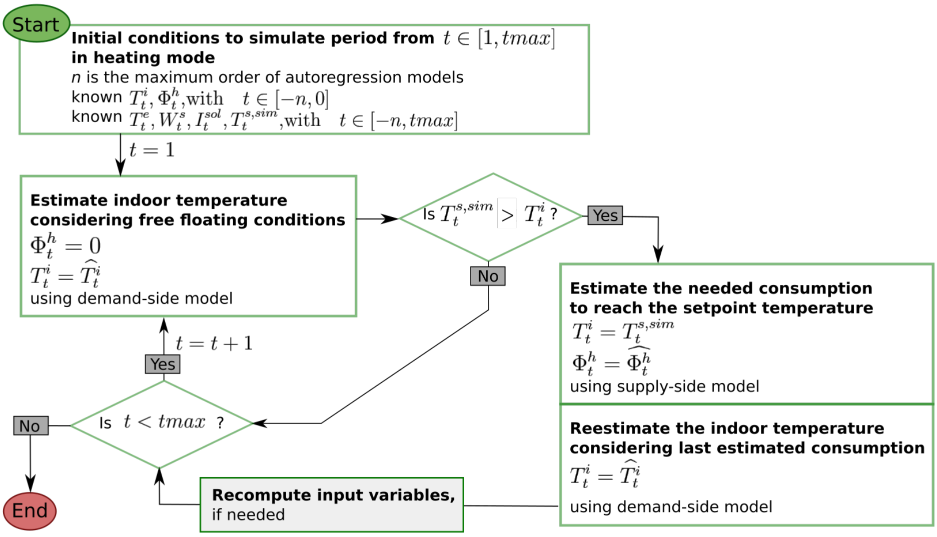

Start
Initial conditions to simulate period from $t \in [1, tmax]$ in heating mode
n is the maximum order of autoregression models
known T_t^i, Φ_t^h,with $t \in [-n, 0]$
known $T_t^e, W_t^s, I_t^{sol}, T_t^{s,sim}$,with $t \in [-n, tmax]$
$t = 1$
Estimate indoor temperature considering free floating conditions
$\Phi_t^h = 0$
$T_t^i = \widehat{T_t^i}$
using demand-side model
Is $T_t^{s,sim} > T_t^i$?
Yes
No
Estimate the needed consumption to reach the setpoint temperature
$T_t^i = T_t^{s,sim}$
$\Phi_t^h = \widehat{\Phi_t^h}$
using supply-side model
Reestimate the indoor temperature considering last estimated consumption
$T_t^i = \widehat{T_t^i}$
using demand-side model
$t = t + 1$
Yes
Is $t < tmax$?
No
End
Recompute input variables, if needed

Figure 3.3: Models coupling flow diagram.

Table 3.1: Columns of A matrix.

Column	Conditions
T_{t-k}^e	$k \in \mathbb{N} \wedge k \leq max(n_{\beta_e(B)}, n_{\omega_e(B)})$
$T_{t-k}^{e,lp}$	$k \in \mathbb{N} \wedge k \leq max(n_{\omega_p(B)}, n_{\beta_p(B)})$
$\widehat{T_{t-k}^i}$	$k \in \mathbb{N} \wedge k \leq max(n_{\phi(B)} n_{\omega_e(B)})$
Ψ_{t-k}	$k \in \mathbb{N} \wedge k \leq n_{\beta_p(B)}$
$\widehat{\Phi_{t-k}^h}$	$k \in \mathbb{N} \wedge k \leq max(n_{\omega_h(B)}, n_{\gamma(B)})$
$S_t^{az,fs,h_{sin}}$	$h_{sin} \in \mathbb{N} \wedge 1 \leq h_{sin} \leq n_{har,az}$
$S_t^{az,fs,h_{cos}}$	$h_{cos} \in \mathbb{N} \wedge 1 \leq h_{cos} \leq n_{har,az}$
$S_t^{el,fs,h_{sin}}$	$h_{sin} \in \mathbb{N} \wedge 1 \leq h_{sin} \leq n_{har,el}$
$S_t^{el,fs,h_{cos}}$	$h_{cos} \in \mathbb{N} \wedge 1 \leq h_{cos} \leq n_{har,el}$
$W_t^{d,fs,h_{sin}}$	$h_{sin} \in \mathbb{N} \wedge 1 \leq h_{sin} \leq n_{har,wd}$
$W_t^{d,fs,h_{cos}}$	$h_{cos} \in \mathbb{N} \wedge 1 \leq h_{cos} \leq n_{har,wd}$
$I_t^{sol,lp}$	-
$W_t^{s,lp}$	-
$T_t^{s,sim}$	-

Algorithm 3.1: Forecasting algorithm and coupling of supply-side and demand-side models

```
Input: Trained supply-side model; trained demand-side model; autoregressive
       orders $n_{\gamma(B)}$, $n_{\beta_t(B)}$, $n_{\beta_e(B)}$, $n_{\beta_p(B)}$, $n_{\phi(B)}$, $n_{\omega_h(B)}$, $n_{\omega_e(B)}$, $n_{\omega_p(B)}$;
       number of harmonics $n_{har,az}$, $n_{har,el}$ and $n_{har,wd}$; the smoothing
       parameters of the low-pass filter $\alpha_e$, $\alpha_s$ and $\alpha_w$; initial indoor conditions
       ($T^i$), weather conditions during the whole evaluation period (outdoor
       temperature $T^e$, wind speed $W^s$, wind direction $W^d$, solar irradiance
       $I^{sol}$, and solar position $S^{az}$, $S^{el}$), the space heating consumption few
       timesteps before the period to be evaluated the hysteresis of the
       thermostat $h$ and, finally, the setpoint temperature ($T_t^{s,sim}$) to apply
       during the evaluation period
Output: The predicted heat consumption ($\widehat{\Phi^h}$) and the predicted indoor
        temperature ($\widehat{T^i}$) considering a setpoint temperature schedule ($T^{s,sim}$)
        during a period $ts \in [0, j]$.
begin
    SET $ts = 0$;
    DEFINE the $A$ input–output matrix ($A \in \mathbb{R}^{j,i}$). The $i$ columns are described in Table 3.1. From now on, variables are referred to columns in $A_{ts,*}$;
    SET the autoregressive terms $Y_{t-k} : k \in \mathbb{N} \wedge k > 0$ of the next variables: $\widehat{T^i}$ ($T^i$ is used), $\widehat{\Phi^h}$ ($\Phi^h$ is used) and $\Psi$ ($T^e$ and $T^i$ are used) at their respective columns in $A_{ts,*}$;
    while $ts \leq j$ do
        $\widehat{\Phi_t^h} = 0$;
        ESTIMATE $\widehat{T_t^{i,lp}}$ using the demand-side model;
        if $\widehat{T_t^{i,lp}} < (T_t^{s,sim} - h)$ then
            SET $\widehat{T_t^i} = T_t^{s,sim} + h$;
            CALCULATE $\Psi_t$ using, among others, last set $\widehat{T_t^i}$;
            ESTIMATE $\widehat{\Phi_t^h}$ using the supply-side model;
            ESTIMATE $\widehat{T_t^i}$ using the demand-side model;
        for $hr \leftarrow 1$ to $max(n_{\omega_h(B)}, n_{\omega_e(B)}, n_{\beta_h(B)}, n_{\beta_h(B)}, n_{\gamma(B)}, n_{\phi(B)})$ do
            SET the autoregressive terms $Y_{t-k} : k \in \mathbb{N} \wedge k > 0$ of the next variables: $\widehat{T^i}$, $\widehat{\Phi^h}$ and $\Psi$ at their respective columns in $A_{ts+hr,Y}$;
        SET $ts = ts + 1$;
    $\widehat{\Phi^h} = A_{*,\widehat{\Phi_t^h}}$;
    $\widehat{T^i} = A_{*,\widehat{T_t^i}}$ ;
```

Algorithm 3.2: Genetic Algorithm for the optimization of the auto regressive orders ($n_{*(B)}$), the low-pass filter (α_*), and the number of harmonics ($n_{har,*}$) to be considered in the transformation of the input variables

```
Input: Hourly space heating consumption, indoor and set point temperature of
       the thermostat and historical weather of the location of the household
       during a period where the boiler is operating. At least 3 months of data
       are required.
Output: Find the optimal auto regressive orders $n_{\gamma(B)}$, $n_{\beta_t(B)}$, $n_{\beta_e(B)}$, $n_{\beta_p(B)}$,
        $n_{\phi(B)}$, $n_{\omega_h(B)}$, $n_{\omega_e(B)}$, $n_{\omega_p(B)}$; optimal number of harmonics $n_{har,az}$,
        $n_{har,el}$ and $n_{har,wd}$; and optimal smoothing parameters of the low-pass
        filter $\alpha_e$, $\alpha_s$ and $\alpha_w$
begin
    DEFINE a test set and a training set (15% and 85%, respectively);
    DEFINE a cross-validation with 8 folds from the training set. Randomly
      select, for each of the folds, a set of 80% of the days for training and 20%
      for validation;
    SET the value ranges, levels and type of variables of the parameters to
      optimize;
    DEFINE an encode–decode technique to convert each single combination of
      parameters to a Reflected Binary Code (RBC) representation, taking into
      account the allowed ranges or levels assigned to each parameter;
    INITIALIZE population with random candidate RBC representations, also
      called chromosomes;
    EVALUATE the related cost of each chromosome using Algorithm 3.3.
      In this step, $\omega$, $\beta$, $\phi$ and $\gamma$ ARX-models coefficients are estimated using
      the least squares method;
    SET $i = 1$;
    while $i \leq MaxIteration$ do
        SELECT multiple chromosomes from the last iteration, giving more
          chances to the ones with lower evaluated cost;
        RECOMBINE pairs of parents;
        MUTATE the resulting offspring in order to obtain a set of candidate
          chromosomes for this iteration;
        EVALUATE the related cost of each chromosome using Algorithm 3.3.
          In this step, $\omega$, $\beta$, $\phi$ and $\gamma$ ARX-models coefficients are estimated
          using the least squares method;
        $i = i + 1$;
    OBTAIN and decode the chromosome with the minimum cost, which con-
      tains the optimal values for $n_{\gamma(B)}$, $n_{\beta_t(B)}$, $n_{\beta_e(B)}$, $n_{\beta_p(B)}$, $n_{\phi(B)}$, $n_{\omega_h(B)}$,
      $n_{\omega_e(B)}$, $n_{\omega_p(B)}$, $n_{har,az}$, $n_{har,el}$, $n_{har,wd}$, $\alpha_e$, $\alpha_s$ and $\alpha_w$.
```

Algorithm 3.3: Cost evaluation of each chromosome

```
Input: A chromosome which contains an RBC representation; training set; test
       set; and the description of the cross-validation folds.
Output: The cost related to the input chromosome
begin
    DECODE the RBC representation to the set of parameters which represent
      the input chromosome;
    TRANSFORM the variables of the raw data set considering the decoded
      parameters. To build a data set with all the needed transformations and
      lagged variables.;
    DUPLICATE $n_f$ times the transformed data set. Each of these items will
      represent a fold of the cross-validation procedure.;
    SPLIT each fold between the training and the validation period specified in
      the input of this algorithm. This procedure aims at avoiding the models
      over-fitting. Since the folds are not randomly selected for each chromo-
      some, the likelihood of the GA to reach a global optima is greater because
      all the chromosomes are trained and validated exactly with the same folds;

    i = 1;
    while $i \leq n_f$ do
        TRAIN the supply-side model (Equation (3.2)) and the demand-side
          model (Equation (3.1)) with the training subset of the $i$th fold;
        VALIDATE the indoor temperature ($\widehat{T^i}$) and the heat consumption
          ($\widehat{\Phi^h}$) using the trained models, the validation subset of the $i_{th}$ fold,
          and the Algorithm 3.1;
        CALCULATE the cost of the $i_{th}$ fold using the Equation (3.6) and the
          validation results;
        i = i + 1
    CALCULATE the total cost of the chromosome, averaging the cost of all
      the folds;
    TRAIN the supply-side model (Equation (3.2)) and the demand-side model
      (Equation (3.1)) with the test set. These models will be an output of the
      algorithm;
    VALIDATE the indoor temperature ($\widehat{T^i}$) and the heat consumption ($\widehat{\Phi^h}$)
      using the last trained models, the test set, and the Algorithm 3.1;
    CALCULATE the cost of the test set using the Equation (3.6) and the
      validation results ;
    CALCULATE the final cost as the mean value between the average of all
      the cross-validation folds cost, and the test set cost.;
    i = i + 1
```

3.2.6 Evaluation of Potential Energy Savings

The developed methodology is suitable for multiple applications. For instance, day-ahead forecasting or Demand Response (DR) services can benefit from this methodology by including it within Model Predictive Control (MPC) procedures. To demonstrate its wider applicability, within the framework of this paper, the assessment of several set point temperature scenarios along a historical period, is performed.

These scenarios are always compared against the Business as Usual (BaU) scenario, instead of against the real data measurements. The reason is that the models errors, even if they are small, may disturb the evaluation of the estimated absolute energy differences. Therefore, it is better to compare between simulated scenarios and to obtain relative energy differences that are affected by the same error model. This strategy is supported by the fact that both the demand side and the supply model residuals fulfil the white noise requirement. The model parameters were trained using a cross-validated framework and, finally, the models were validated over a data set not seen by any of the cross-validation folds. The only requirement to assure an accurate evaluation of the relative energy differences is that the set point temperature, along the training period, should contain different temperature levels. This guarantees proper capturing of the heat dynamics of the households. Therefore, if no excitation is provided to the output variables, no dynamics can be inferred. Equation (3.7) describes the mathematical expression used to evaluate the relative energy differences between a BaU scenario, in which the set point temperature is the same as the measured one, (T^s), and another scenario under evaluation, represented by $T^{s,sim}$.

$$\Phi^h_{savings} = \frac{\sum_{t=1}^{n} \widehat{\Phi_t^h}(T^s) - \sum_{t=1}^{n} \widehat{\Phi_t^h}(T^{s,sim})}{\sum_{t=1}^{n} \widehat{\Phi_t^h}(T^s)} 100\%. \tag{3.7}$$

3.3 Case Study

3.3.1 Case Study Datasets

A real test of the whole methodology was performed over a test pilot case formed by 15 households placed in a north-western area of Spain. Each household is equipped with a condensing gas boiler which is controlled by a smart thermostat.

Both the condensing boiler and the smart thermostat, named BAXI Connect, were manufactured and provided by the company BAXI. In all cases, the distribution heating systems were based on radiators. Other building characteristics as well as occupancy patterns were not known because of data privacy requirements. Figure 3.4 shows a set of pictures of the installation process. It starts with the connection between the control board and the gateway, followed by the removal of the old thermostat and finishing with the switching on of the new smart thermostat. The smart thermostat follows the Open Therm communication protocol to communicate with the gas boiler and a wireless connection to communicate with the household router. The variables transmitted by the thermostat are: the indoor temperature, the set point temperature, the outdoor temperature (boilers equipped with an extra sensor), an indirect estimation of the space heating thermal power, and an indirect estimation of the domestic hot water thermal power. These data are communicated every 60 min and the hourly measurement tolerance corresponds to 1 kWh for the space heating and domestic hot water power and 0.5 °C for the temperature readings. The testing period started in December 2018 and finished in May 2019. However, since the involved customers had to accept the terms and conditions through the BAXI Connect mobile application, the activation was performed sequentially in time. A representative number of connected customers was not achieved until March 2019. Therefore, the analyses performed in this research are limited to this time period, from March to May. The final number of users with accurate data had to be limited to 11 households, selected among the whole population of 15 households. This reduction is due to the lack of data availability for the selected test period and due to the requirement of having a minimum level of excitation of the set point temperature. Several heating and cooling ramps were required for proper model training. Households where the set point temperature was fixed along large periods (several days, weeks, or even months) were discarded.

The IT architecture of this case study is formed by the local smart thermostat which transfers all the data to a central server managed by BAXI. These data were anonymized and communicated through a RESTful API communication layer to the big data analytics cloud. The details of this distributed and big data processing framework are described in [90].

Climate Data

Although some of the households have an extra temperature sensor placed outside the building to provide data on climate-dependent exogenous variables, the amount of gaps and outliers discarded the use of these data. As an alternative, outdoor temperature, wind speed and wind bearing data were obtained from a weather web service managed by the company Dark Sky [91]. These climate data are based on the approximate location of each household (postal code). Additionally, the global incident solar radiation on a planar surface is obtained from the Copernicus European Union's Earth observation program [92], which entails more accurate modeling the solar heat gains of the households.

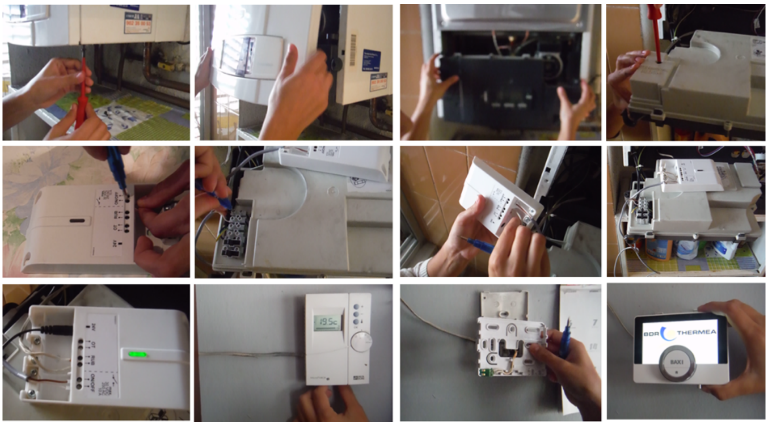


Figure 3.4: Pictures of the installation of the smart thermostat (BAXI CONNECT) in one of the case study households. The top row shows the removal of the front cover of the boiler. The middle row shows the connection with the gateway. The bottom row shows the new smart thermostat installation.

3.4 Results

3.4.1 Detailed Model Validation in One Household

In Figure 3.5, the variables used in the demand and supply-side models of one of the analyzed households are shown. The testing period ranges from 1st March to 31st May 2019. Starting from the top, the dark-green line corresponds to the set point temperature assigned by the tenant. The dark-orange line corresponds to the indoor temperature gathered by the thermostat. The violet line corresponds to the outdoor temperature gathered from dark sky web service [91]. The outdoor temperatures, ranging between 10 °C and 25 °C, are observed along the testing period. The magenta line corresponds to the boiler energy consumption. The light-green line corresponds to the direct normal incident solar radiation. The dark-yellow line represents the wind speed times. The light-brown line corresponds to the difference between the outdoor and the indoor temperatures, only if this is positive. In addition, Figure 3.6 depicts the set point and the indoor temperature for the same household, but in a shorter period. The aim is to show the correct operation of the thermostatic control.

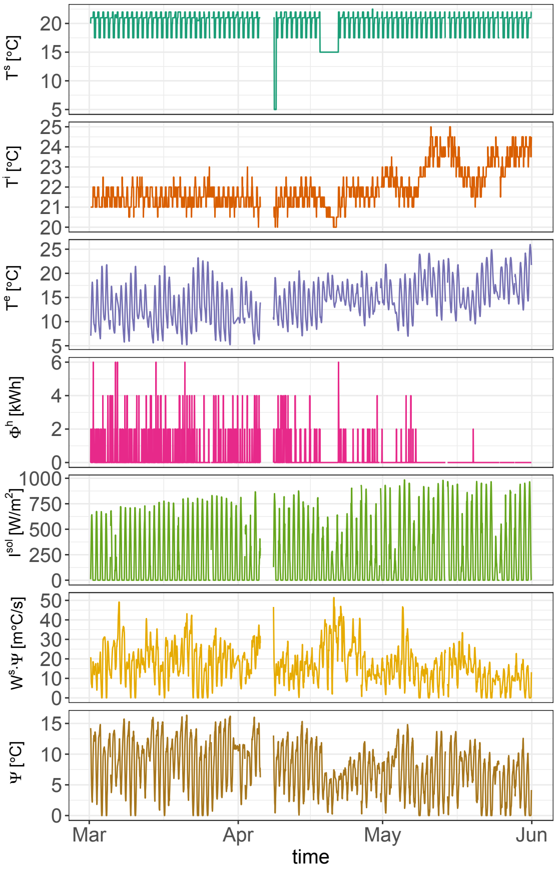


Figure 3.5: Input and output variables of the demand and supply-side models for one household.

Using these initial data sets, a cross-validation process is implemented to identify all the unknown parameters of the two models. The number of folds (n_f) considered was eight, and the percentage of training in each fold was 80%.

The ranges, and the allowed levels considered for the optimization are summarized in Table 3.2. As can be seen, most of the obtained auto-regressive orders have a maximum value of four because beyond this value their statistical significance tends to decrease. However, in the case of the indoor temperature, since this state variable is highly affected by the household thermal inertia, higher orders are permitted in the optimization (7 and 16 in the supply-side and demand-side, respectively), even the optimised values for the case study tend to range between 1 to 4. The ranges of harmonics for the Fourier series are between one and three. These ranges keep the model simple while allowing enough flexibility. To increase the chances of the GA obtaining larger values of the alphas and to address their high sensibility when they have values close to one, an exponential weighted distribution was permitted. The set of optimal parameters for the household in study is described in the last right-hand column of Table 3.2.

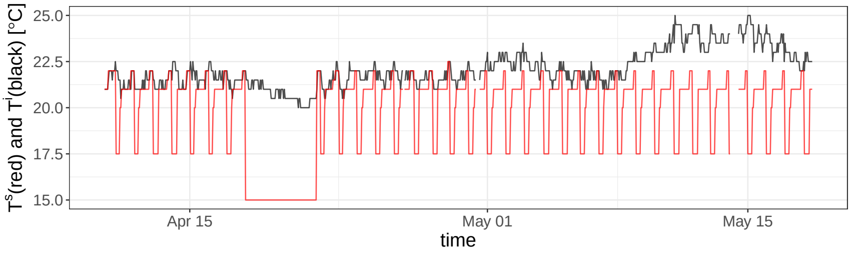


Figure 3.6: Actual set point and indoor temperature for one household.

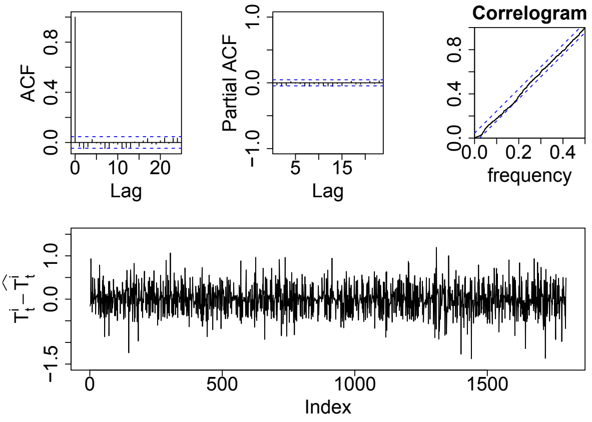


Figure 3.7: Training residuals of the model with T^i as output.

Figure 3.7 shows the residual analysis of the demand-side model of the analyzed household along the training period. As can be observed, the residuals are not auto-correlated, they follow a Gaussian distribution and the variance is homocedastic along the time. These three conditions set that the residuals are independent and identically distributed, meaning they achieve the white noise condition and the model is properly trained and considered as valid. To validate the model with new data, and therefore to assess its forecasting accuracy, the daily aggregated $MAPE$ and $RMSE$ indicators were computed. They yield values are 1.4% and 0.45 °C, respectively. These error ranges are very satisfactory and demonstrate the validity of the model for simulations of long term periods.

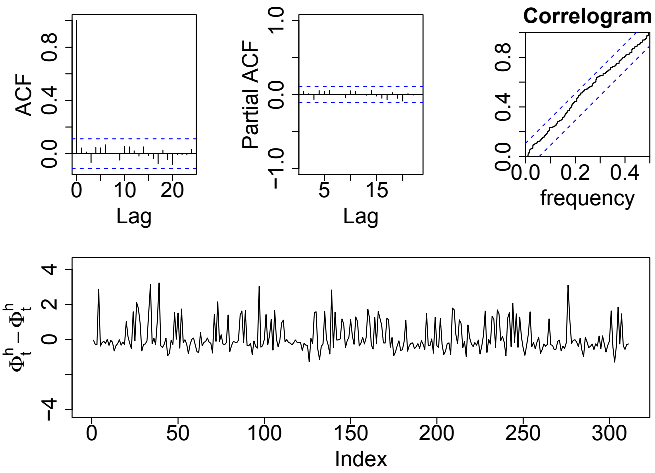


Figure 3.8: Training residuals of the model with Φ^h as output.

Figure 3.8 shows the residual analysis of the supply-side model of the analyzed household for the training data sets. As can be observed, the residuals are not auto-correlated. Even though, they do not follow a Gaussian distribution and the variance is not homocedastic along the time. Thus, the stationarity is not satisfied, which means that the residuals of the supply-side model are not fully white noise. This problem is originated during the gathering process implemented for the energy consumption monitoring using the BAXI smart thermostats. The data storage is made in hourly granularity, but using a high data resolution (2 kWh), not recommendable for the modelling of residential heating systems using complex models, such as ARX. Nevertheless, considering that, and keeping in mind that the usage of this model is coerced to large period predictions (weeks or months), the model is considered properly trained and valid. Even though, as future work of this methodology, other techniques, which may incur in simpler models, should be tested to better deal with the resolution limitation of this devices. To evaluate the accuracy of the model to assess potential energy savings, the $MAPE$ and $RMSE$ were computed at aggregated daily granularity. This is because the tolerance of the space heating consumption readings is too high in relation to the hourly space heating consumption of the households. The computed daily $MAPE$ and $RMSE$ were 37.1% and 4.72 kWh, respectively, for the testing period.

Table 3.2: Model coefficients configuration for each exogenous variable.

Parameter	Type	Values Range	Levels	Weights Distribution *	Optimal Value for Household in Study
$n_{\gamma(B)}$	integer	$\mathbb{N} \in [1, 4]$	4	uniform	1
$n_{\beta_t(B)}$	integer	$\mathbb{N} \in [0, 7]$	8	uniform	5
$n_{\beta_e(B)}$	integer	$\mathbb{N} \in [0, 3]$	4	uniform	3
$n_{\beta_p(B)}$	integer	$\mathbb{N} \in [0, 3]$	4	uniform	0
$n_{\phi(B)}$	integer	$\mathbb{N} \in [1, 16]$	16	uniform	3
$n_{\omega_h(B)}$	integer	$\mathbb{N} \in [0, 3]$	4	uniform	1
$n_{\omega_e(B)}$	integer	$\mathbb{N} \in [0, 3]$	4	uniform	1
$n_{\omega_p(B)}$	integer	$\mathbb{N} \in [0, 3]$	4	uniform	0
$n_{har,az}$	integer	$\mathbb{N} \in [1, 3]$	3	uniform	2
$n_{har,el}$	integer	$\mathbb{N} \in [1, 3]$	3	uniform	1
$n_{har,wd}$	integer	$\mathbb{N} \in [1, 3]$	3	uniform	1
α_e	float	$\mathbb{R} \in [0.00, 0.99]$	20	exponential	0.891
α_s	float	$\mathbb{R} \in [0.00, 0.70]$	14	exponential	0.252
α_w	float	$\mathbb{R} \in [0.00, 0.90]$	18	exponential	0.824
$mode_{I_{sol}}$	discrete	**	3	uniform	linear depending solar position
$mode_{W_s \times \Psi}$	discrete	***	3	uniform	linear depending wind direction
h	float	$\mathbb{R} \in [0.25, 1]$	4	uniform	0.5

Notes: * see Figure 3.9; ** no dependence, linear dependence, and linear depending solar position; *** no dependence, linear dependence, and linear depending wind direction.

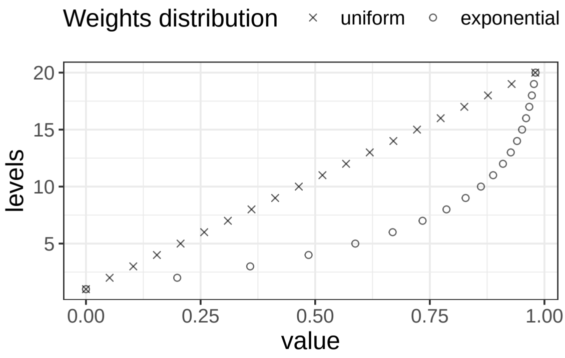


Figure 3.9: The 20 levels for a float parameter ($\mathbb{R} \in [0, 0.99]$) considering a uniform or an exponential weight distribution.

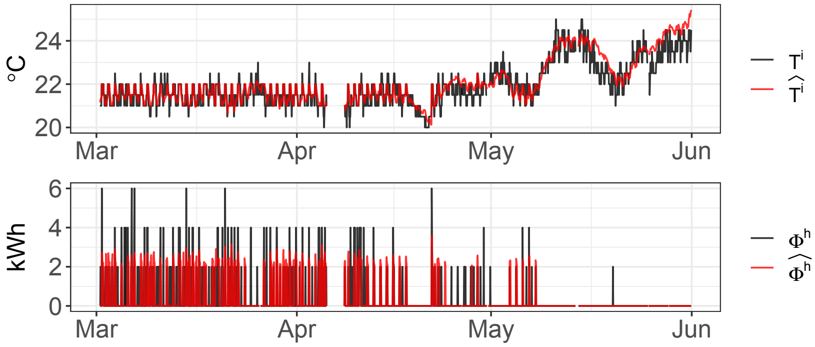


Figure 3.10: Accuracy of $\widehat{T^i}$ and $\widehat{\Phi^h}$ forecasting from 1st March to 31st May using Algorithm 3.1 and considering the real set point temperature ($T^{s,sim} = T^s$). The cumulative consumption over the period is: $\sum \Phi^h = 772$ kWh $\sum \widehat{\Phi^h} = 784$ kWh).

Once the models of the household are trained and validated, forecasts of the indoor temperature and of the space heating consumption are performed, applying the procedure defined in the Algorithm 3.1. The prediction period was between 1st March and 31st May. Figure 3.10 depicts the comparison between measured data of heat consumption Φ^h and indoor temperature T^i, with the black colored line, and the forecasted ones, $\widehat{\Phi^h}$ and $\widehat{T^i}$, with the red colored line. The set point temperature considered in the forecasting ($T^{s,sim}$) is the BaU set point (T^s), which is the original schedule set by the user. As can be seen in Figure 3.5, the set point temperature ranges from 18 °C to 22 °C along the majority of the period. Looking at the results of the simulation, it is notably appreciated that the predicted indoor temperature accurately fits the dynamics of the measured values. However, the supply-model tends to inaccurately predict some of the peaks. As previously mentioned, one of the reasons for this low accuracy is related to the high measurement tolerance of the space heating energy consumption readings. Nonetheless, it is remarkable that the whole-= period aggregated space heating energy consumption difference ($\sum \Phi^h - \sum \widehat{\Phi^h}$) is 12 kWh. That means only 1.55% over-prediction, which can be considered as a good result considering the main goal of this research.The good performance of the models in periods where the household behaves in free-floating mode (with the boiler switched off) is also remarkable.

3.4.2 Model Validation in a Larger Population of Households

The training and validation framework was applied over a set of households, 11 households, with available space heating energy consumption for the period winter–spring 2019. Instead of showing a residual analysis for each of them, two Goodness Of Fit (GOF) indicators were considered. To determine the models parameters of each household, a cross-validation procedure and two prediction strategies were followed. The testing period comprised three months (1st March–31st May). The first strategy was a one-step ahead prediction of each of the models in order to see how the prediction fit the actual data according to an hourly update of the data inputs. This could be understood as the training error of each model. Following this strategy, no error propagation was considered. The second prediction strategy consisted of following the Algorithm 3.1. In this case, in addition to the trained models, the BaU set point temperature and the historical external weather conditions were also considered. The initial conditions for the indoor temperature and space heating initialization were those of 1st March at 00:00:00. Using the second strategy, the error propagation was considered. If the models did not properly characterize the dynamics, the GOF indicators would dramatically increase when compared to the one-step ahead prediction strategy. Both strategies were confronted with the monitored data gathered by the smart thermostat. The GOF indicators were the Mean Absolute Percentage Error ($MAPE$) and the Coefficient of Variation of the Root Mean Square Error ($CVRMSE$). Equations (3.8) and (3.9) describe their mathematical expressions. In these equations, n corresponds to the number of time steps of the whole period, y_t is the measured time series and $\widehat{y_t}$ is the predicted time series.

$$MAPE = \sum_{i=1}^{n} \left| \frac{y_t - \widehat{y_t}}{y_t} \right| \tag{3.8}$$

$$CVRMSE = \frac{1}{\overline{y_t}} \cdot \sqrt{\sum_{i=1}^{n} \frac{(y_t - \widehat{y_t})^2}{n}} \tag{3.9}$$

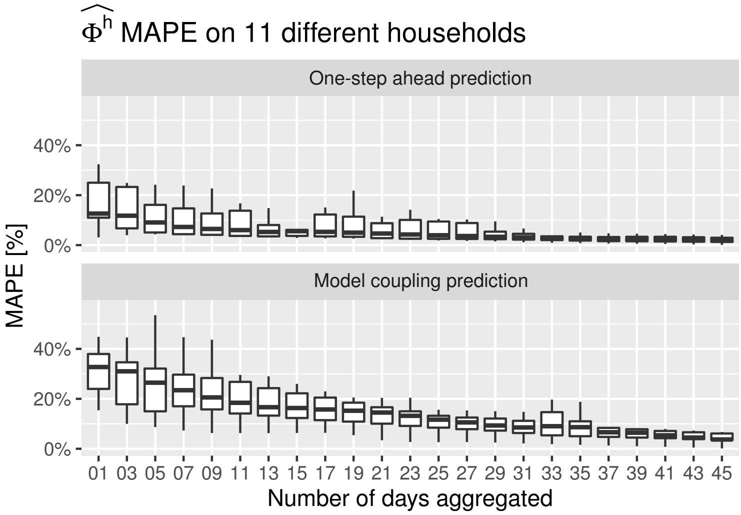


Figure 3.11: MAPE of the space heating consumption of 11 households, aggregating data to daily multiples.

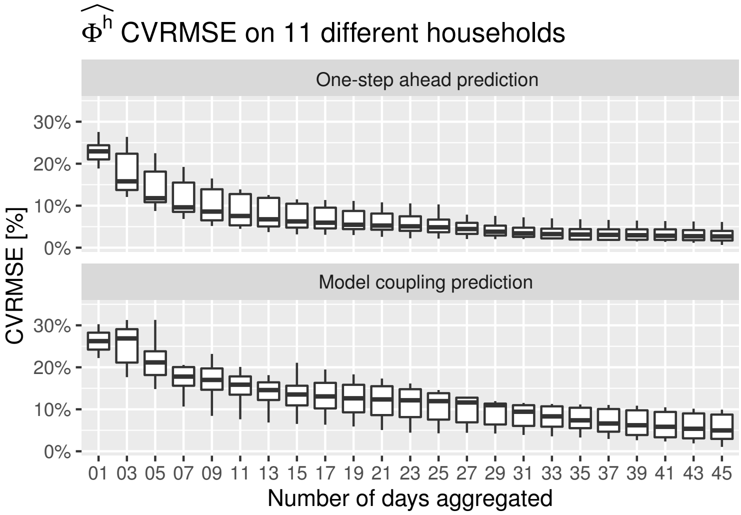


Figure 3.12: CVRMSE of the space heating consumption of 11 households aggregating data to daily multiples.

In Figures 3.11 and 3.12, box-plots of the $MAPE$ and $CVRMSE$ for the space heating energy consumption are, respectively, shown for the 11 households, considering both prediction strategies and the same testing period. The $MAPE$ is only computed when $y_t > 0$. As can be seen, the data are aggregated to several days to understand what the minimum period required to perform this kind of analysis is 30 days. In both, $MAPE$ and $CVRMSE$, the errors evolved similarly, decreasing asymptotically as the aggregation frequency increased. When aggregation periods larger than 30 days are considered, both errors have an average value of less than 10%. Therefore, a minimum period of a month is recommended for the assessment of energy savings scenarios. It can also be concluded that both models are able to correctly characterize the dynamics of households since the error propagation does not increase sharply between both prediction strategies.

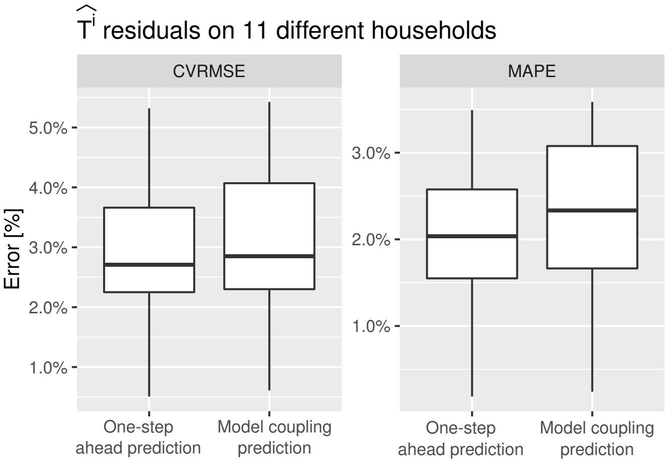


Figure 3.13: Hourly indoor temperature CVRMSE and MAPE of 11 households.

The box-plots of the $MAPE$ and the $CVRMSE$ of the indoor temperature are shown in Figure 3.13. The high accuracy of the demand-side model results in a very well predicted indoor temperature using both strategies. The hourly frequency residuals are lower than 3% on average. This means that the model is capable of accurately modeling the dynamics of the thermal losses and heat gains. This is of high importance since the indoor temperature is the variable used to control the operation of the space heating system of the households.

3.4.3 Assessment of Potential Energy Savings

To envisage a wider applicability of the data-driven techniques developed in this research, the potential energy savings of several set point temperature scenarios over the analyzed household of Figure 3.5 are shown in Figure 3.14. As can be seen, in the period between 1st March and 31st May, the space heating energy consumption is estimated to decrease around 24% if the BaU set point schedule is lowered by 1 °C. If this BaU schedule is lowered by 2 °C, the estimated space heating energy savings reaches around 49%.

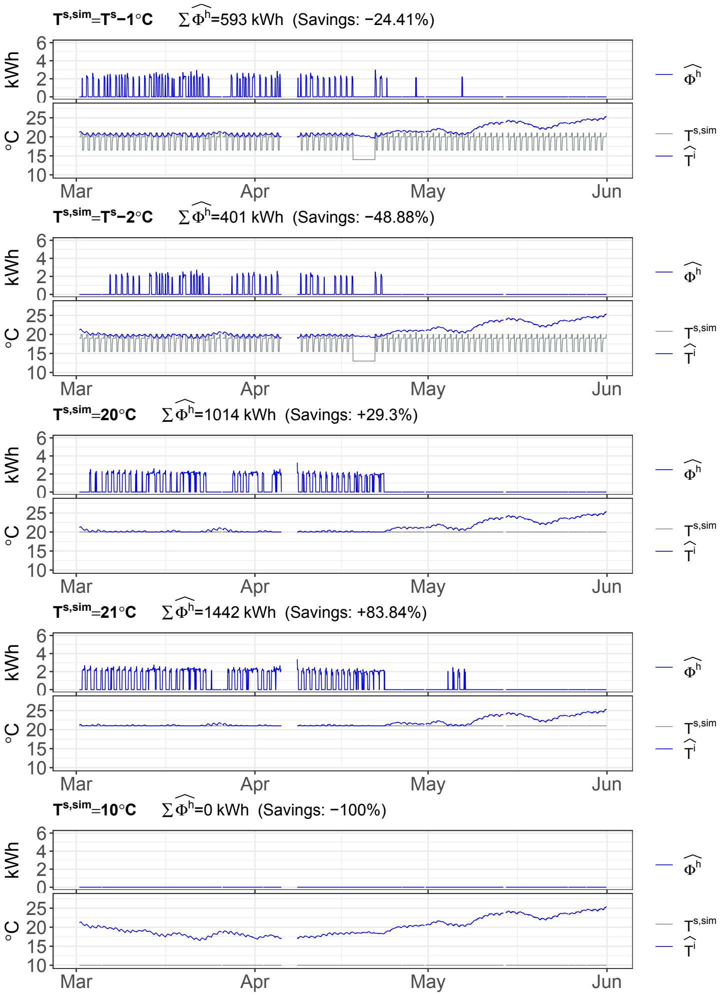


Figure 3.14: Energy potential savings applying different set point temperatures scenarios ($T^{s,sim}$) over one household.

These figures should be considered as approximate since the supply model has a higher error validation than the demand side model. This data-driven methodology also allows the assessment of the response of the space heating energy consumption to time fixed values of the indoor temperature, as recommended by the building code regulations. Two predictions of fixed set point temperatures of 20 °C and 21 °C are generated and shown in the last top down plots of Figure 3.14.

Although some energy savings was expected, the outcomes of these simulations yielded approximate energy consumption increases of around 29% and 84%, respectively. Therefore, for this household, it is not recommended to fix the temperature along the whole period, since this would lead to higher space heating energy consumption. This conclusion is in line with the set point temperature schedules set by the user in the BaU scenario, where 30 to 40% of the hours are set to low set point temperature. These low set point temperature periods correspond to the night time, when no energy gains are present and the outdoor temperatures are lower. In other words, even in case the values of the fixed set point temperature scenarios are lower than the higher values of the set point temperatures in the BaU scenario, the energy demand of the household increases to avoid the drop in the indoor temperature during this non-operational period.

Finally, the free-floating conditions can be also assessed. This facilitates the estimation of the minimum indoor temperature that a household would reach without the operation of the space heating system. In this case, this household would reach a minimum temperature of 16.7 °C along the whole tested period.

An assessment of approximate potential energy savings of the 11 households was performed. Figure 3.15 presents the box-plot of the energy consumption difference between the measured data and the simulated space heating energy consumption over the period between 1st March and 31st May. The first column represents the measured space heating energy consumption versus the predicted one, obtained by applying the Algorithm 3.1 considering the BaU set point temperature. The average difference is −2%, with an interquartile range between 4 and −10%. Since the absolute error is lower than 10%, it can be concluded that the trained models for the 11 households are valid to simulate set point temperatures scenarios. In the second and third columns, the comparison between two set point temperatures scenarios and the BaU set point temperature scenario is shown. Decreases of 1 °C and 2 °C are tested, yielding potential average energy savings of 18.1 and 36.5%, respectively.

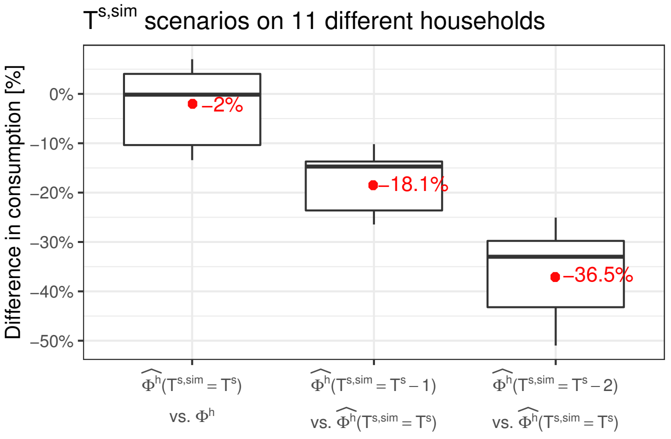


Figure 3.15: Energy potential savings distributions applying two different set point temperatures scenarios ($T^{s,sim}$) over 11 households.

3.5 Conclusions

The study demonstrates high accuracy of the models to predict both the indoor temperature and the space heating energy consumption. However, for this specific use case, since the measurement resolution for the space heating consumption is too high, a minimum aggregated period of 30 days is recommended to properly estimate the potential energy savings scenarios.

The major novelty of the proposed methodology is that it goes beyond the prediction of the heat consumption and the indoor temperature of these systems. The methodology incorporates an optimization algorithm and a control loop which provides the capability to virtually mimic all the possible user controlled modes driven by the set point temperature.

Some direct conclusions can be finally obtained in relation to the potential energy savings which can be achieved if the users decide to modify their usual set point temperature schedule. Average estimated energy savings of 18.1% can be achieved if the usual set point temperature is lowered by 1 °C. Up to approximately 36.5% energy savings can be achieved if the usual set point temperature is lowered by 2 °C.

Chapter 4

Operation and flexibility assessment of direct load control systems in buildings

This chapter has been published as a paper:

Mor, G.; Cipriano, J.; Grillone, B.; Amblard, F.; Menon, R.P.; Page, J.; Brennenstuhl, M.; Pietruschka, D.; Baumer, R.; Eicker, U. Operation and energy flexibility evaluation of direct load controlled buildings equipped with heat pumps, Energy and Buildings 2021, https://doi.org/10.1016/j.enbuild.2021.111484

4.1 Introduction

Renewable energy sources like solar panels and wind turbines are invaluable for transitioning to a fossil-free energy system to mitigate climate change impacts. However, their natural fluctuations introduce significant uncertainty in the power grid. In addition, they transform the present unidirectional centralized system into a bi-directional decentralized system with smaller units and multiple prosumers, increasing the difficulty to achieve power balance [93]. This leads to an increased need for flexibility on the demand side [94, 95] and for new storage capacity [96, 97]. One attractive solution identified to support the transition of power systems is to manage not only the energy supply but also the demand via Demand Response (DR) programs [98, 99]. The principle behind it is to use various economic incentives to shift the electrical loads of end-use customers from times with a high wholesale market price or when the system's security is threatened to other time periods. As has been pointed out in [100], there are predominantly two types of DR programs: i) explicit DR (also called incentive-based); ii) and implicit DR (also called price-based). In Implicit DR, a price signal is sent to the prosumers to motivate their user behaviour change. Explicit DR involves the participation of a third party, who takes action on behalf of a customer by sending an activation signal such that the system behaviours are directly modified. In both DR programs, and considering that nearly 50 % of the total energy consumption of buildings comes from Space Heating (SH)/Cooling (SC) and domestic hot water (DHW), as stated by [101], there is definitely a role that electrically driven Heating, Ventilation and Air Conditioning (HVAC) systems can play.

Although the installation of control devices, communication, control protocols and standardization have improved, DR is currently still rarely implemented in the commercial and even less in the residential sector in Europe [102]. Serale et al. [103] reviewed 161 papers on Model Predictive Control (MPC) in buildings, and revealed that only a fourth considered residential buildings and only a bit more than a fifth compared experimental cases to simulated cases. Kohlhepp et al. [104] performed a thorough review of 16 projects of field tests and demonstrations of applied DR from around the world. Only four projects had more than 100 households, a size large enough to represent load diversity and test resource competition. A singular case of commercially applied DR to large scale residential buildings is run by the French company Voltalis, which manages one of the biggest portfolios of explicit DR services in the world. They follow a strategy of DR based on service curves [105, 106].To our knowledge, they have not published peer-reviewed papers analysing the impacts of this DR strategy or provided a general methodology to evaluate the delivered energy flexibility. In general, there is a lack of test case benchmarks. Comparing the results among case studies with different goals, addressed electricity markets and technology environments is still very challenging.

The few real case applications of DR have brought forth a wide diversity of methodologies to evaluate the energy flexibility that individual or clustered buildings can provide. In many cases, assessment methodologies are focused on the potential energy flexibility at the building design stage. Arteconit et. al [107] is a clear example of defining an indicator of flexibility labelling at the design stage. Finck et. al [108] performed a very detailed analysis of the demand flexibility that power-to-heat systems can deliver. Several flexibility indicators such as available storage capacity and efficiency are enhanced with a flexibility factor, which relates electricity costs in the lower price and higher price periods in a day-ahead electricity market DR scenario. A thermal instantaneous power flexibility indicator is also described. These indicators have a great potential to evaluate the energy flexibility in DR services addressing the ancillary markets. The only weakness is that they were demonstrated in a theoretical simulated environment. Moreover, that research was more focused on developing control strategies and not on the flexibility evaluation itself.

In their hands-on review, Reynders et al. [109] made a valuable contribution in reviewing prior research dealing with definitions and quantification of energy flexibility. One of their main conclusions was that a large share of the performed research practices did not explicitly define or were not focused on quantifying energy flexibility. Yet, they dealt with the development of control strategies and algorithms for specific case studies. They also stated that most of the studies had in common the identification of three general properties of energy flexibility: i) the potential flexibility in several time horizons; ii) the load which can be shifted; and iii) the cost of this flexibility. The authors also deducted that methodologies aimed at quantifying the energy flexibility by analyzing triggered events at specific times have greater strengths when dealing with the flexibility to be delivered by the thermal mass of buildings or energy storage systems. In contrast, methodologies which relied on differences in the accumulated energy profiles are difficult to interpret because they treat systems driven by multiple time constants as a single state system. El Geneidy and Howard [110] performed a detailed analysis of the categories of characteristics that constrain the contracted flexibility potential in homes. Although their results are valuable for defining further DR strategies, they are limited by simplified assumptions and exclusively based on simulated scenarios. Bampoulas et al. [111] conducted a more detailed recent review on studies aiming at defining suitable flexibility indicators. They highlighted that most of these studies were limited to evaluating control strategies and assessing the activating and deactivating of the building's thermal mass. Still, they did not clearly quantify the flexibility potential of HVAC systems.

Following these remarks, Junker et al. [112] developed a novel methodology to characterize the energy flexibility as a dynamic function named the Flexibility Function (FF). This FF enables a Flexibility Index, which describes how a building can respond to certain activation signals. The FF is a step-response function that assumes that the relation between the penalty signal and the power load is linear and time-invariant. Several theoretical cases were presented to validate this proposed FF, demonstrating how the FF enables the quantification of the energy flexibility in different types of buildings. This paper represents a valuable contribution to the field since it establishes a robust methodology to represent, in a normalized manner, the correlation between the penalty signal and the load response. The concept of the FF applies to several building typologies and DR scenarios but specifically addresses implicit DR services. However, the assumption that the dependence of the active power and the activation variable is linear limits its applicability to DR services which can fulfil this requirement. Recently, Junker et al. [113] published a paper presenting a new generic method capable of overcoming the linearity and time dependency of the correlation between the flexibility and the penalty signal. This new method follows the principles of the FF, but it changes the perspective. They developed a non-linear dynamic model based on stochastic differential equations. It is applied to price-based controlled buildings and water towers, showing high robustness, accuracy and scalability to similar business cases. One limitation is that these methods are developed to specifically address implicit DR services driven by penalty signals triggered by one of the stakeholders of the electricity sector. This is very common in many electricity markets, such as the spot electricity market, the intra-day market, or certain ancillary services markets. However, in some explicit DR services, where the activation variable is a power trace to be followed, such as when a commercial aggregator makes bilateral agreements with their Balance Responsive Parties (BRP), both the FF and the flexibility characterization model defined in [112, 113] need to be modified or extended to adapt them to these different kinds of activation variables.

In our research, an extension of the previously developed flexibility characterization procedures is performed, which is the main novelty of the research work. Based on the background knowledge developed by Junker et al. [112], and further improved in [113], new linear regression-based models, designed to characterize the energy flexibility delivered by blocks of buildings, are developed and validated in real cases. These new flexibility models address different implicit and explicit DR scenarios. For example, the activation variable can be the spot market price, the percentage of power to be activated, or a power trace to be tracked. This is also an extra contribution to the paper. One last novelty of the research lies in the fact of developing and applying these flexibility models on clustered residential buildings, ranging from high energy performance detached houses (Germany) to building blocks connected to low-temperature district heating (Switzerland) or a group of buildings formed by small shops, a food market and residential units (Spain). In all the scenarios, the methods were applied to remote-controlled heat pumps with different system configurations.

The rest of the paper is organized as follows. Section 4.2 describes the developed methodology, identifies potential flexibility markets, presents a common methodology for quantifying energy flexibility and describes the models and the new FF formulations. Hereinafter, the three case studies (Spain, Germany and Switzerland) are presented in Section 4.3. They comprise three clusters of buildings with heat pumps remotely driven by MPC procedures. The operation of the DR services and the results are summarized in Section 4.4, where details of the outcomes of the different direct load control tests are presented. The energy flexibility is assessed, through the derived Flexibility models and Flexibility Functions, in this section. Finally, the findings are extensively discussed in Section 4.5, and summarised in Section 4.6.

4.2 Methodology

4.2.1 Identification of the addressed flexibility markets

Different markets exist for the trading of electricity between buyers and sellers. In the day-ahead market, products are traded for delivery on the following day. The intraday market trades products to balance possible deviations from the day-ahead forecast. Balancing or control reserves markets are needed to balance electricity generation and consumption in the short term. Three different types of control reserves markets are available: i) Frequency Containment Reserve (FCR), ii) Automatic Frequency Restoration Reserve (aFRR), and iii) Manual Frequency Restoration Reserve (mFRR). They differ according to the principle of activation, to their bid minimum size and symmetry, and their activation speed. The last category of markets is the Reserve Replacement (RR) market. These capacity mechanisms aim at ensuring the security of supply from a long-term perspective.

In this paper, four of the above-mentioned markets are selected to be addressed through direct load control DR services: i) the Spanish wholesale electricity market (day-ahead); ii) the German operating reserve; iii) the German intraday spot market and; iv) the Swiss imbalance market (aFRR).

In Spain, OMIE is the nominated electricity market operator (NEMO) for managing the Iberian Peninsula's day-ahead and intraday electricity markets. The delivery takes place on the day after the trading day (incl. weekends or holidays), and trading sessions take place in one daily auction 365 days/year. Sale and purchase bids can be made considering between 1 and 25 energy blocks in each hour, with power and prices offered in each block. In the case of sales, the bid price increases with the block number; in purchases, the bid price decreases with the block number. The minimum size is 0.1 MW. The Spanish TSO, Red Eléctrica Española, has developed an information system known as 'System Operator Information System (esios)', specially designed to run all the necessary processes to ensure economic and reliable exploitation of the Spanish Power System in real-time. The esios portal offers an open API where the wholesale electricity prices for the next 24 h are published once the spot market is closed (at 13 h of every day). These electricity prices become the control variable for the direct load control services implemented in the Spanish use case.

Unlike the day-ahead spot market in Germany and Switzerland, the intraday market can be described as a corrector market because the time intervals between trade and activation and the activation period are significantly lower. Thereby, electrical energy is traded in intervals of one hour for Switzerland or 15 min for Germany. In Germany, trades for 15 min intervals can be completed between 15:00 (CET) of the previous day until 5 min before activation [114].

In Germany, four different TSOs are responsible for the reserve markets, and around 60 companies are pre-qualified to deliver operating reserves. Therefore, compared to the spot trade market, there is a highly reduced field of actors. The FCR activation time of a few seconds is very short term. aFRR requires an activation time of less than 30 s and 5 min to reach full power. RR requires 5 min for activation. mFRR and RR are traded daily and bids can be provided in blocks of 4 h. Negative and positive reserve power is traded. As a first instance, positive or negative power is offered with different assigned prices. If an offer is accepted, a working price (e.g. EUR/MWh) is also offered, and the activation occurs according to the working price within a merit order list. The main drawbacks of mFRR and RR are that at least 1 MW of power must be certified. Thereby, an aggregated larger pool operation is necessary. The German operating reserve market, especially mFRR has seen dropping costs within the last years [115, 116], whereas in comparison the amount of energy traded at the EPEX Intraday market has almost doubled from 2014 – 2019 (from 47 TWh to 91.6 TWh) [117], shifting the favourability more to intraday trade. In the German pilot site, activations were carried out by Centrica, an aggregator company situated in Belgium, according to available market data from Belgium. This is justifiable due to the fact, that the spot market products are tradable in between Germany and Belgium [118] as well as the operating reserve market conditions are comparable [119].

The Swiss operating reserve markets are managed only by one TSO (Swissgrid). Compared to Germany, the minimum certified bid of the aFRR and RR markets is 5 MW, making them even less accessible for residential buildings, as a vast pool of assets would be needed. For aFRR, the trading is automated. The products traded are asymmetric and must be available 30 s to 5 min after the notification for a duration of up to 15 min. The size of the aFRR in Switzerland in 2017 was ±380 MW [120]. The high participation of hydropower supply in the reserve markets limits residential DR. The high number of DSOs present in Switzerland [120], each of them with a limited asset pool, also hinders the development of DR services by the DSOs. For the field tests of the Swiss pilot site, the targeted reserve market was the aFRR, as its market constraints are the most accessible for heat pumps. By combining a pool of batteries with fast activation time and heat pumps whose power availability lasts longer, an aggregator could theoretically fulfil the market constraints. The trading in this work was done by Centrica (aggregator), and HES-SO Valais-Wallis carried out the activations in Switzerland. Real trading could not be tested, as it would have required 200 times the capacity offered by the pilot site to reach the minimum bid of 5 MW.

4.2.2 New reference methodology to assess energy flexibility

The methodology to characterize the energy flexibility in a more standardized way follows the initial methodology set out by [112]. This methodology defines a dynamic function, named the flexibility function FF, which characterizes the energy flexibility of any device through the use of penalty signals. In our research, the analysed use cases do not strictly follow the activation of the energy flexibility through penalty signals since they respond to other DR schemes. To address these different DR schemes, we took a broader approach than [112] and implemented a methodology to include other kind of signals and activation variables which are more realistic for the analysed energy flexibility markets. The proposed methodology follows the process shown in Fig. 4.1. As can be seen, the initial point starts with setting up the baseline modelling, which corresponds to the energy performance model of the buildings in a Business as Usual (BaU) scenario. This baseline model is then used to forecast the building energy consumption for the time horizon defined by the activation period. This energy forecasting is integrated into a model predictive control optimization where the activation variable is the output. The cost function depends on the flexible electricity market to be addressed. The activation period is different for each use case and flexible electricity market. It is driven by the optimized activation variable, ranging from a penalty signal, such as the day-ahead price, a percentage of power activation time, or a power trace to be tracked. The active power consumed throughout of activation period is registered. This time series is considered as the dependent variable within the flexibility model. The baseline forecasting and the activation variable time series are defined as the independent variables. The flexibility model is then formulated also to include the corresponding autoregressive terms. The next step consists of training this flexibility model with historical data of the activation period. Once the flexibility model is trained and validated, the i-step prediction is used to define the flexibility function, FF.

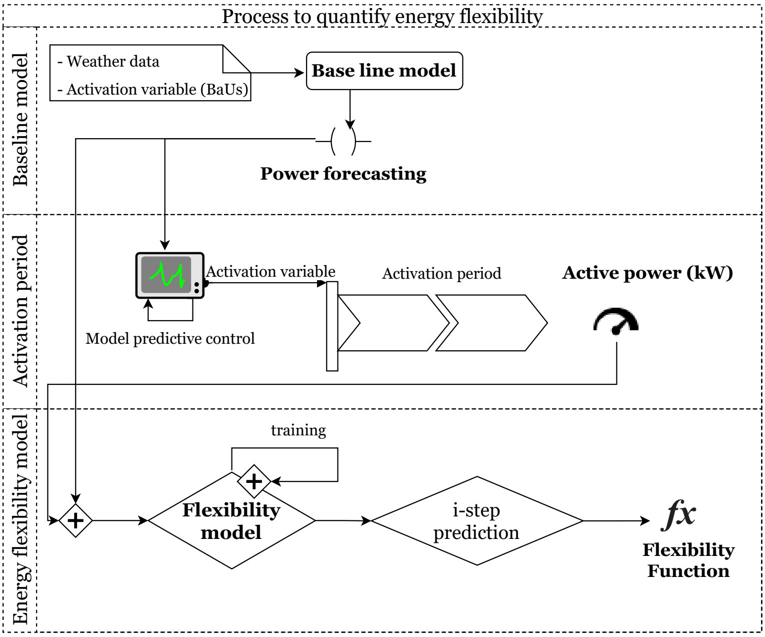


Figure 4.1: The general process to quantify the energy flexibility

.

Baseline modelling

Since the energy flexibility cannot be directly measured, as it represents the activation or deactivation of power usage, it is determined by comparing measured power during the activation period and forecasting the power consumed by the building as if the activation had not taken place. This supposed scenario is called the Business as Usual (BaU) scenario. To determine the energy load forecasting under the BaU scenario, a model of the thermal dynamics and the energy consumption of the building, prior to the activation period, needs to be developed. This model is called the baseline model. The baseline model can be defined as the energy characterization of the starting situation and has a fundamental role in the determination of energy flexibility. In fact, the baseline model allows isolating the effects of the activation variables from the effects of other parameters that can simultaneously affect the energy consumption. To obtain the baseline model, several approaches can be followed:

- Empirical modelling based on a system of differential equations and heat transfer functions
- Grey box modelling based on state-space models
- Data-driven modelling based on transfer function models or machine learning techniques

In this research, the three approaches have been used for the different use cases. The first approach requires detailed models with several monitored variables and a calibration stage to fit with the monitored data. An example of these kinds of calibration processes can be found in [121]. The second approach requires monitoring the state variable (indoor temperature or water tank temperature) and a precise process to identify the unknown parameters. A[122, 88] detailed description of the identification procedure applied over suitable grey box building heat dynamics models is presented. The third approach requires good data quality of a minimum historical period and the measurement of the control variable. Several authors applied this last approach to determine the heat dynamics of buildings. In [123], some of the most common data-driven methods used to develop baseline models are reviewed. The baseline models developed in each use case are described in detail and referenced in the corresponding subsection of Section 4.3 of this document.

Flexibility models

A flexibility model is a regression-based model which aims at finding the correlation among the active power, the activation variable and the power under the BaU scenario. In this research, a data-driven approach is followed based on Autoregressive (AR) models. As previously mentioned, the initial modelling technique is defined by [112] is modified to adapt it to the specific constraints of the different activation variables. The original model by Junker et al. assumes that the active load when exposed to a the penalty signal can be separated into two parts; the load that dynamically responds to the the the penalty, and the non-responsive load (baseload power, in our equations). However, in our case the dynamics due to the active and baseload signal itself are added. Thus, an ARX model is considered based on the initial equation presented in [112]. This allows to better estimate the amount of time and load that can be flipped once an activation, a change of price, or a trace to follow is received. Additionally, it helps to the proper estimation of the rebound effect caused by a change in the penalty, as it considers the thermal inertia available in the system.

In the use cases when the activation variable is the day-ahead electricity price of the wholesale spot market, the model formula is described in Eq. 4.1.

$$\phi_{T_o}\big(B\big)P_t^e = \omega_{T_o}\big(B\big)P_t^b + \Psi_{T_o}\big(B\big)DA_t + \varepsilon_t \tag{4.1}$$

P_t^e is the active power of the system, P_t^b is the predicted baseline power without activations, and DA_t corresponds to the activation variable, the day-ahead electricity price. $\phi_{T_o}(B)$, $\omega_{T_o}(B)$ and $\Psi_{T_o}(B)$ are the parameters of the model. The sub index T_o represents their dependence with one categorical variable, the outdoor temperature. In order to better express this dependency, a 4 hours moving-averaged transformation is applied over the outdoor temperature for the testing periods. This averaged temperature is further split in two levels: [6.67 °C - 12.3 °C] and [12.3 °C - 21.5 °C]. Therefore, the T_o is not used as a exogenous variable of the model. The backward shift operators,B, are defined as $B^k y_t = y_{t-k}$, where y_t is the considered variable (P_t^e, P_t^b,DA_t) at time t and $k \in [0, j]$. Here, j refers to the maximum order allowed to that backward shift operator, B. The ε_t corresponds to the white noise residual of the model at time t.

In the use cases when the activation variable is the percentage of activation time within each time step, the model formula is described in Eq. 4.2.

$$\phi_{bd}\big(B\big)P_t^e = \omega_{bd}\big(B\big)P_t^b + \Psi_{bd}\big(B\big)A_t + \varepsilon_t \tag{4.2}$$

P_t^e is the active power of the system, P_t^b is the predicted baseline power without activations, and A_t corresponds to the activation variable, which is the percentage of time asked for activation within every time step [0%-100%]. $\phi_{bd}(B)$, $\omega_{bd}(B)$ and $\Psi_{bd}(B)$ are the parameters of the model. The sub index *bd* represents their dependence with one categorical variable, the building number. *bd* comprises the categorical values of the building number for this use case [20, 22, 24, 25], and a virtual building that aggregates the power of all of them. Therefore, the building number, *bd*, is not used as a exogenous variable of the model. The backward shift operators,B, are defined as $B^k y_t = y_{t-k}$, where y_t is the considered variable (P_t^e, P_t^b,A_t) at time t and $k \in [0, j]$. Here, j refers to the maximum order allowed to that backward shift operator, B. The ε_t corresponds to the white noise residual of the model at time t.

In the use cases when the activation variable is a trace to be tracked, the power used within the activation period is no longer the model's dependent variable. In Eq. 4.3, the dependent variable is substituted by the difference between the active power, P_t^e, and the baseline power,P_t^b. Whereas, the independent variable of the model corresponds to the difference between the power trace to be tracked, P_t^f and the baseline power, P_t^b. The modified formula is shown in Eq. 4.3.

$$\phi_{s,T_o}(B)\left(P_t^e - P_t^b\right) = \omega_{T_o}(B)\left(P_t^f - P_t^b\right) + \varepsilon_t \tag{4.3}$$

$\phi_{s,T_o}(B)$, $\omega_{T_o}(B)$ are the parameters of the model. The sub-index T_o represents their dependence on a categorical variable, the outdoor temperature. Based on a 4 hours moving-averaged transformation of the outdoor temperature, for the test periods, the results are split into two groups of outdoor temperature levels: [6.5 °C - 15.7 °C] and [15.7 °C - 28.5 °C]. The sub-index s refers to the sign of the trace to be tracked in relation to the baseline power, being equal to 1 when it is positive, equal to 0 when there is no difference with the baseline power, and equal to -1 when it is negative. Therefore, neither the T_o nor the s are used as exogenous variables of the model. The backward shift operators,B, are defined as $B^k y_t = y_{t-k}$, where y_t is the considered variable (P_t^e, P_t^b,X_t) at time t and $k\epsilon[0, j]$. Here, j refers to the maximum order allowed to that backward shift operator, B. The ε_t corresponds to the white noise residual of the model at time t.

Flexibility Functions

The Flexibility Function (FF) can be understood as the impulse response function of each flexibility model since the flexibility models include autoregressive terms of the dependent variables, which cause an influence over the P_t^e when $t \geq 1$.

To do so, an i-step prediction is performed to estimate the impulse response of the models properly.

In the use case when the activation variable is the day-ahead price, the FF is determined based on a positive and a negative change in the day-ahead electricity price (±0.1 €/kWh) for the time steps $n = 15$, 60 and 120 minutes and for a flexibility evaluation period of $i = 480$ minutes. When the activation variable is the percentage of time of activation within each time step, 100 % activation signals for time steps of $n = 1$, 2 and 4 hours are tested along a flexibility evaluation period of $i = 12$ hours. Both use cases follow a similar procedure to determine the FF:

$$t = \big(0, 1, ..., i\big) \tag{4.4a}$$

$$P^e_{t\leq 0} = 0 \tag{4.4b}$$

$$P^b_{t\in\mathbb{N}} = 0 \tag{4.4c}$$

For the day-ahead electricity price as the activation variable:

$$DA = \begin{cases} 0.1, & \text{if positive price change} \\ -0.1 & \text{if negative price change} \end{cases} \tag{4.4d}$$

$$DA_t = \big(\, 0, \underbrace{(DA, .., DA)}_{n\ times}, \underbrace{(0, ..., 0)}_{n\text{-}i\ times} \,\big) \tag{4.4e}$$

$$\phi_{T_o,k=0}\big(B\big)P^e_t = -\phi_{T_o,k\geq 1}\big(B\big)P^e_t + \Psi_{T_o}\big(B\big)DA_t \tag{4.4f}$$

$$\phi_{T_o,k=0}\big(B\big) = 1 \tag{4.4g}$$

$$FF_t = P^e_t = -\phi_{T_o,k\geq 1}\big(B\big)P^e_t + \Psi_{T_o}\big(B\big)DA_t \tag{4.4h}$$

For the percentage of time activation within a time step as the activation variable:

$$A_t = \big(\, 0, \underbrace{(100, .., 100)}_{n\ times}, \underbrace{(0, ..., 0)}_{n\text{-}i\ times} \,\big) \tag{4.4i}$$

$$FF_t = P^e_t = -\phi_{bd,k\geq 1}\big(B\big)P^e_t + \Psi_{bd}\big(B\big)A_t \tag{4.4j}$$

In the use case when the activation variable is a trace to be tracked, the FF is determined by considering a 100 % activation signal of time steps $n = 15$, 30 and 60 minutes for a flexibility evaluation period of $i = 120$ minutes. A multi-step prediction method is used to predict the expected response of ±1 kW of the trace to be tracked. The previous estimate of the flexibility function, $(P^e - P^b)$, is used for the new prediction step. The baseline power is set to $P^b_t = 0$ for $t \in (0, 1, ..., i)$. Here, s is equal to 1 if the activation is pos equal to -1 if it is negative.

$$\left(P^e_{t\leq 0} - P^b_{t\leq 0}\right) = 0 \tag{4.5a}$$

$$\left(P^f_t - P^b_t\right) = \begin{cases} 0, & \text{if } t \leq 0 \\ \left(\underbrace{(s, ..., s)}_{n\ times}, \underbrace{(0, ..., 0)}_{i\text{-}n\ times}\right), & \text{otherwise} \end{cases} \tag{4.5b}$$

$$\phi_{s,T_o}(B)\left(P^e_t - P^b_t\right) = \omega_{T_o}(B)\left(P^f_t - P^b_t\right) \tag{4.5c}$$

$$\phi_{s,T_o\,k=0}(B)\left(P^e_t - P^b_t\right) = -\phi_{s,T_o\,k\geq 1}(B)\left(P^e_t - P^b_t\right) + \omega_{T_o}(B)\left(P^f_t - P^b_t\right) \tag{4.5d}$$

$$\phi_{s,T_o\,k=0}(B) = 1 \tag{4.5e}$$

Considering the flexibility model of Equation 4.3 and the set up described in previous equations, the FF is defined as:

$$\begin{aligned} FF_t &= \left(P^e_t - P^b_t\right) \\ &= -\phi_{s,T_o\,k\geq 1}(B)\left(P^e_t - P^b_t\right) + \omega_{T_o}(B)\left(P^f_t - P^b_t\right) \end{aligned} \tag{4.6}$$

4.3 Case studies

The methodology to evaluate the energy flexibility is applied over three case studies which have in common a direct load control of space heating systems driven by heat pumps:

- Case study of the Spanish wholesale electricity market price as the activation variable. Blocks of buildings placed in North-East Spain (Sant Cugat)
- Case study of the percentage of activation time as the activation variable. Residential households placed in South Germany (Wüstenrot)
- Case study of a trace to be tracked as the activation variable. Blocks of residential buildings placed in Switzerland (Naters)

A new player, called the Cluster Manager (CM), is incorporated in these case studies. CMs are site managers that cluster together with the local energy , which are remotely controlled (e.g. heat pumps). They have technical knowledge of these energy systems and the connected devices (control system, meters, sensors...). They manage these assets and act as the bridge between the aggregator, who bid in the markets, and the end-user. Thus, they do not have to deal with market specifications handled by the aggregator.

4.3.1 Spanish case study: wholesale market price

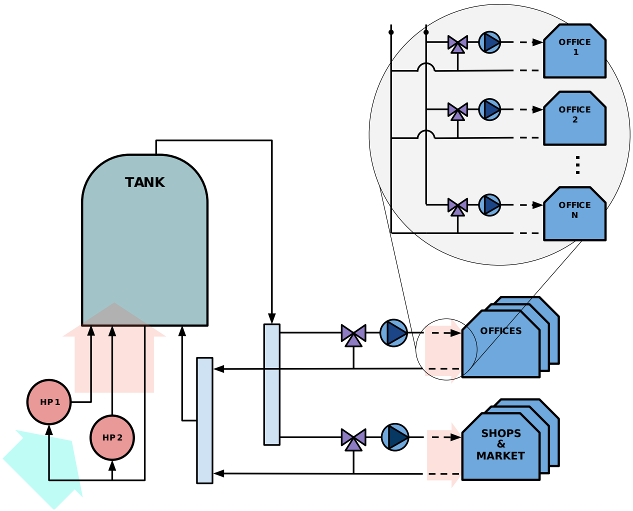


Figure 4.2: Space heating configuration of the Spanish use case. The zoom shows details of the hydraulic distribution ring

.

This case study is a pilot site constituted by buildings that combine apartments, offices, shops and a local food market. They are placed in a city called Sant Cugat, in Northern East Spain. Figure 4.2 shows the space heating and cooling system configuration. It comprises a water storage tank of 3,500 litres fed by two reversible heat pumps accounting for 60 kW of electric power. The heat pumps are controlled by an immersed temperature probe inserted into the bottom of the water tank. The heat pumps deliver thermal energy to the water storage tank through a primary circuit with two hydraulic pumps and external heat exchangers, which follow the same operation schedules as the heat pumps. The water tank provides hot and cold water to two different hydraulic circuits, which transfer this thermal energy to 32 offices, 3 shops and a local food market. These hydraulic circuits are managed by two 3-way motorized valves incorporating a proportional integral derivative (PID) control, leading to variable water volume flow rates. The control variable of the system is the water tank setpoint temperature. Since the two heat pumps do not have variable-speed compressors, they are thermostatically controlled in ON/OFF modes.

The direct load control strategy followed in this use case is based on the augmented heat pumps performance with price information from the wholesale market and weather forecast data for the current and following day. The heat pumps' electrical use adjusts times when the Spanish wholesale market spot price is lower (day-ahead optimization). To make these services operational, a Model Predictive Control (MPC) approach is put into practice. Every day at 00:00, a Genetic Algorithm (GA) optimizes the cost function, which is the minimum daily electricity consumption cost and gets the vector of the setpoint temperature of the water storage tank, T^s_{opt}, for the next 24 hours in the more cost-effective way.

The baseline model is developed based on the third approach mentioned in Section 4.2.2. It is a data-driven approach formed by two ARX models. They define the dynamic energy balance between the electricity load of the heat pumps, the water tank temperature and the thermal energy delivered to the offices, to the shops and the local food market, as well as the thermal losses in the water storage tank and the water distribution rings. More details of this kind of model can be found in [124]. These two forecasting models need, as inputs, day-ahead predictions of the thermal energy consumed by the shops, the offices, and the local food market. Since they form a block of buildings, they can be simplified as a multi-space building formed by several thermal balance nodes. This model is expected to behave highly non-linear in relation to the external temperature and other climate-dependent exogenous variables. Therefore, data-driven models are also used to evaluate their energy performance. After a previous fine-tuning phase, where several machine learning models were evaluated, the Generalised Additive Model GAM, developed by Hastie et al. [125], provided the highest accuracy and was the selected one.

4.3.2 German case study: percentage of activation time

This case study is a pilot site situated in the rural municipality of Wüstenrot in southwest Germany. It consists of a newly built positive energy settlement with 18 residential single and multifamily buildings. These buildings are connected to a low-temperature district heating grid fed by a so-called "agrothermal" – a large scale geothermal - collector. All buildings are equipped with decentralized heat pumps, thermal buffer storage tanks ranging from 175 to 300 litres, radiant floor systems, and photovoltaic (PV) systems of installed power between 6 and 29 kWp per building. In addition, a cloud-based monitoring system is installed for 12 buildings that include all relevant thermal and electrical energy flows. Within those 12 buildings, a local energy management system is installed to control the heat pumps. Figure 4.3 shows a scheme of the energy systems configuration of one of the households. Since different manufacturers provided the heat pumps, some connectivity problems appeared with the interfaces of some of them and the activation was only carried out for four heat pumps manufactured by Tecalor (Typ TTF 10 and TTC05). Two of these heat pumps have a maximum electrical power of 2.38 kW, and two have a maximum power of 3.82 kW. These activations aimed to test the potential and challenges of flexible control of heat pumps from the viewpoint of a flexible service provider.

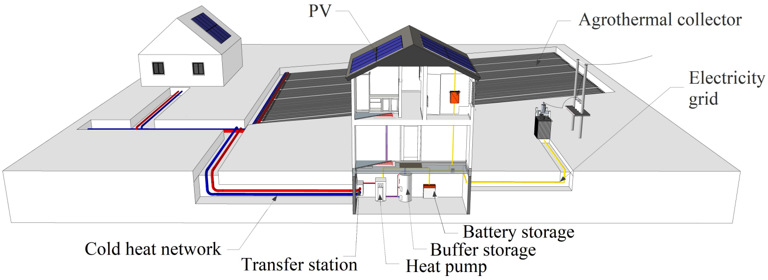


Figure 4.3: Energy systems configuration of one of the single households of Wüstenrot pilot site

The development of the baseline model followed the first approach mentioned in Section 4.2.2. For four of the selected households, a white-box model of each building was generated. More details of the models can be found in [126] and in [127]. They include heat pumps, buffer storage water tanks and control systems. To increase the model's accuracy, a calibration on parameters changeable by the users (indoor setpoint temperature and air exchange rate) with measured data was carried out. Given the unavailability of a baseline for the fifth household, due to inadequate monitoring data, this baseline has been derived from another house which was most similar (same heat pump type and no heating buffer) applying a linear extrapolation based on the historical consumption difference of both. Input parameters for the heat pump control are active power, DHW temperatures and floor heating temperatures. The control strategy is a direct load control over heat pumps on/off.

4.3.3 Swiss case study: trace to be tracked

This case study is a pilot site placed in the municipality of Naters, in Southern Switzerland. It comprises 12 residential multi-family buildings connected to a centralized low-temperature district heating network (anergy network). It represents 166 residential units. The size of the buildings ranges from 4 to 36 residential units per building. The buildings' construction years range from 1919 to 2015. Thus their envelopes have different thermal efficiencies and have either radiators or floor heating systems. Each building is equipped with one or two fixed speed compressor heat pumps, thermal buffer storage tanks for SH and DHW. Hardware components called 'gateways' are installed in each building. They collect, process and export data from the building devices (e.g. heat pumps, electricity meters) to a cloud-based platform that enables remote control of the heat pumps. The gateways installed in this use case do not have the same level of internal intelligence as the management system installed in the German use case. Due to some restrictions in the control interfaces with the heat pumps, only five out of fourteen heat pumps were intensively tested, accounting for a maximum aggregated electricity power of 34.3 kW. Figure 4.4 shows and scheme the energy systems configuration of the multi-apartment buildings.

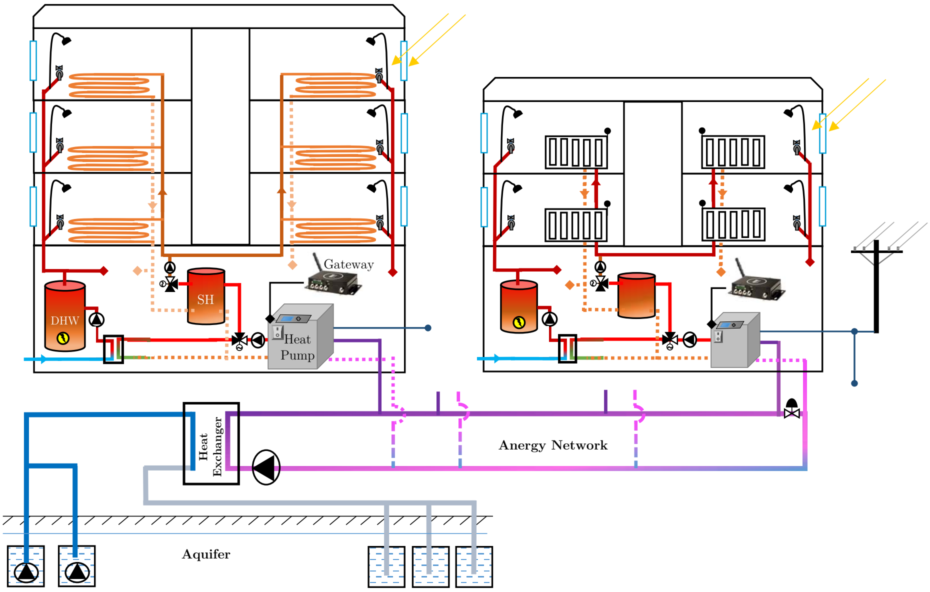


Figure 4.4: Energy systems configuration of one of the multi-apartment buildings of Nater's pilot site

The test aims to confirm the potential and challenges of flexible control of heat pumps in residential buildings from the viewpoint of a flexible service provider. A transactive DR approach was tested (a two-way communication system). Its reliability and performance over consecutive days with multiple DR-events per day was also assessed. The framework can be divided into three steps: i) the site is waiting to provide DR services by running BaU; ii) the aggregator starts negotiating power traces with the CM; iii) once a trace has been agreed on, the CM tracks it with an MPC adapted from the formulation developed by [128]. The baseline trace is modelled based on the third approach mentioned in Section 4.2.2. It is a data-driven approach formed by a Seasonal Autoregressive model (SAR) for each building using the past 3 days' power data. The aggregated baseline for the site is computed by summing up the estimated baseline of each building. The other traces are generated by solving scheduling optimization problems. The control variables of the heat pumps are the SH and DHW temperature set points, which are increased/decreased based on the new values optimized by the MPC.

4.4 Results

4.4.1 Operation of the Spanish case study

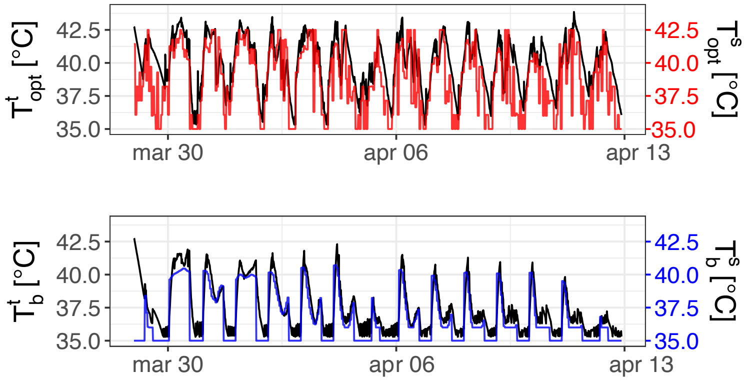


Figure 4.5: Results of the direct load control of the use case of the Spanish wholesale spot market price as activation variable

Figure 4.5 depicts the results of the direct load control applied in the case study where the Spanish wholesale spot market price acts as the activation variable.

An MPC optimization was applied during the activation period, which comprised from March 29th to April 12th 2020.

The upper figure shows, in a black-coloured line, the monitored active water storage tank temperature,T^t_{opt} along the activation period. It is compared with the water storage tank setpoint temperature,T^s_{opt}, the red-coloured line, obtained as the output of the day-ahead optimization performed every day. The T^s_{opt} is the direct control variable that drove the heat pumps performance along the activation period. As can be seen, the water tank temperature follows the optimized setpoint temperature very well. The lower plot shows the simulated baseline forecasting of the water storage tank temperature (black-coloured line), T^t_b, and the corresponding setpoint temperature (blue-coloured line), T^s_b, in the BaU scenario, which is minimum operational temperature level required by the offices, shops and local food market to keep the comfort requirements. The differences in both plots show the effect of the activation. It can be seen that the baseline forecasting usually has two temperature peaks and a second smaller temperature level. In contrast, the optimized temperature shows a single peak that is slightly lagged in time. This time lagging shows the MPC is shifting the higher setpoint temperature values to the periods with lower electricity prices.

Time series inputs for the flexibility model development

In Figure ,4.6 the day-ahead signal price, DA, the forecasting of the baseline power load, P^b, and the active power of the heat pumps, P^e, are shown. Comparing the two time series of power, the differences due to the MPC are appreciable. The bigger differences can be seen for the first days of April, where the active power is concentrated in the lower price hours while the baseline forecasting also consumes in higher prices periods.

Since the objective of this use case is to reduce the cost of the energy consumption of the heat pumps, Figure 4.7 depicts the accumulated cost difference achieved between the active optimized energy performance (black line) and the BaU scenario (red line). The reduction of cost reaches 18 % at the end of the field test operation period. This is an auspicious outcome to consider day-ahead price optimization as an important way to optimize the operational costs of heat pumps systems while offering flexibility to the electricity system.

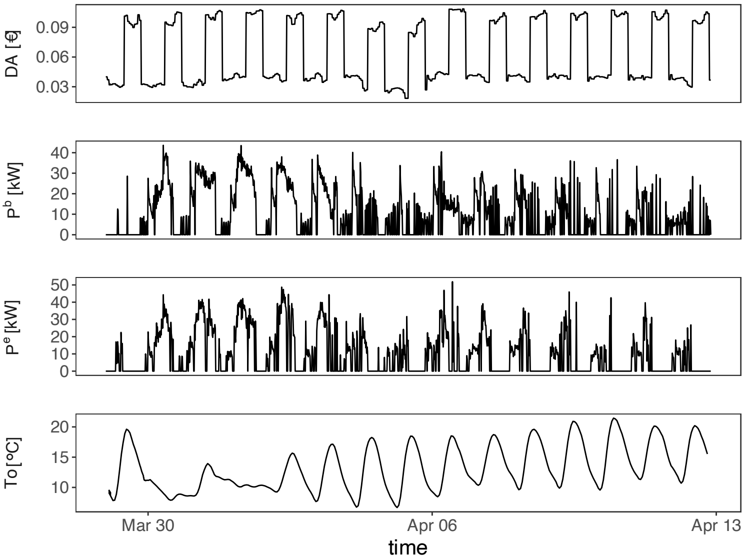


Figure 4.6: Active power, P^e, forecasting of the power baseline, P^b, day-ahead electricity price, DA and outdoor temperature, T_o of the heat pumps of the Spanish use case, during the direct load control operation period

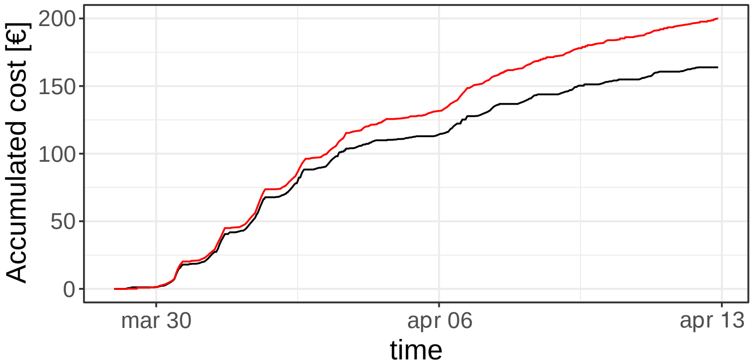


Figure 4.7: The accumulated cost of the active optimized energy performance (black line) compared to the BaU scenario (red line)

4.4.2 Operation of the German case study

Before operating the Tecalor heat pumps, different tests were conducted to verify their control capabilities. An upwards signal of 100 % activation for 30 minutes, followed by a stop of 10 minutes and the second activation of 15 minutes was sent to the heat pump controller. The result is shown in Figure 4.8. The time to start up was 56 seconds from the setpoint to on. Besides, the heat pump needed 15 minutes to reach 75 % of the maximum power. It can also be seen that the activation profile started with a first step increase, followed by a roughly linear ramp. The time to shut down was 1 min 22 seconds, whereas the shutting down profile was a decreasing step function. There is a 20-minute recovery time between switching off and switching the heat pump on again. These factors determine how a flexibility service provider can control the heat pump flexibly and integrate it into a virtual power plant.

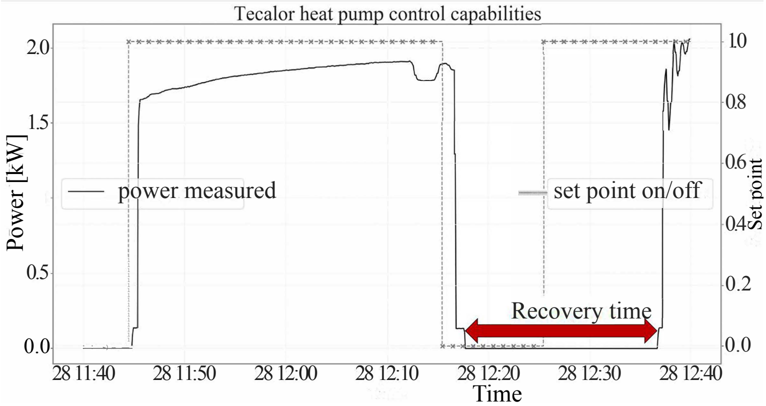


Figure 4.8: Heat pump control capabilities analysis

Another test was performed to assess a stepwise activation. For certain flexibility services, a heat pump may have to deliver a linear increasing or decreasing power curve (e.g. track the TSO's aFRR signal). Since the modulation of the power output of the heat pumps was not possible, the test performed to deliver a linear ramp was based on stacking the deactivation of heat pumps. In this test, 1 minute between each heat pump switching on/off was set up, and a variation time of switching on between 5 to 30 minutes. During this test, 3 heat pumps were available at the case study pilot site. Temperature measurements of both the DHW and the floor heating system were available, allowing us to estimate the available flexibility in the system. The ranking of the heat pumps to switch them on and off was based on the measured temperature in the floor heating circuit, which turned out to be the limiting factor.

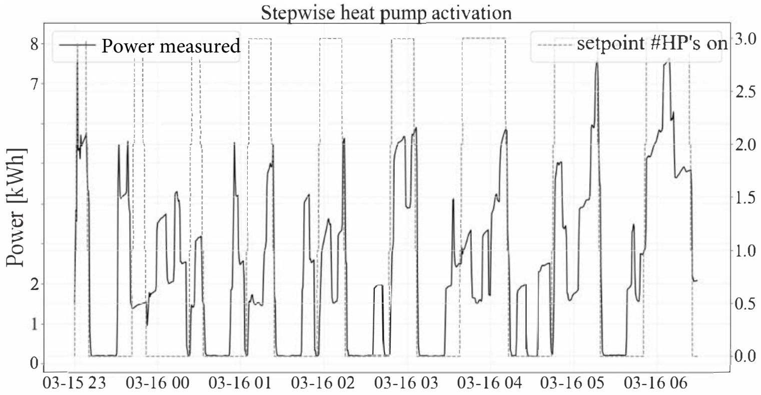


Figure 4.9: Stepwise action of 3 heat pumps in the German pilot site

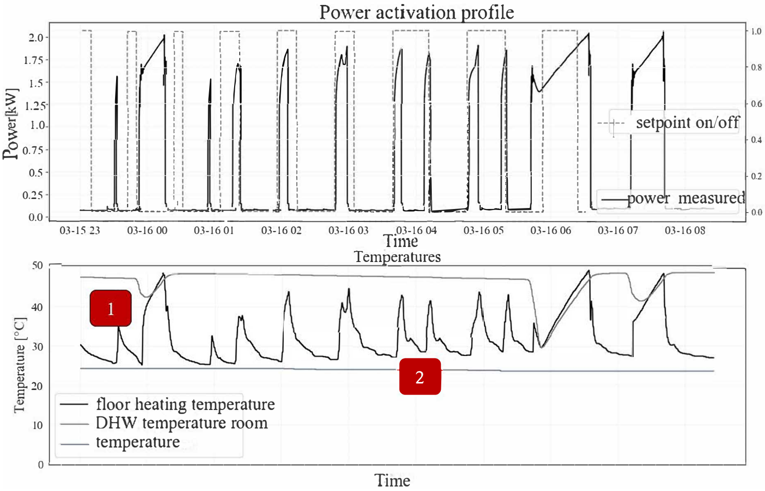


Figure 4.10: Power and temperatures of the heat pump systems along the activation period

This test is shown in Figure 4.9. The results were not satisfactory to deliver a service such as aFRR standalone. This can be explained by the small pool (3 units) and the fact that the heat pumps were often unavailable for (de)activation due to comfort/safety constraints. Furthermore, since the heat pumps controls are driven by load curves that are dependent on the indoor and outdoor temperatures, and the latest was high for the testing period, the heat pump power demand was lower than initially expected. In Figure 4.10 a deeper zoom on the (un)availability causes of one of the heat pumps is shown. Number 1 indicates forced on the situation, which means unavailability of the heat pump. This is due to the DHW temperature dropping below a threshold, forcing the heat pump to switch on for comfort reasons. Number 2 indicates a forced off situation, which means the heat pump is unavailable because the floor heating temperature exceeds the threshold temperature, forcing the heat pump to switch off for comfort/safety reasons. After the temperature drops again below the low threshold, the heat pump can be activated again, as can be seen from the graph.

Looking at the overall results of the performed tests, it has been demonstrated that the flexible operation of heat pumps in the cases study is possible and can be leveraged for multiple flexibility services or energy markets. Nevertheless, important points of attention are: i) the latency to ramp up to full power to ramp down to switch it off, which is around 1 minute; ii) and the recovery time, which is around 20 minutes. Furthermore, the comfort set points and the available storage in hot water tanks or the inertia of the building clearly determine the duration for which the heat pump can be switched on or off.

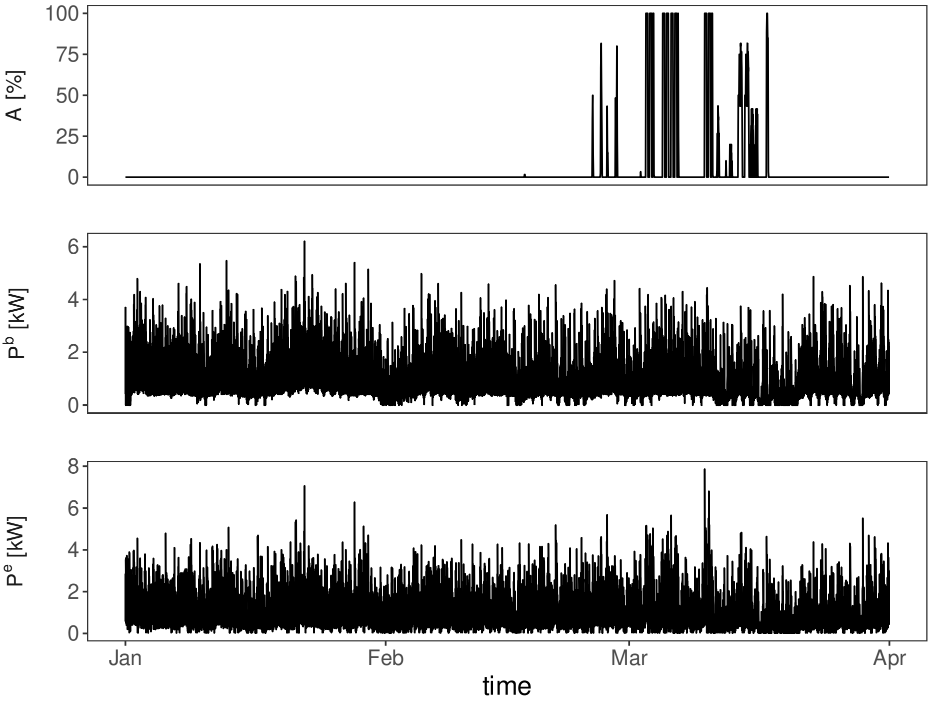


Figure 4.11: Active heat pumps power (P^e), forecasting of baseline power in BaU scenario (P^b) and percentage of activation time in each hour (A) of four households in Wüstenrot pilot site

.

Time series inputs for the flexibility model development

The operation of the case study in Wüstenrot, Germany, was a direct load control of four of the available heat pumps considering activation signals sent by a commercial aggregator. When activation was sent, the heat pumps had to operate for as long as possible during the whole activation period. In this case study, the control variable is the percentage of activation time (ON/OFF) of each building or heat pump (named 20, 22, 24 and 25). The energy flexibility is also analysed from this point of view. Figure 4.11 depicts the operation performance of the heat pumps from February 15th to March 31st. During those days, some activation signals were sent by the commercial aggregator. Therefore, as the actual heat pumps operation was affected by these signals, large differences between active power, P^e, and the forecasting of the baseline power in BaU scenario, P^b, can be appreciated for the activation period.

4.4.3 Operation of the Swiss case study

The use case in Naters, Switzerland, consists of a direct load control of five HPs that consider activation traces negotiated between the CM and the commercial aggregator. When an activation trace is accepted, the heat pumps should track the trace during the whole activation period.

Figure 4.12 represents the results of a day from a week-long test of direct load control services, detailed at the building level. The light grey vertical areas display the 15 minute negotiation periods between the aggregator and the CM. The light red vertical areas display the direct load control periods performed on-site as solutions of the tracking MPC optimization. It is not always easy to assess what a system would have done without direct load control, but coupling set points, temperature and power measurements can visually help. As a reminder, HP's local control works with hysteresis on the temperature of each storage. When the storage temperature drops too far below the setpoint value of the hysteresis, the compressor starts, and the HP runs until the upper value of the hysteresis is met. This is, of course the theory, but unforeseen events can sometimes change this behaviour.

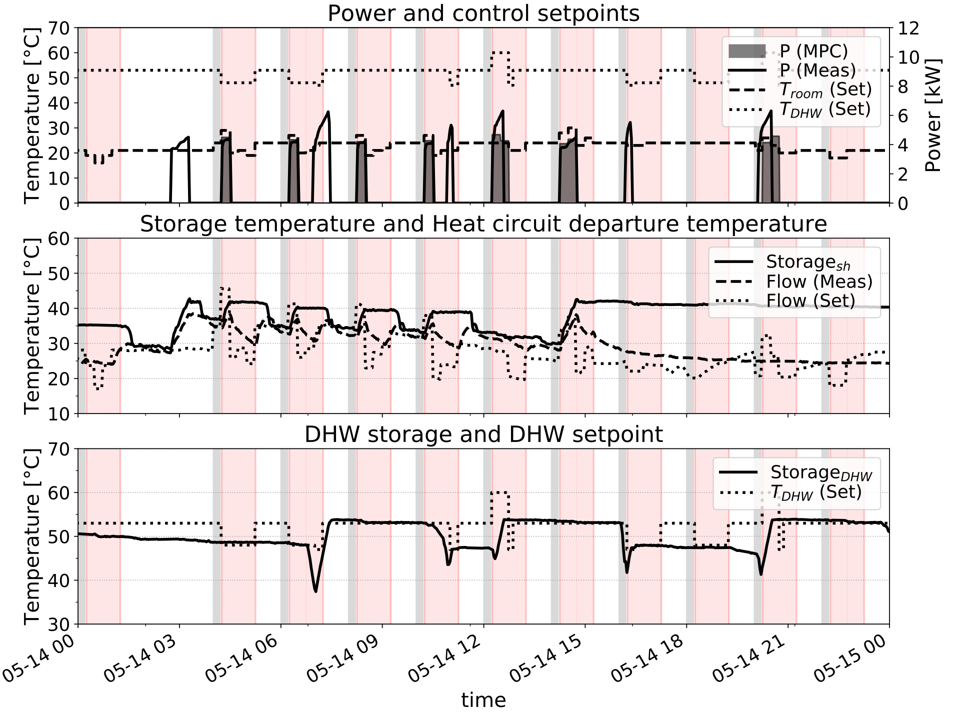


Figure 4.12: Power and temperature variation resulting from direct load control for one building in the Swiss pilot site

The top panel of Figure 4.12 shows both the temperature setpoints used for controlling the HP and the power measurements. The dotted lines correspond to the setpoint values for SH and DHW, respectively. Outside the direct load control periods, the values of those set points are set back to their default values. The solid coloured line displays the measured power consumed by the compressor of the HP. The solid coloured bars are the power consumption given as the solution of the tracking MPC. The middle panel represents the effect of direct load control on SH. The dashed line corresponds to the measured departure temperature of the heating circuit after the 3-way valve. The dotted line represents the theoretical departure temperature of the circuit as given by the heat curve of the HP. It is modelled as a function of the SH setpoint displayed in the top panel and T_o averaged over 3 hours. The bottom panel represents the effect of direct load control on DHW. In Figure 4.12, it can be seen that direct load control of SH perfectly matches the results of the tracking MPC. Instead, for the DHW load, it appears to be more difficult. Having only one sensor to assess the energy state inside the DHW storage tank makes it difficult to predict when a new cycle will occur. For comfort reasons, DHW is always prioritized and setpoints are only reduced to a minimum of 47 °C. Therefore, delaying a DHW cycle for more than 30 minutes is not always possible, as demonstrated for the DR call at 06:00. In the bottom panel, we can see that the storage temperature at the start of the period is low. This is because the setpoints are set to the lowest possible value. At 06:40, a DHW consumption brought the storage temperature below the lower bound of the hysteresis, which starts a new DHW cycle. The DR called at 10:00 is a good example of the usefulness of MPC when dealing with direct load control. When the power traces are generated, the storage tank temperature is maximal. There is only a small chance that a DHW cycle will happen in the next hour. However, within the third 15 minutes interval, a sudden high DHW consumption puts the storage temperature below the lower bound of the hysteresis, and the heat pump starts a new DHW cycle. At 11:00, to avoid deviating further from the trace, the DHW setpoint is reduced, which directly stops the heat pump.

Iterative tracking performance

Figure 4.13 presents the results of a day from a week-long test of direct load control services over all the HPs. The light grey vertical areas display the 15 minute negotiation periods between the aggregator and the CM. The light red vertical areas display the direct load control periods performed on-site as solutions of the tracking MPC optimization.

The top panel of Figure 4.13 displays the aggregated power (blue) of five participating HPs on May 14th 2020. The daily average outside temperature is 18 °C with temperatures above 20 °C from 12:00 to 20:00. Therefore, most HP consumption occurs during the early hours of the day when the outside temperature is still cold. The dashed red lines are the power loads P^f agreed upon by the aggregator and the CM. The selected traces are assumed to be constant over the sampling period of 15 minutes. Each one corresponds to a power trace resulting from a 6-hour forecast scheduling optimization problem proposed by the CM and selected by the aggregator. The bottom panel of Figure 4.13, represents the power deviation $(P^f - P^e)$. When the values are negative, it means that the on-site power was lower than the expected trace, and when they are positive, it means that the power was higher. The relative deviation over the day is -6.4 kWh and the cumulative deviation, computed as the sum of all the absolute deviations, is equal to 32.7 kWh. When high power change occurs as a result of direct load control, high deviation spikes can be observed. The negative spikes correspond to an activation delay of the HPs: Even when conditions for the local controller are met, HP compressors are only started after a 2-minute delay by the local controller. To compensate, the tracking MPCs are launched two minutes before the new actuation periods. As soon as an optimal solution is found, the new setpoints are sent. Setpoints to switch off HPs are sent at the actuation time. HP compressors directly stop when conditions are met, except when an explicit minimum running time is implemented by the local controller. The positive spikes observed can be the result of the monitoring sampling rate of 2 minutes and of the way power is measured: The power consumption of four out of five HPs is not directly measured but reconstructed from operating temperature time series and manufacturer datasheets. The interpolation and the model formulation can sometimes create mismatches.

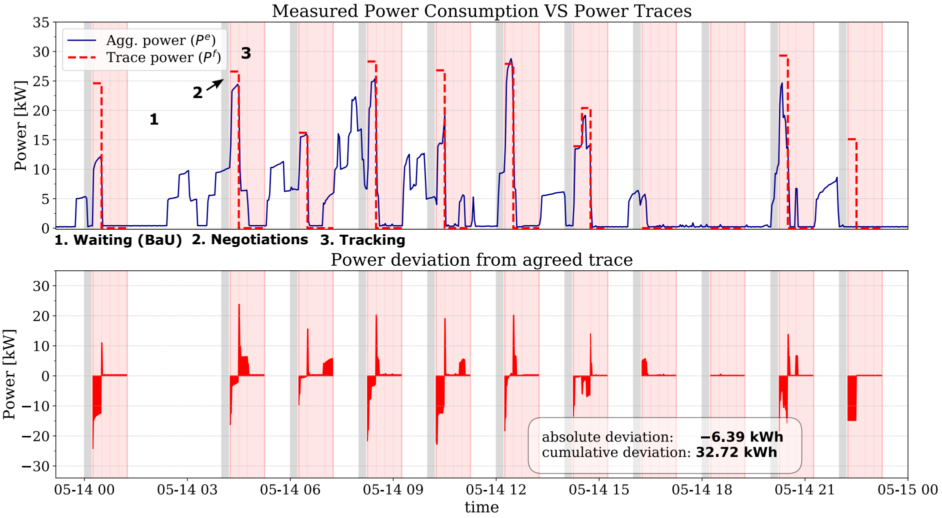


Figure 4.13: Power deviation compared to the agreed-upon traces resulting from the DR calls over a day for a weekly test in the Swiss pilot site

.

Time series inputs for the flexibility model development

In this use case, the objective of the flexibility function is to characterize how flexible the HP consumption was due to the activation trace accepted by both entities in terms of amount and shift in time. In this case study, the control variables are the DHW and SH setpoint temperature of five multi-household buildings. The entity that controls these variables is the CM, which proposes feasible traces that can be fulfilled.

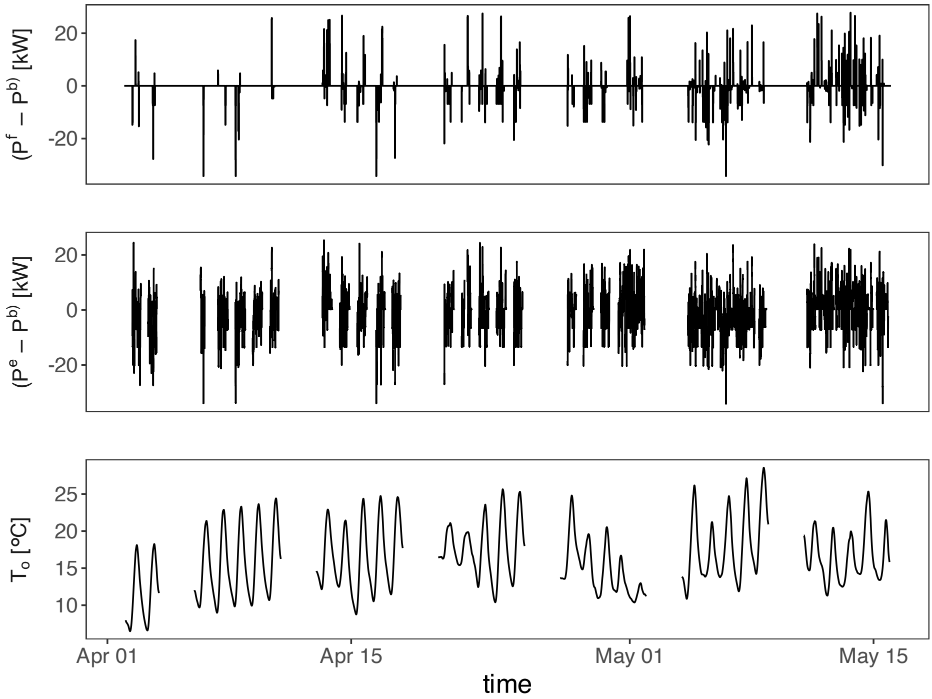


Figure 4.14: Difference between the trace to be tracked, P^f, and the prediction of the baseline, P^b, versus the difference between the active power P^e and the baseline prediction of the baseline, P^b, and the 4-hour moving-averaged outdoor temperature in the Swiss pilot site. The tests in May were week-long tests

.

Figure 4.14 depicts the performance of the HPs from April 3rd to May 15th. The granularity of the monitored data is two minutes, and the power is aggregated over the individual readings of the five available buildings. For this field operation period, multiple activation traces were tested in several operation tests. They are represented separated by gaps in Figure 4.14. In the top panel, the difference between the power trace to be tracked and the baseline forecasting is represented. As expected, significant differences between these two time series are clearly appreciated. The middle panel shows the differences between the active power, P^e, and forecasting of the baseline power,P^b. As in the other graph, the differences show that the heat pumps are following the trace up to a certain level and that these traces have very different patterns than the BaU scenario.

4.4.4 Energy flexibility evaluation and quantification

Training and validation of the flexibility models

For the case study where the activation variable is the Spanish day-ahead electricity price, a training and validation activation period was set up from March 29th to April 12th 2020. The flexibility model of this case study is defined in Equation 4.1. The training of the model to identify the regression parameters was carried out using 90 % of the data. The remaining 10 % of data was used to validate the model with new data and then avoid model overfitting. The Flexibility Function (FF) is finally inferred from this model.

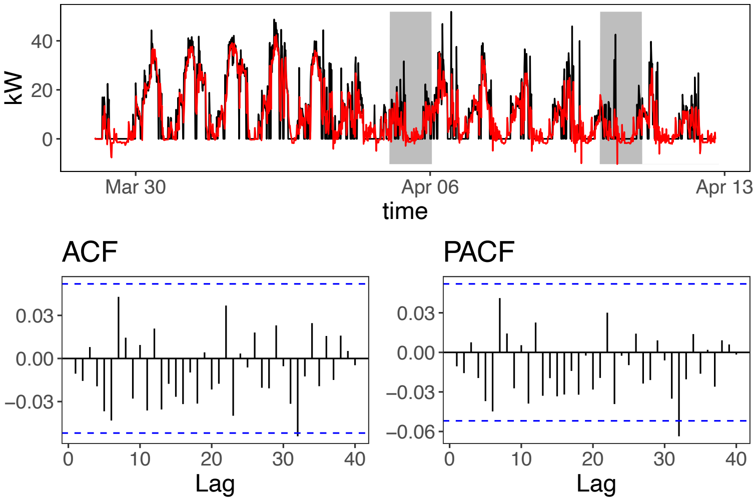


Figure 4.15: Flexibility model for the Spanish pilot site: the upper graph is a comparison of the active power (black line) and the predicted one (red line); the lower graphs show the autocorrelation functions of the training period residuals

The top plot of Figure 4.15 depicts the training and validation periods with white and grey backgrounds, respectively. In this plot, the active power, P^e, is represented by a black coloured line. The forecasting based on the flexibility model is represented by a red coloured line. It can be seen that no significant differences in residuals between the two periods are appreciated; therefore, it is confirmed that overfitting issues were avoided. Additionally, from the two bottom plots, the Auto Correlation Function, ACF, and the Partial Autocorrelation Function, PACF, of the residuals of the training period, do not indicate autocorrelation in residuals.

Therefore, they can be considered i.i.d, and the white noise condition is fulfilled. This is the requirement for a model to be considered valid.

For the case study where the activation variable is the percentage of activation time, in the German pilot case, a training and validation activation period was set up from February 15th to March 31st 2020. The flexibility model of this case study is defined in Equation 4.2. The training of the model was carried out using 90 % of the data. The remaining 10 % of the data was used to validate the model.

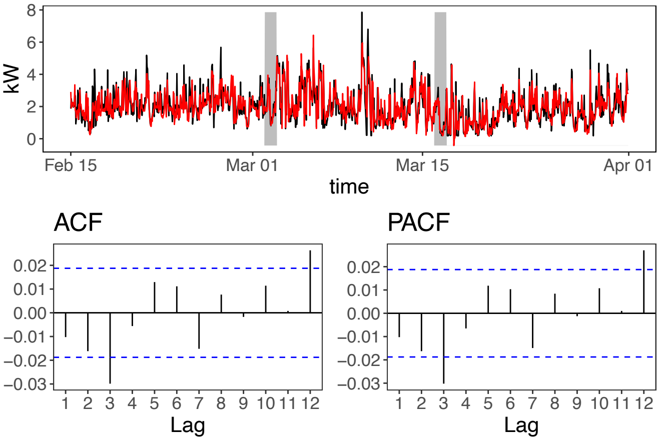


Figure 4.16: Flexibility model for the German pilot site; the upper graph shows a comparison of the active power (black line) and the predicted one (red line); the lower graphs show the autocorrelation functions of the training period residuals

.

In Figure 4.16, the upper graph depicts the training and validation periods with white and grey backgrounds, respectively. In this graph, the active power, P^e, is represented by a black coloured line. The forecasting based on the flexibility model is represented by a red coloured line. It can be seen that no significant differences in residuals between the two periods are appreciated. Although there are two significant spikes in time lags 3 and 12 in the bottom plots of the ACF and PACF, there is no clear indication of autocorrelation in residuals of the training period. Therefore, they can be considered as i.i.d. and then, the white noise condition is fulfilled for this flexibility model.

For the case study where the activation variable is the trace to be tracked, the Swiss pilot case, a training and validation activation period was set up from April

3rd to May 15th in the Swiss pilot site case study. The flexibility model of this case study is defined in Equation 4.3. The training of the model was carried out using 90 % of the data. The remaining 10 % of data was used to validate that the model. In Figure 4.17, the upper graph depicts the training and validation periods with white and grey backgrounds, respectively. In this graph, the difference between the active power and the prediction of the baseline power in BaU, $(P^e - P^b)$, is represented by a black coloured line. The forecasting based on the flexibility model is represented by a red coloured line. It can be seen that no significant differences in residuals are appreciated. Although there is one significant spike in time lag 15, in the bottom plots of the ACF and PACF, there is no clear indication of autocorrelation in residuals of the training period. Therefore, they can be considered i.i.d. The white noise condition is fulfilled for this flexibility model.

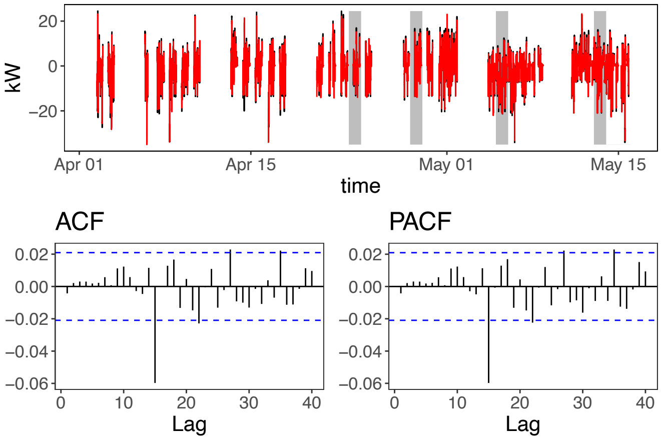


Figure 4.17: Flexibility model for the Swiss pilot site; the upper graph shows a comparison of $(P^e - P^b)$ (black line) and the predicted one performed with the flexibility model (red line); the lower graphs show the autocorrelation functions of the training period residuals

.

Flexibility functions

Figure 4.18 and Figure 4.19 show the obtained flexibility functions, FFs, for the Spanish case study, where the activation variable is the electricity day-ahead Spanish spot market. The activation variable, the day-ahead electricity price, is normalized to activation and deactivation signals of 10 cents.

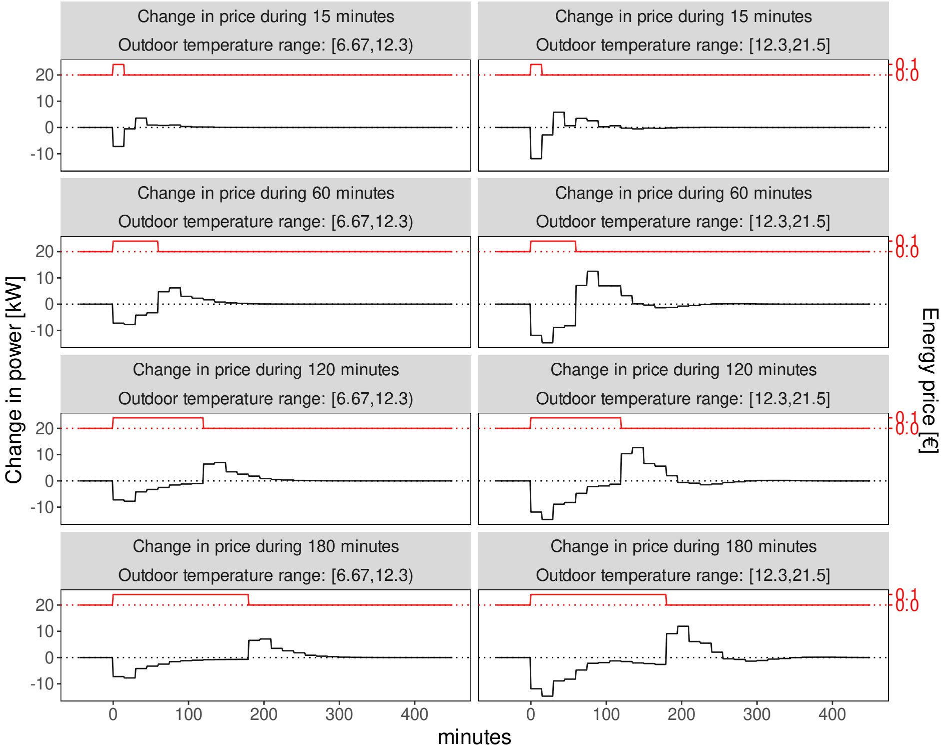


Figure 4.18: *FFs* of the Spanish case study for positive changes of the spot market price

Figure 4.18 shows the obtained *FFs* due to positive signals of different lengths and two different outdoor temperature levels. The left column shows the *FFs* for outdoor temperature ranges between 6.67 ºC and 12.3 ºC. The right column shows the *FFs* for outdoor temperature ranges between 12.3 ºC and 21.5 ºC. It can be seen that the flexibility decreases for low outdoor temperature ranges. When outdoor temperatures are between 6.67 ºC and 12.3 ºC, the average maximum deactivated power reaches -7 kW, and it remains for the first 30 minutes. Then, it increases to -3.5 kW from 30 to 45 minutes, and finally, it linearly increases to -1 kW after 100 minutes of the initial price change. Whereas, when outdoor temperatures are between 12.3 ºC and 21.5 ºC, the maximum deactivated power reaches -11 kW for the first 15 minutes; it decreases to -14 kW after 30 minutes, and finally, it increases up to -1 kW after 100 minutes of the price change. The rebound effect achieves the same maximum power levels for both temperature ranges but in positive. They start just when the activation signal finishes and reach the maximum level within the first 30 minutes after the activation signal ends. Considering the maximum available power of the two heat pumps of 60 kW, this represents maximum flexibility between 11 % and 23 %, with a rebound of the same level, for low and high outdoor temperature ranges, respectively. It can also be concluded that the estimated period where major energy shifts could be done is the starting 30 minutes after the price signal is triggered, in both outdoor temperature ranges. This conclusion is closely related to the thermal capacity of the water storage tank, which is 3,500 litres, and the permitted water tank temperature variation, which is constrained by the indoor comfort conditions in the offices, shops and the local food market.

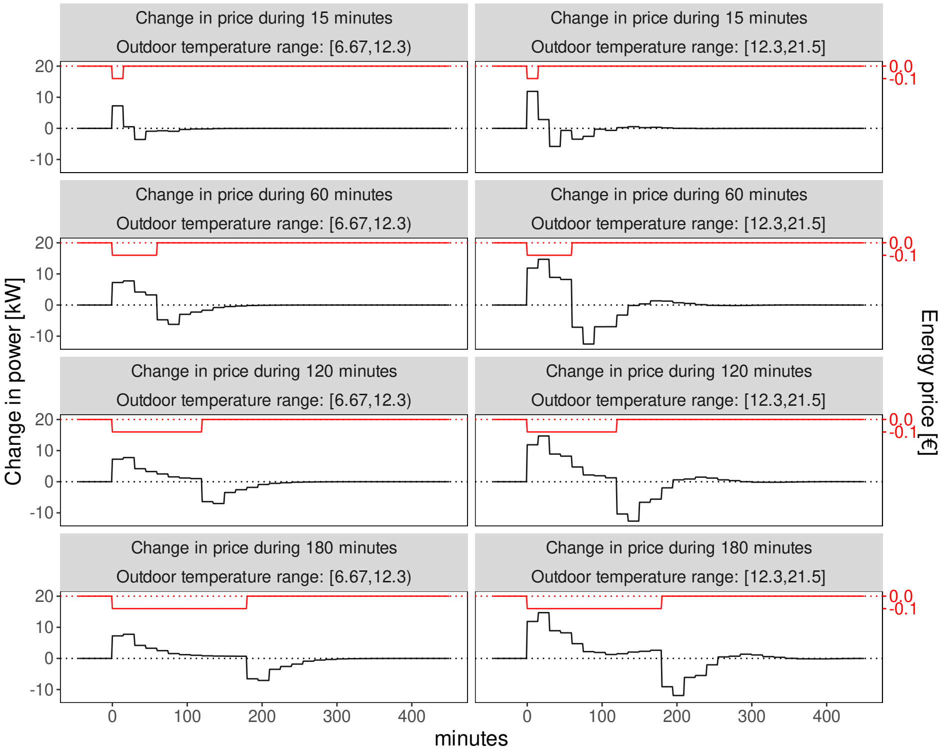


Figure 4.19: *FFs* of the Spanish case study for negative changes of the spot market price

Figure 4.19 depicts the obtained *FFs* due to negative signals of different lengths and the same outdoor temperature levels. The flexibility performance is identical to the case of positive activation but another way around. The rebound effect achieves the same maximum power levels for both temperature ranges but in negative. The same conclusions as in the case of positive signals can be deducted.

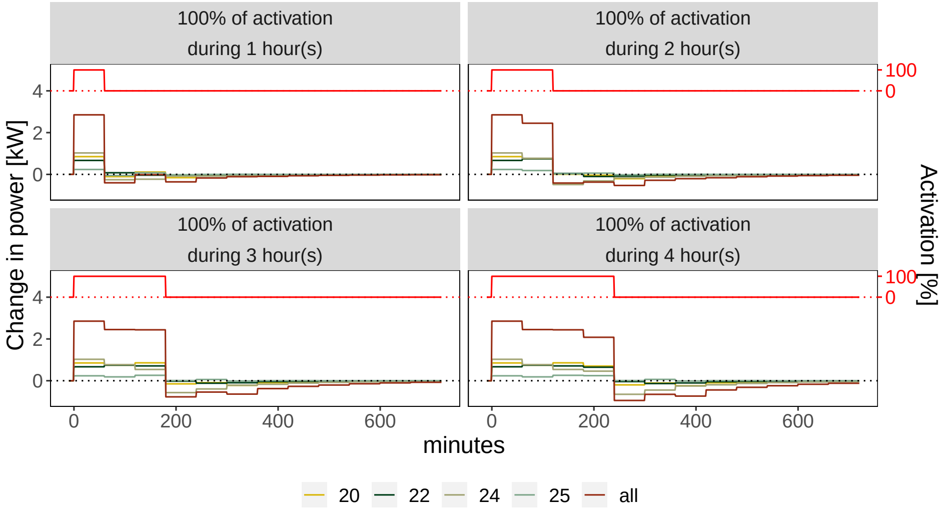


Figure 4.20: FFs of 4 heat pumps of the German case study

Figure 4.20 shows the Flexibility Functions, FFs of four heat pumps and their aggregated power, of the case study where the activation signal is the percentage of activation within an activation period. The activation variable has been normalized to 100 % activation time. The Figure 4.20 represents the FFs of each heat pump/building, named as 20, 22, 24 and 25 in the legend, and the aggregated flexibility of all of them, named as "all" in the legend. Every plot shows a FF for several activation periods ranging from 1 h to 4 h. From this Figure, multiple insights in relation to the achieved flexibility of a cluster of heat pumps can be extracted. The total amount of power flexibility for the cluster of 4 buildings reaches 2.8 kW -on average- for the first hour of activation. And from there, it decreases to 2.3 kW for the second and the third hours of activation. If the activation period is extended to four hours, maximum flexibility decreases to 2 kW. Considering a maximum available power of the four heat pumps of 10.9 kW, this represents maximum flexibility of 25 % for the first hour, 20 % for three hours and 18 % for four hours. After the activation periods, Figure 4.20 depicts a long wave rebound effect of about 20 % of the total active power. Nonetheless, around 70 % of this rebound takes place within the first 3 h after the activation period ends.

In Figure 4.20, it can also be seen that the reactions of buildings 20, 22 and 24 are quite similar and also very similar to the aggregated FFs. However, a very different behaviour happens in building 25 since it seems this heat pump is not activated. This may be due to less flexible indoor comfort conditions.

Figure 4.21 shows the FFs for the swiss case study. In this case, the activation variable is a power trace that should be tracked, and the flexibility is assessed as the deviation towards the traces and towards de predicted baseline in the BaU scenario. In Figure 4.21, the left Y-axis describes the change in power (P^e – P^b) and the right Y-axis describes the change in power due to the trace negotiated with the commercial aggregator (P^f – P^b). The flexibility is analysed for two different outdoor temperature levels; low-to-mid range [6.5 ºC, 15.7 ºC] in yellow and mid-to-high [15.7 ºC, 28.5 ºC] in black. Two types of normalized activation traces of 1 kW (e.g. red signal [-1, 0, 1]) are tested: (1) *Negative*, when the consumption is lower than the baseline, and (2) *Positive*, when the consumption is higher than the baseline. The terms *Negative* and *Positive* used here have to be differentiated from the existing positive (Upward) and negative (Downward) reserve services defined in the market regulation and provided by conventional generators. In this methodology, the term *Positive* refers to an increase in power consumption compared to the baseline, which, from a market perspective, is equivalent to a decrease in power production (negative reserve).

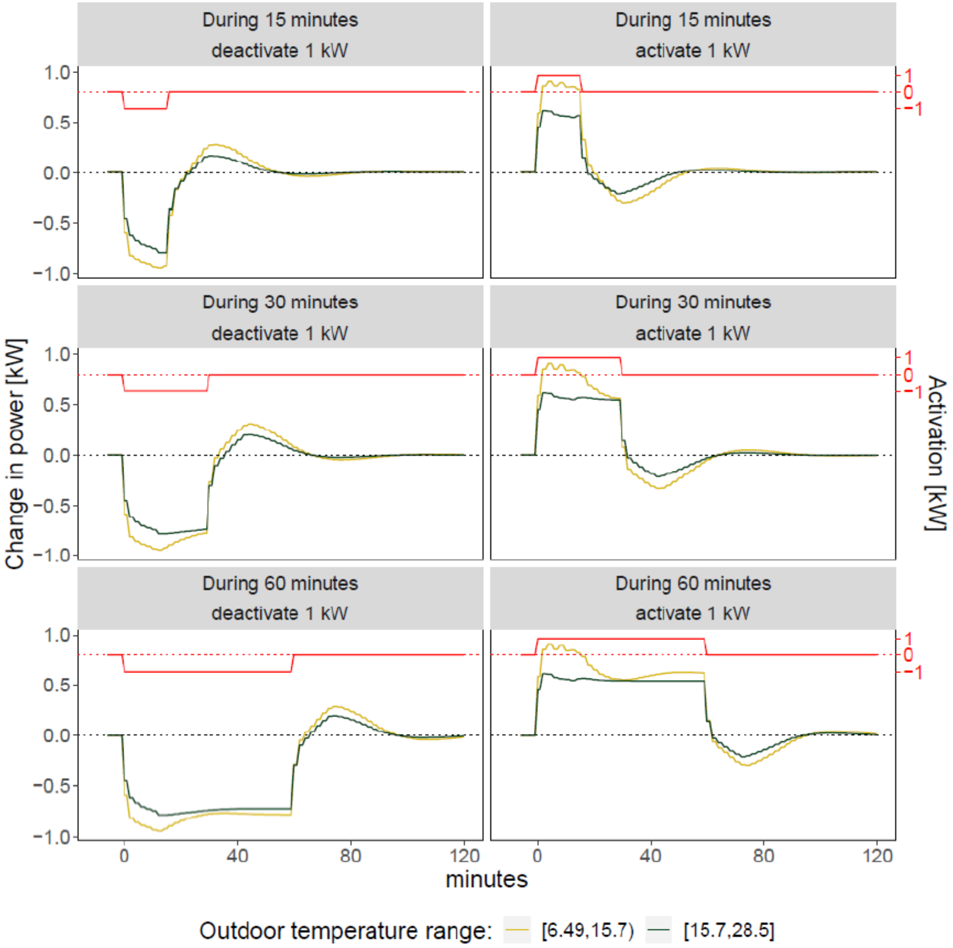


Figure 4.21: Flexibility Function (FF) of a 5-buildings cluster in Naters

In the case of tracking *negative* activation traces (left panels) in low-to-mid outdoor temperature levels (yellow lines), the active power follows 80 to 90 % of the power trace to be tracked for the first 15 minutes, reaching the maximum deactivation peak (98 %) after 13 minutes. Then, the deactivation decreases to 75 % after 30 minutes, maintaining this percentage for 30 minutes more. When tracking a *positive* activation trace, the actual power follows 80-90% of the theoretical activation for the first 15 minutes, then, it linearly decreases to 50 % after 30 minutes and maintains this percentage, with a small rebound (+10 %), up to the 60 minutes. This means that heat pumps involved in this case study, when the outdoor temperature is in the low-to-mid range, can provide the amount of flexibility required by the commercial aggregator for the first 15 minutes. Still, then, the limited availability of thermal energy storage in the building (either for SH or DHW) does not allow for full activation compliance. In both outdoor temperature levels, the rebound effect reaches up to 30 % change in power. It starts just after the activation/deactivation of the trace, and its peak is after approximately 13 minutes. In the case of mid-to-high outdoor temperatures levels, the flexibility peak of the first 15 minutes no longer exists in both *negative* and *positive* traces to be tracked. This can be explained mainly because at these temperature levels; the buildings have less thermal storage capacity and hence less energy flexibility to keep the indoor comfort within the user-defined comfort boundaries. In this case, the system which can still provide a certain level of flexibility is the DHW system, which is thermostatically controlled by the water tank temperature set points. The average compliance of tracking the trace is 60 % along the 60 minutes of activation in the case of positive and 75 % in the case of negative traces. The rebound effects follow the same path as in the lower temperatures case but with smaller peaks of around 25 % of the change in power.

4.5 Discussion

Some specific conclusions can be drawn from the operation of the DR services in each of the three pilot sites:

- The direct load control of the heat pumps of the Spanish pilot site achieved 18 % of accumulated cost savings at the end of the testing period (2 weeks). This is a promising result to demonstrate the benefits of optimising the operational costs of heat pumps through augmented performance with price information from the wholesale market forecast data.

- In the German pilot site, it was demonstrated that using the flexibility of the heat pumps allowed to optimize the heating energy cost on the day-ahead energy market. This flexibility also enabled balancing a BRP's portfolio and optimization on the balancing market. With a limited number of heat pump assets and only ON/OFF control, it was impossible to deliver linear power ramps based on the stacking of heat pumps.

- In the Swiss pilot site, a success of 91 % heat pump activation for the transactive DR approach and 50-95 % fulfilment of the activation traces was achieved for the testing period. The results are strongly correlated with the external temperature. Mid-range outdoor temperature conditions offered more flexibility, as highlighted by the higher activation success and the *FF* closer to 100 % of the theoretical activation.

The developed standard methodology for assessing the flexibility allowed to compare results from the different DR use cases and gave the necessary support for cross-comparison of the most significant energy flexibility indicators. Some specific conclusions can be deducted for the achieved flexibility in each pilot site:

4.5.1 Spanish case study

Considering the peak power of the heat pump system, the maximum flexibility achieved was between 11 % and 23 %, depending on low or high outdoor temperature ranges, respectively. A contrary rebound effect at the same level was achieved in both cases. Table 4.1 summarizes the achieved active flexibility for this pilot site:

Table 4.1: Achieved active power and rebound effect defined by the *FF* in the Spanish use case

Activation time [min]	Maximum change in power [kW]	Maximum power rebound [kW]
Flexibility with low outdoor temperature [6.6 ºC ≤ T ≤ 12.3 ºC]		
positive/negative change of price		
t ≤ 30	-7/7	6/-6
30≤ t ≤ 45	-3.5/3.5	3/-3
45≤ t ≤ 100	linear increase	linear decay
t > 100	-1/1	0.8/-0.8
Flexibility with high outdoor temperature [12.3 ºC ≤ T ≤ 21.5 ºC]		
Positive/negative change of price		
t ≤ 15	-11/11	8/-8
t ≤ 30	-14/14	12/-12
30≤ t ≤ 45	-8/8	6/-6
45≤ t ≤ 100	linear increase	linear decay
t > 100	-1/1	0.8/-0.8

4.5.2 German case study

In the German pilot site, considering a maximum aggregated power of 10.9 kW, 25 % of flexibility was achieved for the first hour. For activation of three hours, it was reduced to 20 %, and it finally decreased to 18 % for activation of four hours. A long wave rebound effect of about 20 % of the total activated power appears in all cases. However, around 70 % of the total rebound effect occurs within the first 3 h after the activation period ends. Table 4.2 summarizes achieved active flexibility for this pilot site:

Table 4.2: Achieved active power and rebound effect defined by the FF in the German use case

Activation time [min]	Maximum change in power [kW]	Maximum power rebound [kW]
Flexibility under a 100 % positive activation signal		
t ≤ 60	2.8	-0.8
60≤ t ≤ 180	2.3	-0.5
180≤ t ≤ 240	2	exponential increase
t > 240	—-	0.0

4.5.3 Swiss case study

The heat pumps involved in the Swiss case study can provide the amount of flexibility required by the commercial aggregator for the first 15 minutes. Still, then, the limited availability of thermal energy storage in the buildings does not allow for full activation compliance. In both outdoor temperature levels, the rebound effect reaches up to 30 % change in power. The average compliance of tracking the trace is 60 % along the 60 minutes of activation in the case of positive activation traces and 75 % in the case of negative ones. Table 4.3 summarises the achieved active flexibility for this pilot site:

Table 4.3: Achieved active power and rebound effect defined by the *FF* in the Swiss use case

Activation time [min]	Maximum change in power [%]	Maximum power rebound [%]
Flexibility with low outdoor temperature [6.49 ºC ≤ T ≤ 15.7 ºC]		
positive/negative trace to be followed		
t ≤ 15	85/-98	-40/40
15≤ t ≤ 30	linear decrease/-75	linear increase / linear decay
30≤ t ≤ 60	60/-75	0/0
Flexibility with high outdoor temperature [15.7 ºC ≤ T ≤ 28.5 ºC]		
Positive/negative change of price		
t ≤ 15	60/-75	-25/25
15≤ t ≤ 30	60/-75	linear increase / linear decay
30≤ t ≤ 60	60/-75	0/0

4.6 Conclusions

This study confirms that thermostatically controlled heat pumps represent a huge potential for DR flexibility depending on conditions. Furthermore, it is possible to manage clusters of heat pumps to respond to requests for DR flexibility. In addition, it has been proven that forecasting and optimization algorithms can be tailored to the particularities of each system configuration (e.g. HP interface, HP installation, and temperature sensors).

The operation tests performed in three European pilot sites demonstrated that the flexible operation of heat pumps in the field is possible and can be leveraged for multiple flexibility services or energy markets. However, several problems need to be addressed with most legacy systems. In general, those systems do not provide fully interoperable connectivity with the heat pump, resulting in constraints to the control and less flexible systems. Additionally, it has been confirmed that outdoor conditions, configured set points and the available thermal storage, both in hot water tanks or inertia in the building, determine the duration for which the heat pump can be switched on or off. Another important conclusion from this research is that a new player, called the Cluster Manager (CM), is essential to assure a successful operation of the DR services in real market scenarios.

Chapter 5

Electricity load characterization of districts

This chapter has been published as a paper:

Mor, G.; Cipriano, J.; Martirano, G.; Pignatelli, F.; Lodi, C.; Lazzari, F.; Grillone, B.; Chemisana, D. A data-driven method for unsupervised electricity consumption characterisation at the district level and beyond, Energy Reports 2021, https://doi.org/10.1016/j.egyr.2021.08.195

5.1 Introduction

Enhancing energy efficiency has become a priority for the European Union [129]. Several policies and initiatives aim to improve buildings' energy performance and collect data of sufficient quality on the effect of energy efficiency policies on building stock across Europe. Knowledge about the energy characteristics of buildings and their occupants' usage is essential to define and assess strategies for energy conservation.

For the last years, dynamic measured data has been massively accessible for a significant part of the European building stock, especially electricity consumption [130]. Besides, accurate location-based data such as weather, cadastre and socio-economic conditions became available with the explosion of governmental open data platforms and price-competitive weather online services. Given the recent advances in machine learning and big data processing, we are in an excellent position to develop and validate statistically-based methodologies capable of inferring, with no human interaction, the main energy features contained in the available data sets to determine how buildings perform and how their occupants consume energy at the local level. The outcomes of these data-driven methodologies can become essential to understand the building stock energy dynamics and, therefore, to support the transition to renewable and distributed generation at district or regional levels. A recent study [131] has shown the necessity to explore energy efficiency solutions for buildings at the local aggregated level (e.g. district, neighbourhood, city, region). The implementation of local Energy Conservation Measures (ECM) and the increase of in-situ renewable generation in buildings are key factors to satisfy energy security and limit global warming in future. This local geographical level is large enough to infer prior unknown patterns of energy consumption and to address several ECM scenarios or, at least, to support decision-making in setting up energy transition plans. Additionally, this is the geographical scale where most of the urban transformations in Europe occur and where the newest instruments for financing energy efficiency strategies in the building sector exist.

In literature, the energy characterisation based on modelling groups of buildings is named building stock modelling. Three major typologies of groups of buildings exist residential, industrial and services. Each of them corresponds to its own building archetypes, uses and occupancy patterns. Two main approaches for building stock modelling can be identified: top-down and bottom-up methods. Lagevin et al. [132] provided an extensive and updated literature review based on Swan and Ugursal [133] classification methods. They extended it by considering three major developments of the last ten years: big data, increased computing power, and new modelling techniques. The bottom-up approach begins with a detailed representation of a system's constituent part that is further aggregated to the whole-system level. In this case, building archetypes are used to characterise each building or a sample of buildings. The outcomes or the key performance indicators (KPIs) are scaled up to summarise the whole building stock of the analysed area. By contrast, top-down approaches begin with an aggregated view of the overall stock of the area, which is then disaggregated into subsequent sub-systems. In this approach, the energy performance of groups of buildings is analysed as a black box, in statistical terms, defined as a large sink with inputs and outputs following historical trends.

In both bottom-up and top-down approaches, energy characterisation of existing buildings at multiple geographical levels (district, city, region) can be used to understand trends in energy use, to correlate the energy consumption to characteristics of the territory and to identify specific locations where there are buildings with poor energy performance. Nonetheless, it is often difficult to obtain this characterisation, which can be tackled from different viewpoints, with widely varying accuracy and associated costs. Traditionally, in the case of bottom-up approaches, the characterisation of the energy performance of a given region is performed employing Building Energy Simulation (BES) models. In these cases, a calibration of the simulated data against real monthly or annual energy consumption data should be considered since the energy performance gap between simulated and real data should be minimised. Although these models are robust, this type of calibration procedures usually ignore the changes in the behaviour of the users over time, and in many cases, the dynamics between the real consumption and the climate conditions are not properly captured. Moreover, in several methodologies, a subset of representative buildings should be considered to depict the archetype of a particular region. Therefore, this model could experience large biases against reality if the sample is not statistically significant or the calibration procedure is not properly implemented. These limitations can result in high inaccuracies in the estimates of energy performance. For the last years, data-driven techniques have been applied to bottom-up approaches to overcome the limitations of simulation-based procedures. Abbasabadi and Ashayeri [134] presented a review paper where several data-driven techniques for urban energy modelling are classified. They detected that the future tendency should integrate data-driven models and simulation-based models, as each of them provides interesting advantages. In Voulis et al. [135], urban electricity demand modelling was tested for Dutch municipalities, where a combination of multiple data sets (reference electricity demand profiles, local customers composition data and aggregated local annual demand data) were used to train a regression model for local electricity demand prediction with an interesting application for local renewable energy transition plans [136]. Kontokosta and Tull [137] developed a predictive energy use model at the building, district, and city scales using training data from energy disclosure policies and predictors from the widely available property and zoning information. Their method was validated in New York, and the results demonstrated that electricity consumption could be reliably predicted using real data from a relatively small subset of buildings. In contrast, natural gas use presented a more complicated problem given the bimodal distribution of consumption and infrastructure availability. An interesting conclusion from this paper is that Ordinary Least Squares (OLS) methods perform better when applied to district and city scales, compared to other statistical techniques, such as Random Forest (RF) or Support Vector Machines (SVM). Oliveira Panão and Brito [138] developed a bottom-up approach to model the aggregated hourly electricity consumption based on a Monte Carlo model. They used probability distribution functions of the building stock characteristics, web surveys for user behaviour characterisation and energy consumption data from national statistics and smart meters data sets as input of the model. The Mean Average Percentage Error (MAPE) and the Coefficient of Variation of the Root Mean Squared Error (CVRMSE) obtained during the validation of the hourly prediction against actual data are 11% and 16%, respectively. Using data from Gothenburg, Osterbring et al. [139] proposed a methodology for building-stock energy characterisation based on characteristics of the buildings, energy performance certificates, building envelope geometries from 2.5D GIS models and measured energy.

In other cases, building stock models are used as a toolbox for specific applications. For instance, in the case of Spain, a study from Rodriguez et al. [140] showed the possibilities to mitigate energy poverty in low-income districts by combining Photo-Voltaic (PV) generation and building thermal storage using actual data and calibrated deterministic models. In this case study, the authors estimated an improvement in thermal comfort of households of up to 33% in winter and 67% in summer by using individual heat pumps and the surplus production of the district PV system. Furthermore, Gouveia et al. [141] estimated the regional energy poverty vulnerability index for Portugal at the civil parish level, based on socio-economic data, building stock characteristics, actual consumption data and theoretical consumption using the EN ISO 13790 approach.

The novelty of our research lies in the development of a data-driven technique to characterise the electricity consumption of large areas at the district level (e.g. postal code level in Spain) and upper levels, with the particularity that actual hourly consumption is considered, which makes it quite innovative considering actual state of the art. Besides, an innovative implementation of multiple statistical techniques to model the buildings stock energy consumption is performed. It is based on inferring knowledge from actual weather data, aggregated consumption data from smart meters and building stock and socio-economic characteristics data. The aim is to obtain normalised energy trends and KPIs to describe the energy consumption of each analysed region - e.g. yearly consumption per built area or monthly-averaged daily load curve due to heating or cooling needs. This characterisation requires the implementation of modelling techniques that segment the total energy consumption into different weather-dependent and non-weather-dependent components, well-described in Section 5.4.

Ideally, the main final energy fuel types related to buildings should be taken into account in the building stock characterisation. The International Energy Agency (IEA), estimate that globally in 2019, and by order of importance, the main fuel types used in buildings are: electricity (32.4%), natural gas (23.4%), traditional biomass (18.5%), oil (10.5%), renewable energy (5.9%), commercial heat (4.9%) and coal (4.1%). However, multiple issues still exist nowadays regarding the availability of energy consumption datasets at the needed aggregation levels, both in terms of geographical resolution and time-frequency. Therefore, considering the broader implementation of the Advanced Metering Infrastructure (AMI) for electricity consumption in certain EU countries, it is much more feasible to obtain detailed sets for electricity than for the rest of them. In summary, and as a first validation of the data-driven characterisation methodology presented in this chapter, electricity consumption has been considered as the only one to be characterised due to the problems in obtaining detailed data for the other main resources.

In literature, an electricity consumption segmentation at the household level using clustering techniques was developed by Kwac et al. [142]. This work helps to determine that the methodology presented in this paper need to integrate an interpreter of similar daily seasonalities, as they may not be directly related to calendar features, but to time-varying changes in the general behaviour of the consumers.In Gouveia et al., [143] energy consumption data profiles from smart meters were used to detect active behaviour regarding space heating and cooling using the deviations from normal behaviour and survey data on socio-economic conditions, building structure, equipment and use. Even though the relatively small sample of participants (19 households with survey and smart meter data), this research enlighten the necessity to consider the non-linearity between consumption and outdoor temperature, either for cooling and heating usages. In our research, multiple cooling and heating change-point temperatures along the day are considered as rectifiers of the model outdoor temperature regressors. The objective is to linearise their relationship, and thus, model properly their influence considering linear regression models. Furthermore, a first order low pass filter accounts for the thermal inertia of buildings, which helps to boost the model accuracy, especially when are based on data frequencies higher than daily (e.g. hourly).In more recent literature, several authors applied advanced energy signatures to model daily thermal consumption to characterise the linear and non-linear heat usage dependency on outdoor temperature, wind and solar irradiation [144]. Similar techniques are applied in our research, focusing on the characterisation of building stock instead of individual households. Furthermore, in Wang et al. [145], regression and machine learning techniques were also used to detect how electricity use was influenced by weather and COVID-19 lockdowns over three large metropolitan areas city-scale aggregated forecasting (Los Angeles, Sacramento and New York). The daily models' forecasting accuracy was between 4-6% of CVRMSE. In our research, similar accuracy is reached 4-12% of CVRMSE, highly depending on the number of consumers aggregated on each case. Even though, and considering the 4h-frequency aggregation considered in our analysis, the increase in error compared to the daily aggregation is very low. The results are also more accurate than the 16% CVRMSE obtained in Oliveira Panao et. al research [138].

Besides the definition and implementation of the methodology, a validation case study is presented in Section 5.5. The outcomes are shared through a Shiny web dashboard [146] that depicts multiple plots related to the electricity consumption characterisation for each postal code and interactive maps to benchmark the whole set of KPIs, among other visualisations. The Spanish province of Lleida (>12500 km^2) is the area selected for the case study. Section 5.2 extensively describe the main data sources used for the case study validation. The final goal is to provide a geographically aggregated characterisation methodology for building performance and usage trends of electricity consumption, both for the residential and public/tertiary buildings.

5.2 Input data

This section explains the data requirements, gathering, cleaning, and transformation procedures needed to successfully characterise the electricity consumption over the case study in Spain. Moreover, it defines the initial requirements to implement this methodology in other countries or use cases.

5.2.1 Cadastral data

Buildings characteristics are gathered from national cadastral datasets. The data format used by these entities across EU countries is harmonised using the INSPIRE Buildings theme [147]. In the case of Spain, the massive downloadable public information of cadastral datasets is available through ATOM files [148], where Geography Markup Language (GML) files regarding "buildings" and "building parts" can be obtained for all the Spanish municipalities. Those files contain a set of georeferenced information for each building and, depending on the type of information described. Each variable could be grouped in:

1. Geometry information, including information about 2D geometries of the building parts, gross floor area, number of floors above and below ground.
2. Typology information, including variables, such as the major current use, the total number of dwellings and building units.
3. Construction information, including the actual conditions of the building and the year of construction.

Even if the amount of information is pervasive, it has to be considered that multiple drawbacks exist when using cadastral data gathered through ATOM files. In the case of the variables belonging to groups 2 and 3, it should be considered that many data inaccuracies can exist compared to the real conditions. Some of the encountered issues are:

- Problems dealing with buildings with several main uses (services + residential, or industrial + services), as only one use is related to each building.
- Non-realistic dwelling areas based on the gross floor area, due to the influence of large parking and/or community areas.
- Some building information is not available for all the regions (Buildings located in the countryside vs those located in cities). For instance, in certain rural areas of the Lleida province, up to 30% of buildings without current use information.

To avoid unrealistic estimations when aggregating this data to postal code geographical level, some filters were considered - e.g. subtract ground floors and basements from the total gross area in residential buildings with more than three floors.

5.2.2 Socioeconomic data

The economic status and the demographics indicators considered in this methodology are gathered through national statistics institutes. In the case of Spain, this data can be obtained from an experimental project of the Spanish Statistical Office (INE), named "Household income distribution map" [149]. This project proposes constructing statistical indicators of the level and distribution of household income at the municipal and census tract geographical levels from the link between INE's

Access toll name	Time-of-use structures (n$^{\underline{\text{o}}}$ periods)	Contracted power range	Main usage
2.0	A (1) DHA (2) DHS (3)	< 10 kW < 10 kW < 10 kW	All-kind of dwellings, houses, small-sized shops or offices
2.1	A (1) DHA (2) DHS (3)	≥ 10 and < 15 kW ≥ 10 and < 15 kW ≥ 10 and < 15 kW	Big-sized houses medium-sized shops or offices
3.0	A (3)	≥ 15 kW	Public buildings, or big-sized shops, or office buildings
3.1	A (3)	< 450 kW (high voltage)	Industrial buildings

Table 5.1: Electricity tariffs description in the Spanish market

demographics information and the tax data from the National and the Autonomous Treasuries. Some of the indicators obtained at the census tract geographical level are the average income per person and household, the income primary sources, the income quantile 80 and 20 ratio, the number of inhabitants, the average population age, the percentage of people under 18 and over 65, the number of people per household, the percentage of single households, and the Gini index.

5.2.3 Electricity consumption data

Datadis platform [150] supplies the historical hourly electricity consumption aggregated by postal code, economic sector, tariff and DSO for Spain. This platform is participated by most Spanish DSOs, who provide electricity services to around 28 million consumption points. The aggregated hourly consumption is gathered through the Datadis API, which requires authentication using an FNMT electronic certificate [151] of a legal entity. On average, most of the postal codes contain two years of historical data. The aggregated information for each obtained item through the API is the hourly consumption and online contracts.

In Spain, the electricity tariffs available through Datadis during the period represented in the case study (from beginning 2018 to mid-2020) are specified in Table 5.1.

Data within the same economic sector sometimes contains gaps, multiple energy trends, and seasonality between different tariffs. Due to this fact, a synthetic tariff is created, named "all", weighting its values using the number of contracts per each of the tariffs. This aggregated tariff improves the representativeness of each postal code when the results are visualised over a map.

Even considering the use of aggregated consumption data at a postal code level, which alleviates the influence of poorly measured data at some particular site, some problems were detected during the initial quality checks. Hence, it became mandatory the implementation of a data cleaning process before modelling steps. In essence, the outlier filtering avoids any measure which accomplishes, at least, one of the following conditions:

1. Hourly consumption equal to 0. It is certainly impossible to have zero consumption considering that several contracts are aggregated per each postal code.

2. Hourly consumption lower than the maximum feasible contracted power, depending on the tariff restrictions. For instance, the contracted power must be lower than 10 and 15 kW, respectively, for 2.0 and 2.1 tariffs.

3. Hourly consumption is six times higher than the 3rd quartile of all the historical consumptions.

4. Hourly consumption outside the right-aligned moving average plus-minus three moving standard deviations, considering a window of 15 days.

5.2.4 Weather data

Outdoor weather conditions are obtained through the Dark Sky API service [91] for the whole area in analysis. In essence, the historical weather data for the same period is downloaded for each of the postal codes considered. The most important variables in our analysis are the outdoor temperature and wind speed.

5.2.5 Geographical levels

Data used in the framework of this energy characterisation is related to multiple geographical levels. In this subsection, each of the available geographical levels is described. Moreover, in the data integration section, it is described how all data sets are normalised to the same level, which is a necessary step to analyse the datasets.

Building level

Data referenced to this level contains the exact location where the building is physically placed. Cadastral data is an example of a dataset with this geographical level. Beyond cadastral information, and mainly due to privacy issues, there are not many other open datasets available at this level. It is worth mentioning that this geographical level would be the most interesting due to its flexibility for aggregation purposes. For instance, characterisation results could be easily aggregated by streets, blocks of buildings, neighbourhoods or custom aggregations which could provide differences within the census tract or postal code levels.

Postal code level

The postal code is a code that is assigned to different areas or places in a country. Initially, it was a code to facilitate and mechanise the delivery of mail. It usually consists of a series of digits, although in some countries, it includes letters. In the case of Spain, it is composed of the province code (two first digits) and then three more digits which represent each different postal code. The institution that defines them is the "Sociedad Estatal Correos y Telégrafos, S.A.". Many other companies, or even the government, widely use this geographical level to refer their data to its location. It strikes a good balance between anonymity, simplicity and detail. The shape of each postal code is obtained from KML files [152].

Census tract level

Census tracts are the lowest level units for disseminating statistical information and are also used to organise electoral processes. Being basically operational in nature, they are always defined by more or less fixed sizes: the number of statistical surveys that an interviewer agent can distribute and collect for population counting purposes in the time of one or two months, or the number of people who can vote in a ballot box without crowding on an election day.

The most updated shapefiles of the census tract in Spain are obtained from the National Statistical Office [153].

For urban areas, the census tract level offers much more detail than the postal code one. The number of building blocks inside a certain census tract is much lower than in the postal code. However, for rural areas, the representativity of both levels is very similar, as they usually represent areas of similar size.

5.3 The architecture of the solution

The implementation of this methodology consists of combining and analysing multiple layers of data, as shown in Figure 5.1. Considering that this information has heterogeneous characteristics, both in terms of frequency, geographical reference and typology, one of the mandatory aspects regarding the cross-analysis is the harmonisation of these layers. Specific aggregations and transformations are done for each input dataset. For instance, GML files of cadastre data are transformed to tabular data and aggregated to several geographical levels to correlate cadastral information to socioeconomic conditions, electricity consumption and weather data. Python 3.8 [154] is used to extract, transform, and load data processes, using QGIS 3.10 [155] as a backend to analyse geospatial data. Regarding the electricity characterisation model, it is implemented in R 4.1 [156]. All these scripts store the raw, intermediate and final results to a MongoDB 4 non-relational database [157].

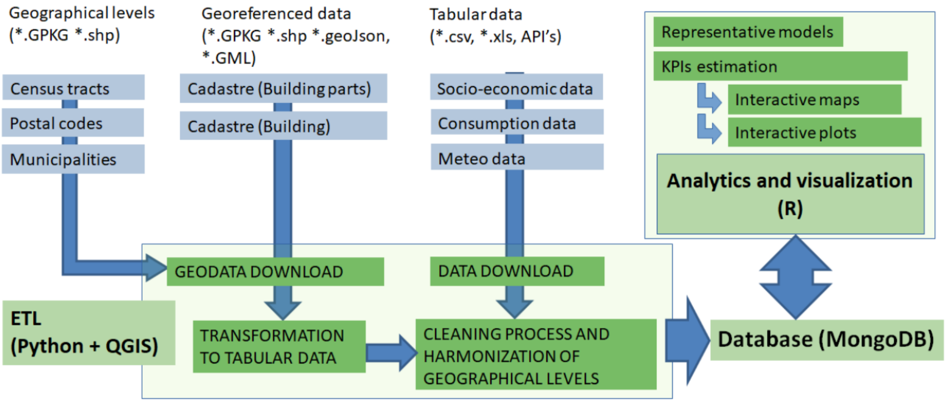


Figure 5.1: General view of the data flow and the architecture of the software

The relationships and transformations among the different databases are depicted in the UML model shown in Figure 5.2, where the classes are named by the name of the provider and the name of the collection, separated using ":". In the case of intermediate or final classes used by the data analytics backend or by the frontend to visualise results, the provider's name is "beegeo". The calculations considered for the aggregations to higher geographical levels are explained following SQL format in yellow notes.

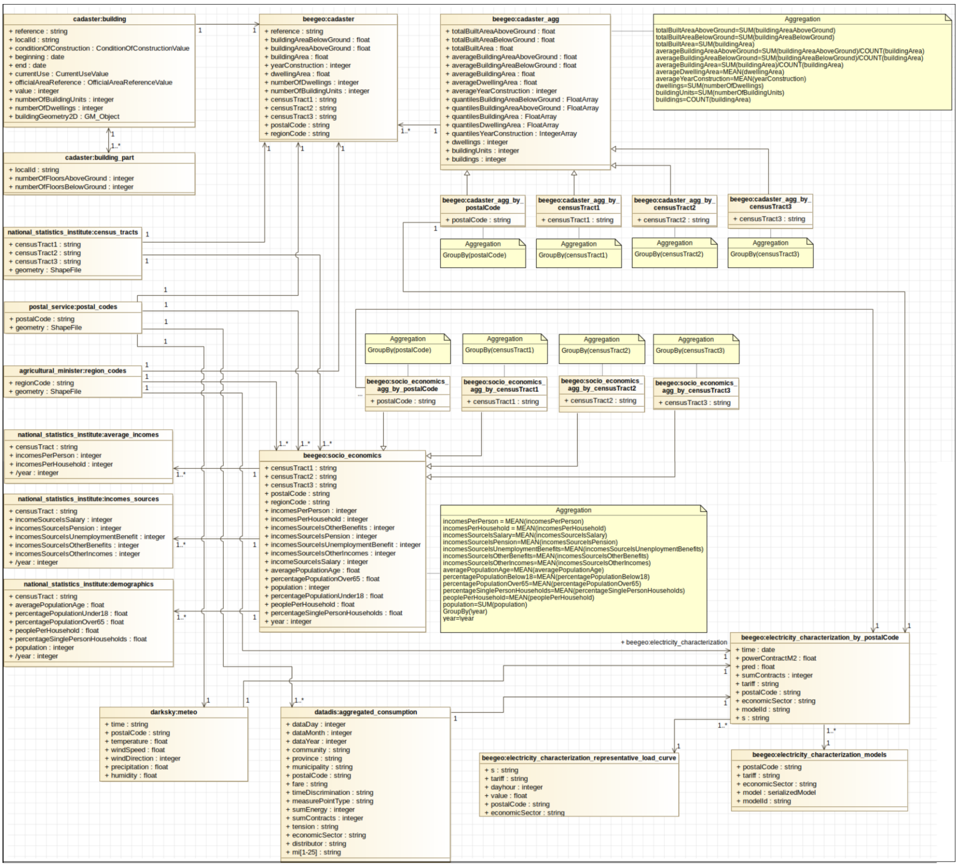

cadaster:building
+ reference : string
+ localId : string
+ conditionOfConstruction : ConditionOfConstructionValue
+ beginning : date
+ end : date
+ currentUse : CurrentUseValue
+ officialAreaReference : OfficialAreaReferenceValue
+ value : integer
+ numberOfBuildingUnits : integer
+ numberOfDwellings : integer
+ buildingGeometry2D : GM_Object
beegeo:cadaster
+ reference : string
+ buildingAreaBelowGround : float
+ buildingAreaAboveGround : float
+ buildingArea : float
+ yearConstruction : integer
+ dwellingArea : float
+ numberOfDwellings : integer
+ numberOfBuildingUnits : integer
+ censusTract1 : string
+ censusTract2 : string
+ censusTract3 : string
+ postalCode : string
+ regionCode : string
beegeo:cadaster_agg
+ totalBuiltAreaAboveGround : float
+ totalBuiltAreaBelowGround : float
+ totalBuiltArea : float
+ averageBuildingAreaAboveGround : float
+ averageBuildingAreaBelowGround : float
+ averageBuildingArea : float
+ averageDwellingArea : float
+ averageYearConstruction : integer
+ quantilesBuildingAreaBelowGround : FloatArray
+ quantilesBuildingAreaAboveGround : FloatArray
+ quantilesBuildingArea : FloatArray
+ quantilesDwellingArea : FloatArray
+ quantilesYearConstruction : IntegerArray
+ dwellings : integer
+ buildingUnits : integer
+ buildings : integer
Aggregation
totalBuiltAreaAboveGround=SUM(buildingAreaAboveGround)
totalBuiltAreaBelowGround=SUM(buildingAreaBelowGround)
totalBuiltArea=SUM(buildingArea)
averageBuildingAreaAboveGround=SUM(buildingAreaAboveGround)/COUNT(buildingArea)
averageBuildingAreaBelowGround=SUM(buildingAreaBelowGround)/COUNT(buildingArea)
averageBuildingArea=SUM(buildingArea)/COUNT(buildingArea)
averageDwellingArea=MEAN(dwellingArea)
averageYearConstruction=MEAN(yearConstruction)
dwellings=SUM(numberOfDwellings)
buildingUnits=SUM(numberOfBuildingUnits)
buildings=COUNT(buildingArea)
cadaster:building_part
+ localId : string
+ numberOfFloorsAboveGround : integer
+ numberOfFloorsBelowGround : integer
beegeo:cadaster_agg_by_postalCode
+ postalCode : string
beegeo:cadaster_agg_by_censusTract1
+ censusTract1 : string
beegeo:cadaster_agg_by_censusTract2
+ censusTract2 : string
beegeo:cadaster_agg_by_censusTract3
+ censusTract3 : string
Aggregation
GroupBy(postalCode)
Aggregation
GroupBy(censusTract1)
Aggregation
GroupBy(censusTract2)
Aggregation
GroupBy(censusTract3)
national_statistics_institute:census_tracts
+ censusTract1 : string
+ censusTract2 : string
+ censusTract3 : string
+ geometry : ShapeFile
postal_service:postal_codes
+ postalCode : string
+ geometry : ShapeFile
agricultural_minister:region_codes
+ regionCode : string
+ geometry : ShapeFile
Aggregation
GroupBy(postalCode)
Aggregation
GroupBy(censusTract1)
Aggregation
GroupBy(censusTract2)
Aggregation
GroupBy(censusTract3)
beegeo:socio_economics_agg_by_postalCode
+ postalCode : string
beegeo:socio_economics_agg_by_censusTract1
+ censusTract1 : string
beegeo:socio_economics_agg_by_censusTract2
+ censusTract2 : string
beegeo:socio_economics_agg_by_censusTract3
+ censusTract3 : string
national_statistics_institute:average_incomes
+ censusTract : string
+ incomesPerPerson : integer
+ incomesPerHousehold : integer
+ /year : integer
national_statistics_institute:incomes_sources
+ censusTract : string
+ incomeSourceIsSalary : integer
+ incomesSourceIsPension : integer
+ incomesSourceIsUnemploymentBenefit : integer
+ incomesSourceIsOtherBenefits : integer
+ incomesSourceIsOtherIncomes : integer
+ /year : integer
national_statistics_institute:demographics
+ censusTract : string
+ averagePopulationAge : float
+ percentagePopulationUnder18 : float
+ percentagePopulationOver65 : float
+ peoplePerHousehold : float
+ percentageSinglePersonHouseholds : float
+ population : integer
+ /year : integer
beegeo:socio_economics
+ censusTract1 : string
+ censusTract2 : string
+ censusTract3 : string
+ postalCode : string
+ regionCode : string
+ incomesPerPerson : integer
+ incomesPerHousehold : integer
+ incomesSourceIsOtherBenefits : integer
+ incomesSourceIsPension : integer
+ incomesSourceIsUnemploymentBenefit : integer
+ incomesSourceIsOtherIncomes : integer
+ incomeSourceIsSalary : integer
+ averagePopulationAge : float
+ percentagePopulationOver65 : float
+ population : integer
+ percentagePopulationUnder18 : float
+ peoplePerHousehold : float
+ percentageSinglePersonHouseholds : float
+ year : integer
Aggregation
incomesPerPerson = MEAN(incomesPerPerson)
incomesPerHousehold = MEAN(incomesPerHousehold)
incomesSourceIsSalary=MEAN(incomesSourceIsSalary)
incomesSourceIsPension=MEAN(incomesSourceIsPension)
incomesSourceIsUnemploymentBenefits=MEAN(incomesSourceIsUnenploymentBenefits)
incomesSourceIsOtherBenefits=MEAN(incomesSourceIsOtherBenefits)
incomesSourceIsOtherIncomes=MEAN(incomesSourceIsOtherIncomes)
averagePopulationAge=MEAN(averagePopulationAge)
percentagePopulationBelow18=MEAN(percentagePopulationBelow18)
percentagePopulationOver65=MEAN(percentagePopulationOver65)
percentageSinglePersonHouseholds=MEAN(percentageSinglePersonHouseholds)
peoplePerHousehold=MEAN(peoplePerHousehold)
population=SUM(population)
GroupBy(\year)
year=\year
+ beegeo:electricity_characterization
beegeo:electricity_characterization_by_postalCode
+ time : date
+ powerContractM2 : float
+ pred : float
+ sumContracts : integer
+ tariff : string
+ postalCode : string
+ economicSector : string
+ modelId : string
+ s : string
darksky:meteo
+ time : string
+ postalCode : string
+ temperature : float
+ windSpeed : float
+ windDirection : integer
+ precipitation : float
+ humidity : float
datadis:aggregated_consumption
+ dataDay : integer
+ dataMonth : integer
+ dataYear : integer
+ community : string
+ province : string
+ municipality : string
+ postalCode : string
+ fare : string
+ timeDiscrimination : string
+ measurePointType : string
+ sumEnergy : integer
+ sumContracts : integer
+ tension : string
+ economicSector : string
+ distributor : string
+ mi[1-25] : string
beegeo:electricity_characterization_representative_load_curve
+ s : string
+ tariff : string
+ dayhour : integer
+ value : float
+ postalCode : string
+ economicSector : string
beegeo:electricity_characterization_models
+ postalCode : string
+ tariff : string
+ economicSector : string
+ model : serializedModel
+ modelId : string

Figure 5.2: UML of the used data model

The implementation of this UML representation is made using a combination of open-source analytics and data storage technologies that allow validating the methodology over the province of Lleida. The visualisation is made using a Shiny frontend application [146], which has been developed on purpose for this case study. In general, the data prompted into this web application is always read from the MongoDB database. However, some of the normalisation calculations are computed on-demand using the serialised characterisation models estimated in the analytics backend. The web application is mounted on Docker containers, hence it should be prepared to be horizontally scalable, which is an interesting feature for future deployment of the application, either for Spain or other EU countries. The time period extends from the beginning of 2018 until June 2020, but the ETL processes are prepared to recursively obtain new data as soon as it becomes available online. To sum up, the web application is divided into four tabs: KPIs on a map, Characterisation, Benchmarking and KPIs correlation.

5.4 Electricity characterisation method

The characterisation methodology consists in the execution of the following steps per each region, tariff, and economic sector under analysis:

- Clustering the daily load curves to infer the most representative usage patterns.
- Estimate a regression model of the electricity consumption using calendar features, clustering results and weather conditions as exogenous variables.
- Disaggregate the raw electricity consumption in baseload, holidays, heating and cooling components.
- Calculate the performance KPIs.

5.4.1 Clustering model

A clustering of the daily load curves for each postal code combination, tariff and economic sector is performed to detect similar usage patterns. The representative groups obtained should be used along the algorithm to increase the reliability of the characterisation due to the consideration of the multiple seasonality's that could not be related to calendar variables or weather conditions.

Clustering can be achieved using various algorithms, which differ in their way to define the constituents of a cluster and how to find them efficiently. The best-suited clustering algorithm depends on the particular data set and the intended use of the results. In this study, the achieved outcome of the clustering technique is to obtain a model to define the typical usage patterns for each case based on the original consumption time series.

The first step is to encode the input data appropriately to the usage pattern recognition. To do so, the original hourly frequency is resampled to 4 hours, as the objective is to cluster daily load curves based on their approximate peak and valley consumptions - e.g. morning consumers, double-valley consumers, or nightly consumers. Then, two normalisations and one encoding procedure are considered:

1. Conversion of the original consumption time series (Q^{abs}) to a daily relative consumption time series (Q^{rel}). $Q^{rel}_t = \frac{Q^{abs}_t}{\sum_{t \in day} Q^{abs}_t}$.

2. Generation of a matrix of days (day) as rows, and parts of the day (dh) as columns, using the daily relative consumption time series.

3. Transformation of the values using a Z-score normalisation, which improves the performance of the clustering algorithm.

   $Q^{z,rel}_{day,dh} = \frac{Q^{rel}_{day,dh} - mean(Q^{rel}_{dh})}{sd(Q^{rel}_{dh})}$

Among the different clustering techniques, distribution-based clustering is chosen because it is the one that most closely resembles the way energy measurement data sets are generated by sampling random objects from a distribution. The distribution of every observation is specified by a probability density function through a finite mixture model of G components, as shown in Equation 5.1.

$$f(x_i;\Psi)) = \sum_{k=1}^{G} \pi_k N(\mu_k, \Sigma_k) \tag{5.1}$$

Where $\Psi = \left\{\pi_1, ..., \pi_{G-1}, \mu_1, ..., \mu_G, \Sigma_1, ..., \Sigma_G\right\}$ are the parameters of the mixture model. $N_k(x_i;\mu_k,\Sigma_k)$ is the k_{th} component Gaussian density for observation x_i with parameter vector (μ_k, Σ_k). $(\pi_1, ..., \pi_{G-1})$ are the mixing weights or probabilities (such that $\pi_k > 0$, $\sum \pi_k = 1$. And G is the number of mixture components (in the model-based approach to clustering, each component is associated with a group or cluster). Assuming that G is fixed, the mixture model parameters Ψ are usually unknown and should be estimated. In the case described above, it is assumed that all component densities arise from the same parametric distribution family: the Gaussian. Thus, clusters are ellipsoidal, centred at the mean vector μ_k and with geometric features such as volume, shape and orientation, determined by the covariance matrix Σ_k. The mixture of multi-dimensional Gaussian probability distributions that best fit the input dataset is estimated via the expectation-maximisation algorithm for maximum likelihood estimation. The covariance (Σ_k) structures for parameter estimation of Gaussian mixture models are the following:

- Spherical: variance is equal in all directions (where the directions are the daypart columns of the input matrix)
- Diagonal: each direction has a different variance
- Ellipsoidal: allows covariance terms to orient ellipse in different directions plus constraints regarding shape and volume of the Gaussian density functions

The Gaussian Mixture Model is computed for G clusters between 2 and 10. The optimum total amount of clusters is selected using the Integrated Completed Likelihood (ICL) criterion, and the model fit is done using the Bayesian Information Criterion (BIC). The key difference between the BIC and ICL is that the latter includes an additional term (the estimated mean entropy) that penalises clustering configurations exhibiting overlapping groups.

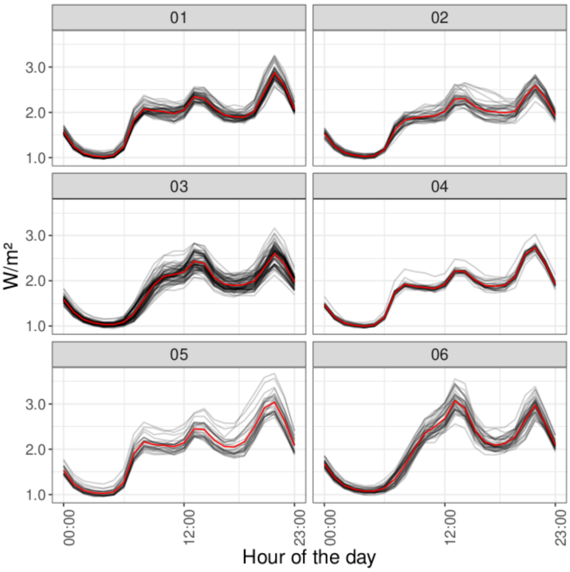


Figure 5.3: Clustering of the daily load curves, only using days which are presumably not affected by weather conditions. These six profiles represent the usage patterns of the case study

Finally, an important point regarding the usage pattern detection is that to infer patterns not accounting for the weather dependence or holidays component, a clustering-classification approach with a different subset of days is considered. The clustering technique explained above is used to detect the patterns from a subset of the daily load curves when low, or even null, weather dependence is expected (during March, April, May, September, October, and November).

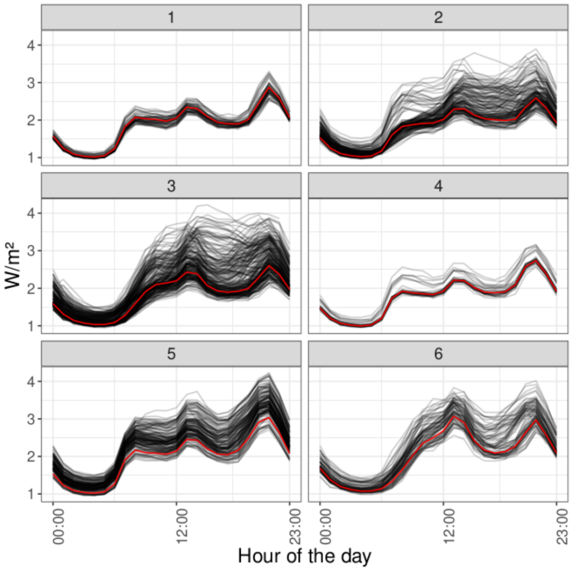


Figure 5.4: Classification of the complete series using the representative usage patterns detected with the clustering technique

Subsequently, in a second step, a classification of the rest of the daily load curves is predicted using the clustering model obtained in the first stage. An example of the clustering results is depicted in Figure 5.3. The red curves correspond to the usage patterns, and the black ones are the actual daily loads during the training phase of the clustering procedure. Using the same representation, the results of the classification stage are depicted in Figure 5.4, where the whole period, including winter and summer seasons, are considered. As it can be seen, the weather conditions' influence tends to increase energy consumption in certain usage patterns. However, in all cases, they tend to maintain the relative shape.

5.4.2 Regression model

The technique used to characterise the electricity consumption consists of a penalised multiple linear regression model. The terms of this model are explained more in detail in the following subsections. However, in essence, the consumption is decomposed into multiple parts: the usage patterns estimated with the previous clustering-classification technique; the calendar features, which allow modelling the hourly and weekly baseload patterns; and the weather features, which enable to estimate the increase in consumption when severe weather conditions occur. Equation 5.2 describes the major components of the penalised regression model.

$$Q_t^e = (B_t \times s_t) + (H_t \times dh_t) + (C_t \times dh_t) + \varepsilon_t \tag{5.2}$$

Where Q_e^t is the electricity consumption at instant t; B_t are the baseload terms interacting with the usage patterns (s_t), H_t and C_t are the weather dependence terms during heating and cooling periods interacting with the hour of the day (dh_t). Lastly, ε_t is the error term of the model, where $\varepsilon_t \sim N(0, \sigma^2)$ and i.i.d.

Baseload terms

The baseload component is one of the most significant parts of electricity consumption. The formal definition of baseload consumption consists of the minimum level of demand on an electrical grid over a span of time. However, in the framework of this methodology, it is understood as hourly consumption with no weather dependence at all. Hence, the baseload component only depends on the representative usage pattern and the calendar variables of a certain day. Given the regression model presented, differences in consumption along the week and the day are considered. For both of them, a Fourier series describing the weekly and daily cycle was used. This decomposition transformation reduces the dimension of the fitting problem in the cases where input variables are periodic. The baseload terms are described in detail in Equation 5.3.

$$B_t = \omega_b + S_{N_d}(p_t^d) + S_{N_w}(p_t^w) \tag{5.3}$$

$$S_{N_d}(p_t^d) = \sum_{n=1}^{N_d} \omega_{b,d,n,cos} \cos(2\pi n p_t^d) + \omega_{b,d,n,sin} \sin(2\pi n p_t^d) \qquad p_t^d = \frac{dh_t}{24} \tag{5.4}$$

$$S_{N_w}(p_t^w) = \sum_{n=1}^{N_w} \omega_{b,w,n,cos} \cos(2\pi n p_t^w) + \omega_{b,w,n,sin} \sin(2\pi n p_t^w) \qquad p_t^w = \frac{wh_t}{168} \tag{5.5}$$

Where ω_b is the linear intercept; $S_{N_d}(p_t^d)$ and $S_{N_w}(p_t^w)$ are the Fourier series of the daily and weekly cycles, where $\omega_{b,w,n,cos}$, $\omega_{b,w,n,sin}$, $\omega_{b,w,n,cos}$ and $\omega_{b,w,n,sin}$ are the coefficients estimated within the regression model, N_d and N_w are the number of harmonics of both series, and finally, p_t^d and p_t^w are the relative part the day or the week at instant t. The dh_t and wh_t variables mean the hour of the day and the hour of the week at instant t. The advantage of using the Fourier series is that it avoids the use of an excessive number of dummy variables which would require the fit of all-possible combinations (24 + 168 dummy variables, in the case of fitting the regression model using an hourly-frequency dataset, multiplied by the number of usage patterns detected in the clustering step). This transformation reduces the fitting problem to the number of harmonics considered (normally, between 3 and 5 harmonics per cycle), which are enough to infer the underlying correlation between the electricity consumption and the seasonal cycle without a considerable loss of information. Additionally, an interesting feature of the Fourier series transformations is that, in some sense, it coerces the regression to maintain a relationship between closer parts of the cycle and between the beginning and the end of the cycle itself.

Weather dependence components

Besides the baseload terms, heating and cooling dependent components account for the consumption related to weather conditions, energy performance and characteristics of the buildings, and Heating, Ventilation, and Air Conditioning (HVAC) systems operation.

These components estimate the increase in consumption due to weather severity. They are important to understanding electricity consumption and infer characteristics of how the reference building/dwelling in a certain zone is composed and operated. Ideally, one of the most interesting building characteristics that could be inferred using this type of modelling is the building envelope's Heat Transfer Coefficient (HTC). This coefficient highly depends on the considerations made during its definition. For instance, depending on the inclusion of certain phenomena, such as ventilation or air leakage, the HTC can be different. If ventilation and air infiltration are not considered, the HTC is calculated considering the energy transfer through the building envelope, i.e. all the surrounding surfaces of the building in contact with the outdoors, ground or other buildings. If they are considered, the energy transfer due to ventilation and air infiltrations is included in the HTC definition. Furthermore, to estimate HTC some variables are needed, such as indoor temperatures or performance characteristics regarding the HVAC systems installed in the buildings. Without this additional information, it becomes nearly impossible to estimate the HTC. Therefore, in the framework of this methodology, instead of characterising the HTC as a heat flow rate quantification, it is estimated as the change in electricity consumption, compared to the baseload, due to a variation in indoor-outdoor temperature difference. To do so, and considering that only the wind speed and the outdoor temperature are available, multiple-input transformations over these features account for the different interactions between the electricity consumption and the outdoor conditions.

The first transformation considers the temperature differences between a theoretical balance temperature and the actual outdoor temperature. The main reason is to overcome the non-linearities between the outdoor temperature and consumption. Furthermore, different balance temperatures are considered during the heating and cooling season, and during multiple parts of the day. This feature helps the model to characterise certain situations better. For instance, regions that require heating and cooling needs at the same time or significant differences of weather dependence along the day. The increase in consumption due to an increase of this feature tends to be more related to ventilation systems without heat recovery units or window operations. Physically, it could be translated into the colder or hotter outdoor air, compared to indoor air, which enters the building, increasing HVAC systems energy consumption.

The second transformation uses the product of the wind speed and the theoretical temperature difference obtained by the first transformation to correlate consumption and the air infiltrations caused by the infiltration of outside air into a building, typically through cracks in the building envelope, doors, windows, and chimneys. This infiltration is caused by wind, negative pressurisation of the building, and air buoyancy forces, commonly known as the stack effect. In general, the higher the product between wind speed and indoor-outdoor temperature difference, the more energy consumption is experienced due to air infiltrations. Making a similar interpretation as in the first transformation feature, HVAC systems need to increase consumption to maintain the normal indoor thermal conditions.

Finally, the third transformation is the consideration of low pass filters in the inputs of the model. Due to building inertia and heat transfer through the envelope, the indoor temperature of buildings does not react instantly to changes in the outdoor temperature. Then, to linearise the correlation between energy losses and energy consumption, a first-order low pass filter of the outdoor temperature T^o with a certain α parameter is considered. This tuned temperature is called $T^{o,lp}$, and, afterwards, it is transformed using the same differential process used in the first transformation. The low pass filter retains the slow undisturbed variations (signals with a low frequency), while the fast variations are damped (filtered). It allows transforming the temperature, used as input in the models, into a variable that better represents the system's dynamics, enhancing the model fitness. This transformation assumes that the dynamics of the buildings can be described by lumped parameter RC (resistance-condenser) models. In turn, this assumption means that the response in consumption due to envelope energy transfers can be modelled as a first-order low pass filter. To summarising, the space heating and cooling terms are mathematically described in Equations 5.6 and 5.7.

$$H_t = \omega^+_{h,lp} T^{h,lp}_t + \omega^+_h T^h_t + \omega^+_{ah} A^h_t \tag{5.6}$$

$$C_t = \omega^+_{c,lp} T^{c,lp}_t + \omega^+_c T^c_t + \omega^+_{ac} A^c_t \tag{5.7}$$

$$\begin{aligned} T^{h,lp}_t &= (T^{bal,c}_{dh_t} - T^{o,lp}_t) d_{s_t} \\ T^h_t &= (T^{bal,h}_{dh_t} - T^o_t) d_{s_t} \\ A^h_t &= W^s_t T^h_t d_{s_t} \end{aligned} \qquad \begin{aligned} T^{c,lp}_t &= (T^{o,lp}_t - T^{bal,c}_{dh_t}) d_{s_t} \\ T^c_t &= (T^o_t - T^{bal,c}_{dh_t}) d_{s_t} \\ A^c_t &= W^s_t T^c_t d_{s_t} \end{aligned}$$

$$T^{o,lp}_t = \begin{cases} \alpha T^o_t & \text{if } t = 0, \\ \alpha T^o_t + (1-\alpha) T^{o,lp}_{t-1} & \text{if } t > 0. \end{cases} \qquad \alpha = 1 - e^{-t_{sampling}/(2\pi\tau/24)}$$

$$d_{s_t} = \begin{cases} 1 & \text{if weather dependence in } s_t, \\ 0 & \text{if no weather dependence in } s_t. \end{cases}$$

Where: $\omega^+_{h,lp}$ is the always-positive linear coefficient for the heating dependent term that considers the thermal inertia of the reference building ($T^{h,lp}_t$), which is related to the heat losses through the envelope and is calculated as the difference between balance heating temperature ($T^{bal,h}_{dh_t}$) at the portion of the day (dh_t) and the low-pass filtered outdoor temperature ($T^{o,lp}_t$) at instant t; ω^+_h is the always-positive linear coefficient for the raw heating dependent term (T^h_t), which is usually related to ventilation heat losses, and it is calculated as the difference between balance heating temperature ($T^{bal,h}_{dh_t}$) at the part of the day (dh_t) and the raw outdoor temperature (T^o_t); ω^+_{ah} is the always-positive linear coefficient for the heat losses due to air infiltrations (A^h_t), which is the wind speed (W^s_t) multiplied by the raw heating dependent term (T^h_t); $\omega^+_{c,lp}$ is the always-positive linear coefficient for the cooling dependent term that considers the thermal inertia of the reference building ($T^{c,lp}_t$), which is related to the heat gains through the envelope and is calculated as the absolute difference between balance cooling temperature ($T^{bal,c}_{dh_t}$) at the part of the day (dh_t) and the low-pass filtered outdoor temperature ($T^{o,lp}_t$); ω^+_c is the always-positive linear coefficient for the raw cooling dependent term(T^c_t), which is usually related to ventilation heat gains, and it is calculated as the difference between balance cooling temperature ($T^{bal,c}_{dh_t}$) at the part of the day (dh_t) and the raw outdoor temperature (T^o_t); ω^+_{ac} is the always-positive linear coefficient for the heat gains due to air infiltrations (A^c_t), which is the wind speed (W^s_t) multiplied by the raw cooling dependent term (T^c_t). Besides, the α value of the low-pass-filtered outdoor temperature depends on the $t_sampling$, which is the number of measures per hour of consumption time series Q^e, and the τ thermal time constant, which defines the number of hours that the synthetic building reacts over a certain change in outdoor temperature. Last but not least, all the temperature differentials and air leakage terms are multiplied by a dummy variable which coerces weather dependence terms to 0 if a certain usage pattern has no weather dependence (ds_t).

Impact of holiday seasonality

After the first tests of the implementation, the authors detected that the influence of holidays tends to generate significant change points in electricity consumption for certain regions, sectors and periods of the year. In most cases, the holidays periods occurred in correspondence of national holidays, Fridays or Mondays between national holidays and weekends, winter and summer weekends, and the summer vacations. However, it was difficult to find a feature that linearly correlates the holidays component of the electricity consumption with the different local festivities of every region along the year. As a first attempt, some of the features that could be used are the number of tourists, second homes occupancy, or hotel bookings at the postal code level and daily frequency. However, this information was impossible to find at the desired aggregation levels. Therefore, another strategy is considered in the final implementation. The data-driven characterisation model is fitted using only those days that are not suitable to be holidays. Then, the whole period is predicted using the trained model and the residuals between the actual and predicted data during the holidays period are considered as the holiday's component. In addition, this holidays dependence component is estimated only when a difference of at least 20% is detected between the RMSE of the holidays / non-holidays period.

Impact of COVID-19 lockdown periods

The Covid-19 Spanish lockdown, during the period from March 15th to June 21st 2020, significantly affected the energy consumption either in residential, industrial or public sectors. Changes in business activities, user behaviour and building occupancy caused this situation. For the presented case study, the time period analysed depends on the availability of electricity consumption data for each postcode. In general, the evaluated period comprised mid-2018 to mid-2020. Thus, the data used to validate the characterisation methodology was fully affected by this lockdown period. A set of terms have been introduced into the regression model to quantify the decrease or increase in consumption due to the lockdown. They basically add an interaction of the lockdown period to the baseload terms and a set of re-adjusted weather dependence coefficients during the period. Another consideration made during this period is that holidays effect on energy consumption must be fixed to zero, as people should have stayed at home for those periods, except in particular cases.

Training of the model

The electricity time series considered during the training phase changes slightly depending on the economic sector considered. It clearly depends on the most representative area factor for each economic sector, as the characterisation outcomes are further compared among different regions. The built area normalisation becomes a key factor in assessing the energy performance of buildings. The ratios considered for each location and existing tariffs are the following:

- Residential sector:

$$Q^e = \frac{\text{Total consumption}^{residential}}{\text{number of contracts}^{residential} \times \text{average dwelling area}} [W/m^2]$$

- Industrial / Agriculture / Offices / Retail sector:

$$Q^{e} = \frac{\text{Total consumption}^{sector}}{\text{number of contracts}^{sector} \times \text{average building area}^{sector}} [W/m^{2}]$$

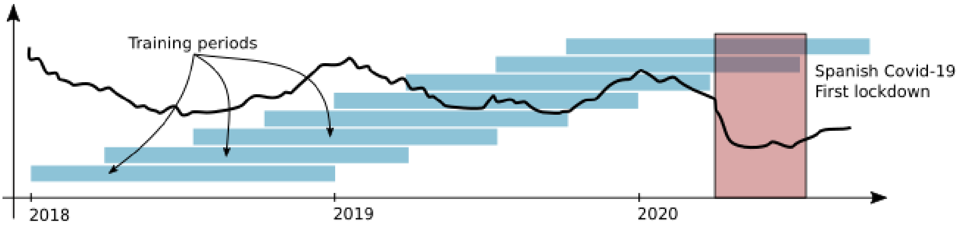


Figure 5.5: Model training periods to characterise the evolution in time of the dependencies

The model's training is recursively performed every three months over a one-year window, as is shown in Figure 5.5. This procedure provides information on how the reference building is evolving in time. So, the characterisation coefficients become, in some sense, time-variant. To decrease the computational time, the original hourly frequency of the input time series is resampled to 4 hours.

Regarding the estimation of the unknown terms, most of them are inferred through the maximum likelihood technique implemented in the penalised function of the R package Penalised [158], where the whole regression formula is estimated. However, several coefficients cannot be solved using this methodology, as they are variables that transform the model inputs themselves. Examples are the thermal time constant of the reference building, the number of harmonics of the Fourier series, or the balance temperatures, among others. The optimisation of these coefficients is made using a Genetic Algorithm (GA) that iterates and evolves chromosomes (in this case are the binary representation of the parameter values to optimise), minimising a cost function, which in this case is the Root Mean Square Error (RMSE) of the predicted consumption versus the metered consumption data. As a required initial input for the GA, a range of feasible values for each parameter to estimate is defined. In the case of $T_{d}^{bal,h}h$, the heating balance temperature range goes from 10 to 22 $°C$, in steps of 0.5. For $T_{d}^{bal,c}h$, the cooling balance temperature ranges between 18 to 30 $°C$, in steps of 0.5. The building thermal inertia parameter (τ) ranges between 1 to 48 hours in steps of 1. Finally, the boolean activators for the weather dependence in each daily seasonality (d_s) can be 0 or 1. In each training period, the initial parameters considered for the GA optimisation are the ones obtained in the last period, that is the reason to increase the number of maximum iteration permitted in the case of the first training period (50 vs. 20), when no initial values are available. The population considered in the GA is 300 for each iteration and the elitism in set to a 5%.

Known terms and time series: Q^e, s, p^d, p^w, dh, wh, T^o, W^s and $t_{sampling}$.

Unknown terms for each usage pattern: ω_b, $d_s(*)$, $\omega_{b,d,n,sin}$, $\omega_{b,d,n,cos}$, $\omega_{b,w,n,sin}$ and $\omega_{b,w,n,cos}$.

Unknown fixed terms: $\tau(*)$, N_d and N_w.

Unknown terms for each day part: $\omega^+_{h,lp}$, ω^+_h, ω^+_{ah} ,$\omega^+_{c,lp}$, ω^+_c, ω^+_{ac}, $T^{bal,h}_{dh}(*)$ and $T^{bal,c}_{dh}(*)$.

(*) Estimated using a genetic algorithm optimiser

5.5 Case study results

Rather than summarise in detail the results over the whole province of Lleida (Spain), which might be investigated in future studies, consumers in the residential sector of postal code 25006 are selected to show the intermediate and final results obtained during the validation procedure. This helps to focus on each of the results obtained concerning the models' accuracy and the estimated KPIs linked to the energy performance of buildings and usage patterns of their occupants.

5.5.1 Characterisation of a postal code

The postal code analysed is related to the Zona Alta neighbourhood in the city of Lleida. It is known as one of the most well-being districts in Lleida, at least compared to those near the city centre. Some of its socio-economic characteristics are household incomes of 36,498€ per year, incomes quantile 80-20 ratio of 3.23 (one of the highest of the province, which means there are large differences between low and high salaries), an average population age of 47.42 years, with 26.95% of people older than 65 and 13.59% under 18.

Estimated model

The accuracy of the models for each of the tariffs and evaluation periods are detailed in Table 5.2 and 5.3.

For both of the selected metrics, the average accuracy (MAPE: 5,04%, CVRMSE: 6,51%) is very high considering the characterisation purposes of this methodology. Even dealing with 4h-frequency predictions, the accuracy level reaches the state-of-the-art forecasting techniques at the city-scale level and daily aggregation. Figure 5.6 shows the energy signature between the 4h-resampled real observations and the predictions of the models. It has been proved that the predictions capture the main trend of the original data, and even the variance is extremely similar.

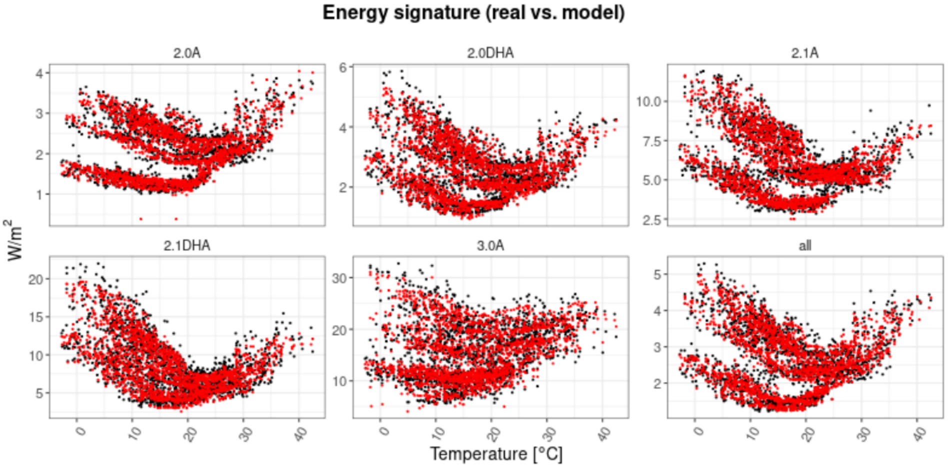


Figure 5.6: Predicted 4-hourly aggregated energy signature versus actual data

The weather-related coefficients are depicted in Figure 5.7. Dark blue lines correspond to the characterisation coefficients between June 2019 to May 2020 and the yellow ones from July 2018 to June 2019. In the Y-axis, the different weather dependence coefficients in heating and cooling modes are depicted. U_{raw} heating values are the ω_h^+ model coefficients depending dh_t (hour of the day), U_lp heating values are the $\omega_{h,lp}^+$ model coefficients depending the dh_t , I^air heating values are the ω_{ah}^+ model coefficients depending the dh_t , T^{bal} heating values are the heating balance temperature depending on the dh_t , τ is the thermal time constant of the building, U_{raw} cooling values are the ω_c^+ model coefficients depending on dh_t , U_lp cooling values are the $\omega_{c,lp}^+$ model coefficients depending on the dh_t , I^{air} cooling values are the ω_{ac}^+ model coefficients depending on the dh_t , and T^{bal} cooling values are the cooling balance temperature depending on the dh_t.

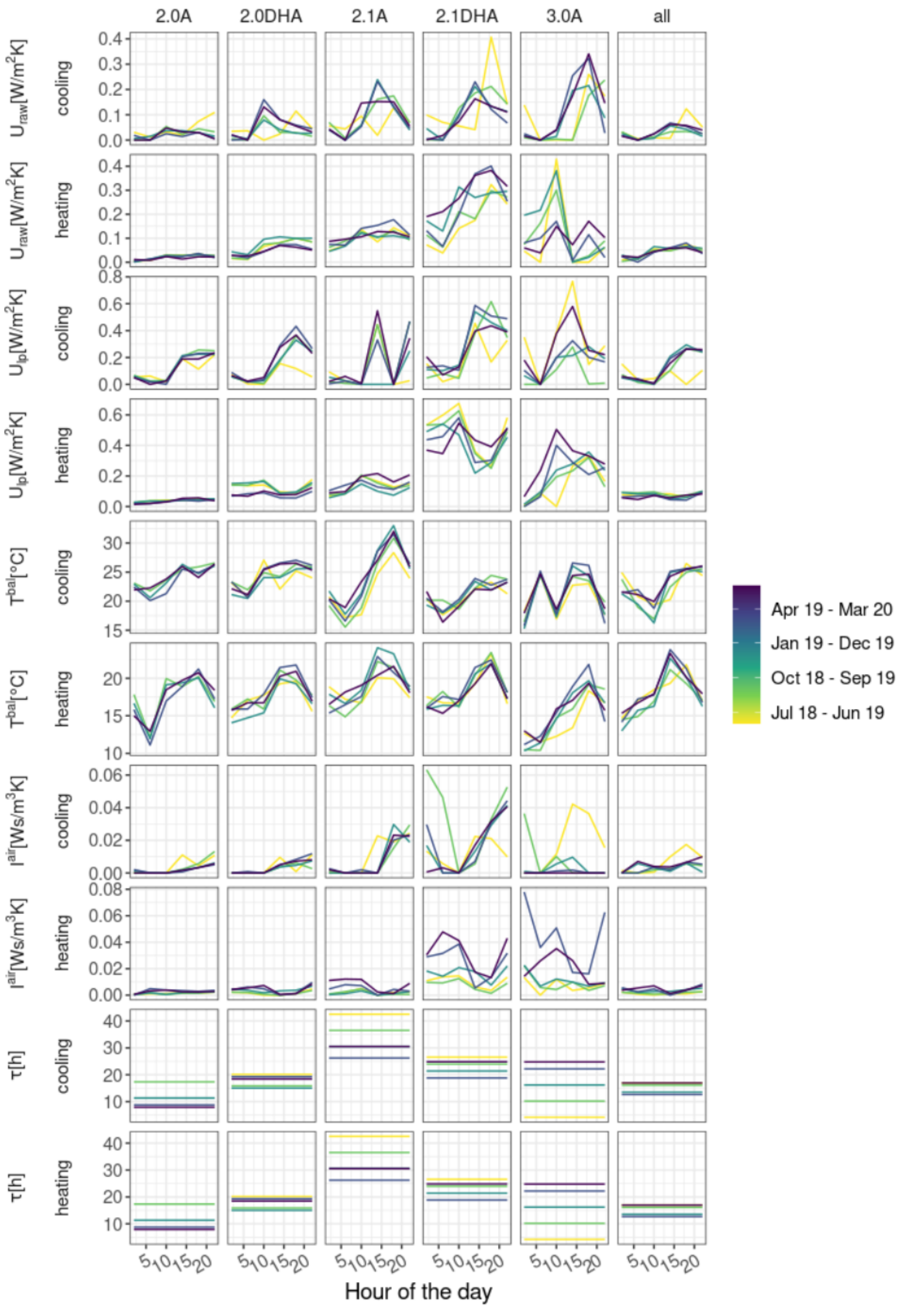


Figure 5.7: Weather-dependent characterisation parameters of the model

Period - MAPE [%]	2.0A	2.0DHA	2.1A	2.1DHA	3.0A	all
June 2018 - May 2019	4,52	7,05	5,78	8,28	7,03	5,33
Sept. 2018 - Aug. 2019	4,31	7,65	5,90	9,30	6,89	5,02
Dec. 2018 - Nov. 2019	4,18	6,25	5,56	8,36	5,92	4,73
Mar. 2019 - Feb. 2020	4,37	5,95	6,35	9,79	6,52	5,34
June 2019 - May 2020	4,15	5,32	5,57	8,69	7,35	4,77

Table 5.2: Mean Average Percentage Error (MAPE) over distinct periods and tariffs

Period - CVRMSE [%]	2.0A	2.0DHA	2.1A	2.1DHA	3.0A	all
June 2018 - May 2019	5,75	8,53	7,34	9,99	8,94	6,45
Sept. 2018 - Aug. 2019	5,68	9,08	7,56	10,68	8,55	6,55
Dec. 2018 - Nov. 2019	5,65	8,06	7,40	10,27	7,84	6,27
Mar. 2019 - Feb. 2020	5,95	7,73	8,25	12,40	8,61	7,03
June 2019 - May 2020	5,56	7,06	7,06	11,03	8,58	6,27

Table 5.3: Coefficient of Variation of the Root Mean Squared Error (CVRMSE) over distinct periods and tariffs

It can be seen that the coefficients across different tariffs vary largely and tend to be higher the more electricity is consumed by the tariff customers. This is a normal effect, as customers with 2.1 and 3.0 tariffs tend to have more domestic appliances or electrical driven HVAC equipment in their households. One of the most interesting insights is that space heating and cooling dependencies tend to differ widely along day time, responding with more emphasis to weather conditions during sunlight hours. Moreover, the estimated balance temperature helps to understand the most common HVAC operation schedule during a typical day, or, in other words, how people or energy managers tend to set the thermostats. Additionally, differences in the thermal time constant show variations in building's envelope characteristics between tariffs. At first glance, it seems that the 2.1A tariff is more related to higher thermal inertia buildings, which could also be related to better-insulated buildings. Regarding the baseload characterisation, each usage pattern's daily and weekly profile and tariffs are obtained using the model parameters.

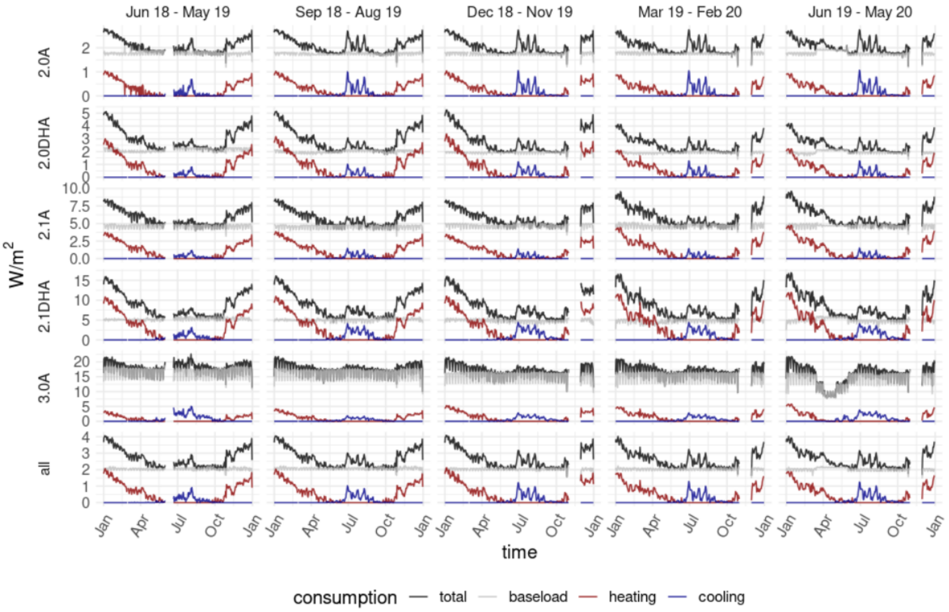


Figure 5.8: Daily electricity disaggregation results over distinct periods and tariffs

In summary, using the developed regression model, the decomposition of the three main components of buildings electricity loads (baseload, space heating and cooling) is made for the whole period of data within each of the evaluation periods (from June 2018 to May 2019, and from June 2019 to May 2020). In the web application, the results of this disaggregation are much better represented using interactive plots. However, to show the results in paper form, the Figure 5.8 represents the daily disaggregation and the total consumption. To compare the yearly evolution between different periods, the X-axis represents the months from January to December.

From Figure 5.8, it can be noted that, in all the cases, the significance of the baseload consumption is much higher than the weather dependence components. Also, the high variance in the baseload component in tariff 3.0A corresponds to the weekdays-weekends variation. Another detail that can be seen in this plot is the impact of the Covid-19 lockdown in Spain during the months from March to May of the last evaluation period, especially in the case of tariff 3.0A, where allegedly some business buildings/dwellings are integrated into the residential sector subset of the Datadis database. The evolution of the heating and cooling components through the year seems to fulfil the expected behaviour during a natural year, considering the total consumption series and the climate data of the case study area. However, it is noted that the reference building of tariff 2.1DHA has a major impact in terms of heating dependency. So, it can be interpreted that customers with this tariff have more electricity resourced heating systems compared to the customers with other tariffs.

Summarised KPIs

Once the characterisation model is technically fitted, a set of KPIs is defined to compare different areas, even when certain conditions differ widely from the type of users, weather conditions, or building characteristics. To do so, simple units and plots were chosen to represent the model results.

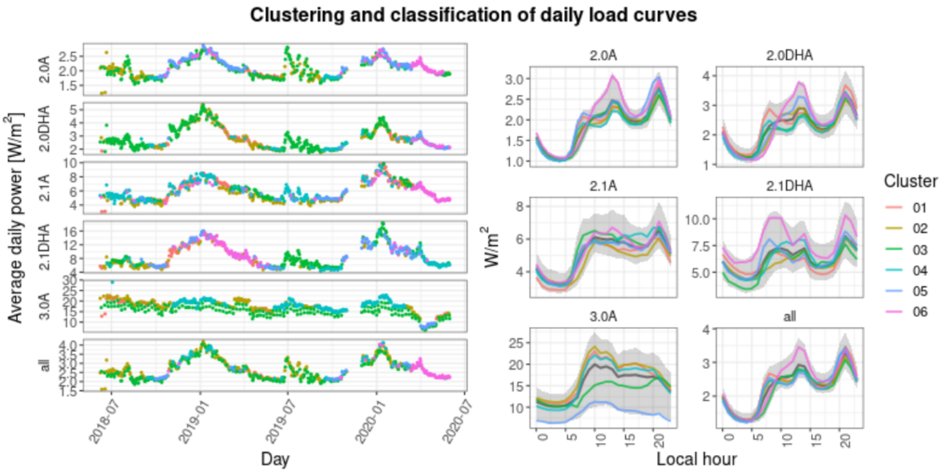


Figure 5.9: Usage patterns detected over distinct tariffs

The results of the clustering and classification of the usage patterns are illustrated in Figure 5.9. In the right pane, the different usage patterns in multiple colours are depicted, and in grey, the interval of daily load curves at confidence 95% is shown. In the left pane, the daily classification is represented, and it can be observed that some patterns have continuity in time. Hence, they tend to evolve over time, depending on certain conditions that interact with energy consumption.

These conditions are related to the weather, part of the year, holiday seasons and other unknown variables.

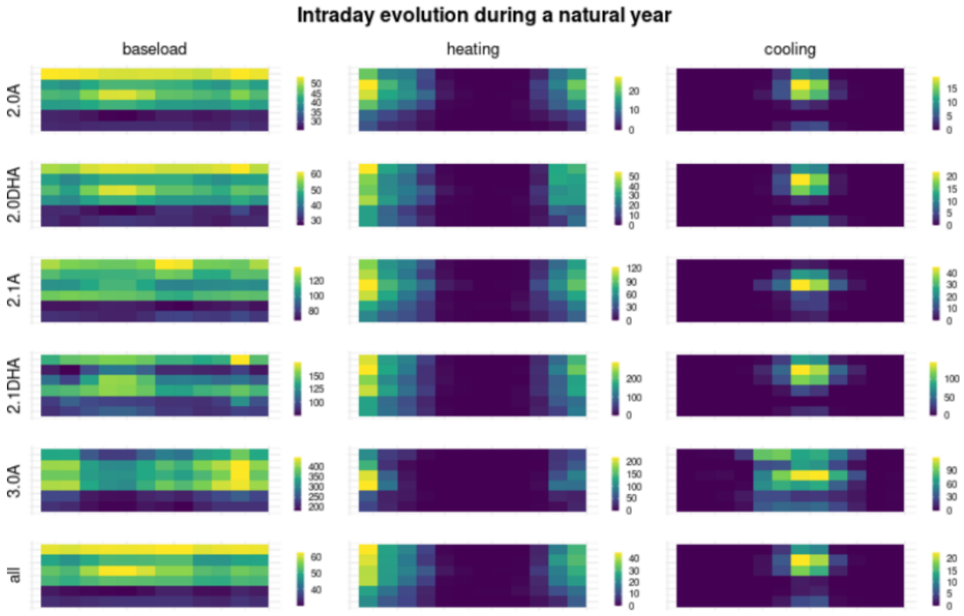


Figure 5.10: Intraday summarised electricity disaggregation results over a natural year and distinct tariffs

The heat map shown in 5.10 uses the most updated characterisation model (Trained with data from July 2019 to June 2020) to show the average kWh/year contribution of each electricity component by tariff through a natural year (X-axis, each step is one month) and the different parts of the day (Y-axis, each step are four hours). It can be seen that in the case of baseload, it seems that, during the Covid-19 confinement, it has been incremented by about 20% during the daytime period from 12 h to 16 h. This can be related to more people in their homes interacting with electricity-driven cooking systems during lunchtime. In contrast, 3.0A customers decrease their consumption drastically during those months. Regarding the heating and cooling components, it can be observed that the different intraday dependencies along different tariffs and months of the year (see Figure 5.10). Maybe, again, the 3.0A customers clearly behaved significantly different in terms of cooling dependency compared to customers with other tariffs. Besides increasing the understandability of the distribution between components and their evolution in time, the Figure 5.11 represents the relative disaggregation, on a natural year basis, between the baseload, the heating and the cooling components, and the impact of holidays and Covid-19 lockdown on the total consumption.

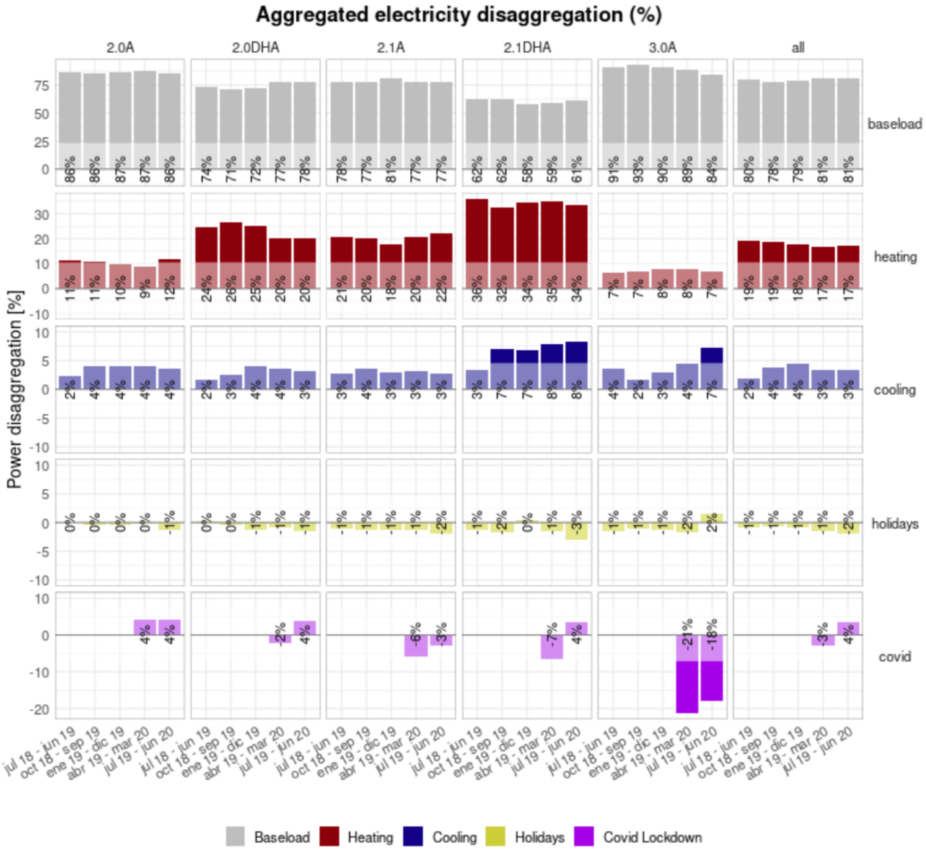


Figure 5.11: Yearly-aggregated relative segmentation of the electricity consumption over distinct periods and tariffs

For instance, concerning tariff 2.0A and the first period July 2018 to Jun 2019: the baseload component represents approximately 86% of the total annual consumption, the heating component the 11%, the cooling component represents 2%, and the holidays do not contribute at all. In this case, the Covid-19 lockdown had a shallow impact during the lockdown period (March 15th to June 21st 2020). Another conclusion is that the evolution of the different components in time is rather similar. However, large differences can be detected between different tariffs, and this corresponds to the different users/building typologies that characterise each tariff.

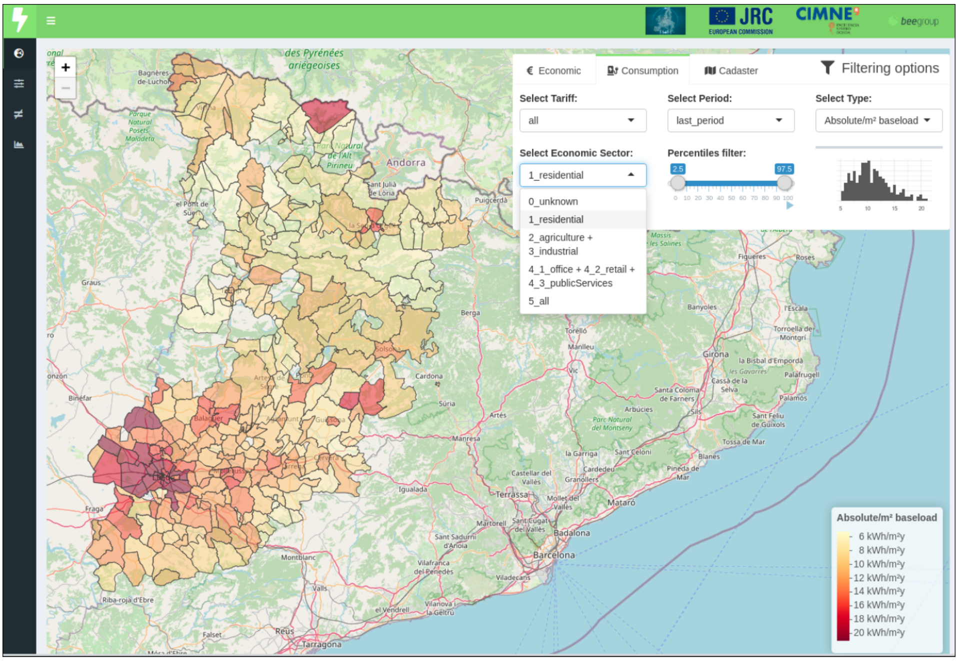
Besides the relative disaggregation, the web application also provides the point of view of the absolute consumption contribution in kWh per natural year. Using this representation, a decrease in total consumption for tariffs 3.0A and 2.0DHA is detected, especially the former, which is much more affected by the Covid-19 lockdown (approx. -20%, according to the relative segmentation results). Then, in general, for the rest of the tariffs, the same amount of total consumption during the whole evaluation period is observed.

5.5.2 Results at a province level

The characterisation results over the whole province will be described in further research publications. However, to show the web dashboard created for this purpose, a set of examples are described in the following paragraphs. This validation has been launched on a single server equipped with a 12-core 3.6 GHz CPU and 32 GB RAM. The execution of the model training algorithm and the calculation of all the KPIs related to all the historical periods available and all combinations of economic sector, postal codes, and tariffs available within the province of Lleida, took 18h. Once the aggregated consumption dataset of the whole month is gathered, the analysis can be reassessed, considering the new data, in less than 2.5 h. It means that the batch calculation on the same conditions for all the Spanish provinces would take less than six days. This computational cost is totally affordable considering the low cost of this type of server and a monthly basis update of the characterisation.

Figure 5.12 depicts the home section of the dashboard, whose purpose is to give a clear and simple visualisation of all the estimated consumption KPIs, cadastre information and socio-economic indicators on a map. The visualisation can be filtered by tariffs, economic sectors, periods, percentiles ranges. An interesting feature is a tiny histogram representing the distribution of values of the variable depicted on the map, especially when outliers can generate useless colouring legends.

The characterisation tab, shown in Figure 5.13, represents the complete assessment of the electricity consumption of a specific postal code and economic sector selected over the map. Several of the plots shown in this tab are interactive versions of the summarised KPIs explained in the subsection above, such as information about the model accuracy, the usage patterns detected and the disaggregation results in several time aggregations. The user can go deeper into the most common electricity uses over a certain geographical area.



Select Tariff: all

Select Period: last_period

Select Type: Absolute/m² baseload

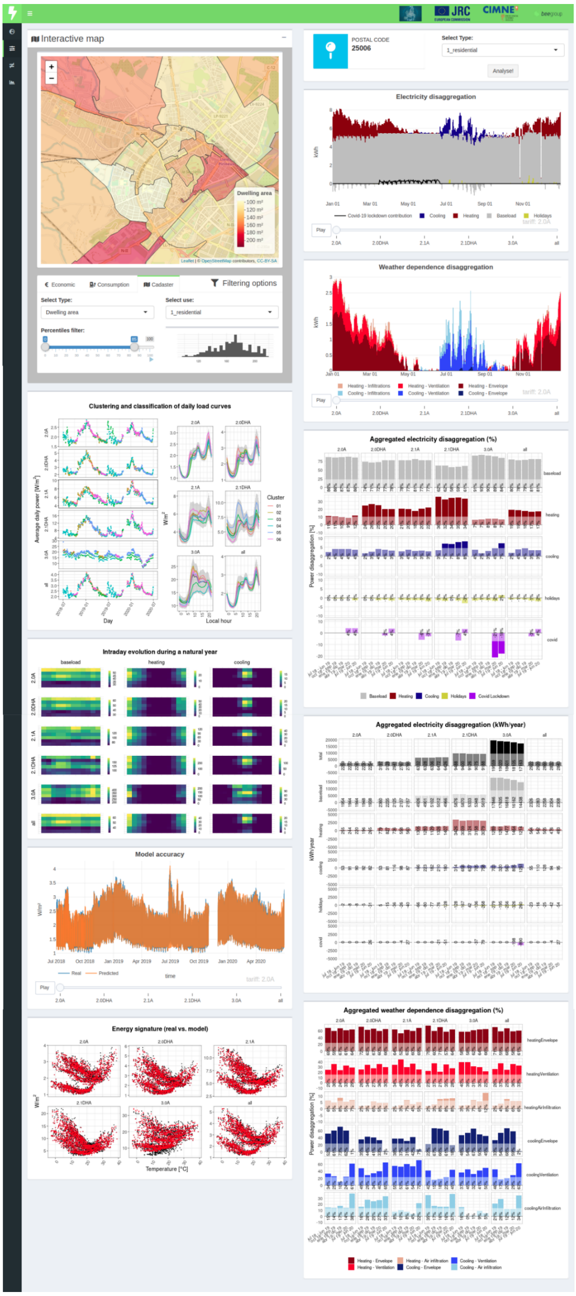
Select Economic Sector: 1_residential

Percentiles filter: 2.5 – 97.5

Figure 5.12: Web application - "KPIs on a map" tab

In Figure 5.14, the benchmarking tab is depicted, where the objective is to exploit the usage of the characterisation models to compare in detail two postal codes. This comparison is made by the estimated electricity components, normalising the results of the second postal code to the weather conditions and building/dwelling sizes of the first one. This normalisation procedure means that the divergence in electricity consumption should be caused by the difference in the energy performance of buildings, alternative usage patterns in electric devices, or by a different HVAC systems operation in cooling and heating electricity consumption components. In parallel, intraday differences along a natural year between the baseload consumption, and the impact of holidays and the Covid-19 lockdown period, are also represented.

Finally, Figure 5.15 shows the tab that allows cross-correlating all the KPIs to understand tendencies and relations between them, providing a wider interpretation of the territory and understanding if the variation of a certain cadastre or socio-economic indicator has a significant correlation to another estimated energy consumption KPI. For instance, it could be inferred if there is a relation between holidays periods contribution to the energy consumption and average percentage of single households, or the average annual incomes per person.



POSTAL CODE 25006

Select Type: 1_residential

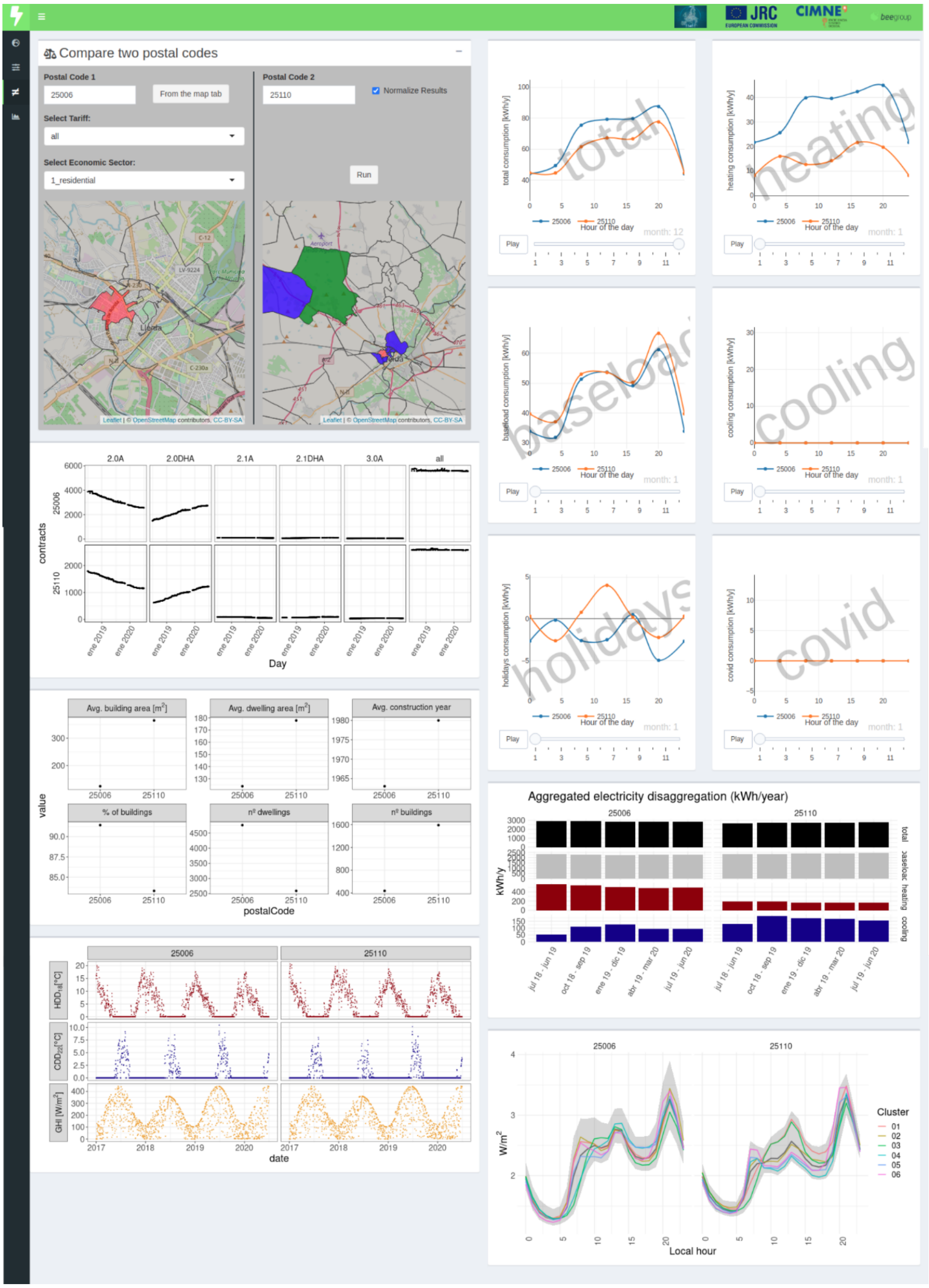
Analyse!

Select Type: Dwelling area

Select use: 1_residential

Percentiles filter:

Figure 5.13: Web application – “Characterisation” tab



Compare two postal codes

Postal Code 1: 25006

From the map tab

Select Tariff: all

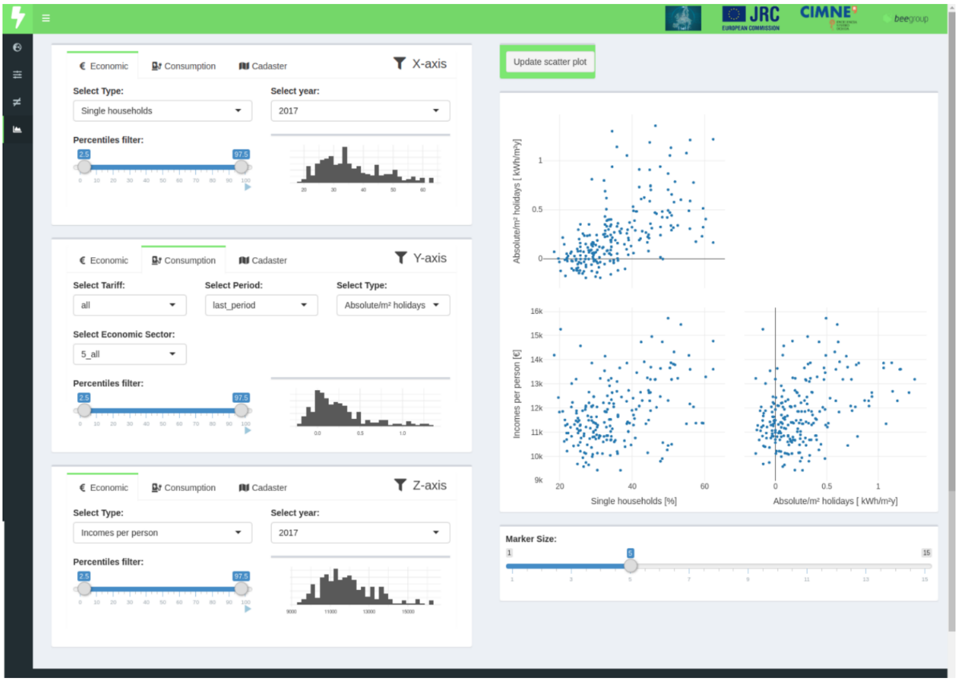
Select Economic Sector: 1_residential

Postal Code 2: 25110

☑ Normalize Results

Run

Figure 5.14: Web application – “Benchmarking” tab



Economic | Consumption | Cadaster — X-axis

Select Type: Single households

Select year: 2017

Percentiles filter: 2.5 – 97.5

Economic | Consumption | Cadaster — Y-axis

Select Tariff: all

Select Period: last_period

Select Type: Absolute/m² holidays

Select Economic Sector: 5_all

Percentiles filter: 2.5 – 97.5

Economic | Consumption | Cadaster — Z-axis

Select Type: Incomes per person

Select year: 2017

Percentiles filter: 2.5 – 97.5

Update scatter plot

Marker Size: 5

Figure 5.15: Web application - "KPIs correlation" tab

5.6 Conclusions

A methodology to characterise actual electricity consumption of large geographical areas has been developed, implemented and validated. It has been proven that the segmentation of the aggregated electricity time series provides multiple interesting possibilities to estimate KPIs related to energy performance buildings and occupants usage trends.

Moreover, it has been developed as an open-source platform able to extract information from publicly available data sources. This platform is split into two main parts: a back-end and a front-end. The former gathers, transforms and stores the data into databases. These data are accessible to data analysis tools designed to model the buildings' electricity consumption only using high-frequency time series data of actual consumption and weather data as the main inputs. The latter visualises the KPIs and the obtained outcomes through a purpose-built web application.

This research demonstrated that implementing this type of data-driven methodologies is feasible for large regions in Spain. Still, other European countries can also apply it as long as similar open data sources are available. The list of possible applications that could use the methodology and the web platform is pretty extensive, targeting different types of beneficiaries, from the public to the private sector.

Chapter 6

General discussion and conclusions

As said in Chapter 1, globally, the buildings' energy consumption meant 30% of all-sectors energy consumption during 2019. Hence, as it is one of the most important sectors, only surpassed by the industrial one (37%), it has been envisaged that a better predictability and energy characterisation of the building sector is a key factor to optimise the demand and maximise the usage of renewable energy sources.

This Thesis enlightens the importance of applying statistical learning techniques for a successful global energy transition in the buildings sector. Several techniques were used in this Thesis. Some are penalised linear models, autoregressive models, clustering techniques, optimisation algorithms, and several model feature transformations. In all cases, these methods were used to represent physical phenomena, such as space heating consumption at dwelling level, the operation of building-level HVAC systems, or the total energy consumption of districts.

It has been proved that even the geographical level of the applications was drastically divergent (dwelling level, building level, and district level), these techniques were handy for a wide range of applications. For instance, forecasting and simulation of energy consumption or thermal comfort conditions, estimation of energy flexibility in buildings, or inference of unknown characteristics contained intrinsically in the energy-related data.

In the following sections, each chapter discussion, conclusions and possible future work are summarised.

Common data analytics platform

The research presented in Chapter 2 indicates the importance of improving consumer energy efficiency by implementing a data management platform that is efficient and scalable. According to different pilot groups and users, the electricity savings achieved during the EMPOWERING project were 2 to 12%. However, improvements in the behavioural aspects of energy use have considerable potential. Customers' motivation also seems to be important, as better results were achieved when customers were involved. Energy savings are motivated by many reasons, and money savings is only one of them. In addition, environmental concerns, government regulations, social policies, and technological restrictions are powerful reasons why future services should diversify to increase their impact.

Furthermore, implementing this ICT platform meant an essential starting point for the applications presented in this Thesis (Chapters 3, 4, 5). For example, using a common data model to structure and store the data properly helps the reusability of analytical functionalities initially coded for another service or case study.

Modelling of thermal consumption in buildings

In Chapter 3, a methodology to virtually emulate the performance of thermostatic load controlled systems relying on statistical learning models derived from the information gathered by smart thermostats. Two regression-based models are developed: one with the supplied energy as the dependent variable (supply-side model), and another one with the indoor temperature as the dependent variable (demand-side model). Multiple exogenous variables, such as outdoor temperature, solar radiation, wind speed and wind direction, are considered in addition to multiple input transformation techniques that enhance these models' accuracy. Finally, a control algorithm, driven by the setpoint temperature, is implemented to couple both models and to be able to estimate the energy consumption and the indoor temperature when several setpoint temperature schedules are applied.

The methodology is validated in real cases within the winter season. One of the first findings is that the methodology used to train and couple the models, as well as the thermostatic control emulation, can be fully applicable to any space heating or cooling system as long as it is thermostatically controlled and a minimum historical data period is available. It has been shown that the models can accurately predict both the indoor temperature and the amount of energy used for space heating.

However, due to the high measurement tolerance of space heating consumption in this use case, a minimum of 30 days is recommended to determine the potential energy savings.

Among its major innovations, this methodology does more than predict heat consumption and indoor temperature. A mathematical optimization algorithm and control loop is integrated to simulate virtually all user-controlled modes driven by a setpoint temperature.

Another important finding of this research is that the analysed households' free-floating conditions can also be assessed accurately. This gives the opportunity, for instance, to estimate, during the winter season, the lower indoor temperature that a household would reach without the operation of the space heating system.

Even though this methodology's limitation is related to data quality requirements when the models are trained, during this training period, the setpoint temperatures of the buildings need to be excited in the range of evaluation of the setpoint temperature scenarios. This excitation generates dynamic changes in indoor temperature and heat consumption that the data-driven models subsequently infer. That means a minimum period of historical data of setpoint temperatures within the range of normal operation of indoor temperatures and space heating consumption is required.

Some conclusions can be drawn regarding the potential energy savings that may be achieved if users modify their usual setpoint temperature schedule. First, it has been demonstrated that lowering the usual set point temperature by 1°C can result in an average energy savings of 18.1%. Indeed, up to approximately 36.5% of energy savings can be achieved if the usual setpoint temperature is lowered by 2°C.

A further potential application of this research would be using this methodology as a forecasting toolbox for the short-term prediction of the impact, over the energy consumption and the indoor thermal conditions, of several set point temperature scenarios. For instance, this methodology could be used as the modelling part of a Model Predictive Control (MPC), which aims at minimizing the electricity cost of thermostatically controlled heat pumps due to market signals or at increasing the benefit of on-site renewable energy production (e.g., PV panels) while maintaining indoor comfort.

In fact, a similar modelling infrastructure has been implemented in the Sant Cugat pilot site presented in Chapter 4. In that case, a four-model strategy has been used. Firstly, using an ARX model for the water tank temperature modelling. Secondly, modelling, with another ARX model, the electric consumption provided by the heat pumps. And finally, using two GAM models for the energy demand forecasting of the office building and a local market.

Energy flexibility of building blocks

Furthermore, Chapter 4 confirmed that thermostatically controlled heat pumps could represent a huge potential for DR flexibility. It has been demonstrated that it is possible to manage clusters of heat pumps to respond to requests for DR flexibility. In addition, it has been proved that forecasting and optimization algorithms can be tailored to the particularities of each system configuration (e.g. HP interface, HP installation, and temperature sensors). Based on tests at three European pilot sites, it appears that heat pumps can be operated playing with the building inertia or storage tanks and are capable of being leveraged for multiple flexibility services. Nonetheless, many legacy systems have several issues that need to be resolved. These systems generally do not provide full interoperability with the heat pump, causing them to be restrictive and less flexible. Additionally, it has been confirmed that outdoor conditions, configured set points, and the available thermal storage strictly determine the duration for which the heat pump can be activated. This research also showed that the figure of a Cluster Manager plays a significant role in providing successful interoperability between the final users and the Aggregator under real-world market conditions.

Although the developed methodology to assess the flexibility in the different pilot sites shows promising outcomes to demonstrate its scalability and wider application, some procedures' limitations to determine the FF need further research. These limitations are mainly related to the non-accurate incorporation of the dynamic variability of the flexibility and the dependencies between the active energy and the activation variable. Both have been addressed in this research by including the autoregressive terms in the model. However, this procedure is not accurate enough and can miss some of the non-linearities. Therefore, some improvements should be addressed. As an example, recent papers [113] opened alternative methodologies to address these non-linearities in price-based DR schemes. These complementary approaches should be investigated in real practice experiences. Finally, simpler and more cost-efficient computational methods to evaluate the flexibility potential of large amounts of buildings and HVAC systems need to be further developed to assure a seamless connection with commercial practices of Aggregators and Cluster Managers in already existing European energy flexible markets.

Territorial energy consumption characterisation

In chapter 5, the development, implementation, and validation of a methodology to characterize actual electricity consumption in large geographical areas, such as districts, has been completed. Several methods were combined to infer significant relationships from the initial dataset. Initially, a clustering technique was used to detect the most representative daily seasonalities. Afterwards, a penalised linear regression methodology was applied to model the aggregated energy consumption per unit floor area in terms of calendar seasonalities and weather conditions. According to this research, multiple interesting KPIs related to buildings' energy performance and occupants' usage trends can be estimated from aggregated electricity time series.

Furthermore, this methodology has been implemented as an open-source platform to extract information from publicly available data sources. As is normally done in web applications, this software is split into two main parts: a back-end and a front-end. The former gathers, transforms and store the data into databases, where the latter version of the ICT architecture presented in Chapter 2 is used. The latter visualises the KPIs and the obtained outcomes through a purpose-built web interface. It has been shown that it is feasible to implement these types of data-driven methodologies for large regions in Spain. Even so, other European countries are also able to benefit from it as long as similar open data sources become available. Applications that could be built using the methodology and the web platform are quite varied, aiming at different types of beneficiaries:

1. Public authorities interested in improving the understanding of the energy consumption flows within their territory, producing better planning and optimal integration of renewable energies, prioritising the ECM implementation at the local level, or assessing ECM impacts over districts or regions.

2. Private companies aim to improve their marketing strategies based on the existing links between the territory and the electricity consumption trends.

During the validation of this methodology, an attempt has been made to include energy resources such as gas, biomass and oil in the analysis. So then, regardless of the implantation rate of the different energy resources in the building equipment (heating boilers, chillers, cooking equipment, domestic hot water), the interpretation of this characterisation could be understood as the performance of all buildings and their occupants against the final energy consumption. Nonetheless, the actual availability of big datasets containing high-frequency gas, biomass or oil consumption is meagre, especially for the residential sector. This point is significant in Spain, where the validation was conducted, and only electricity consumption data is available for many customers. However, this fact should evolve positively to implement global energy data-driven characterisation techniques in the mid and long term due to the pronounced tendency to electrify all-kind of building systems and the strong implementation of advanced meters for gas consumption.

To summarise, practical applications that could use the outcomes presented in this characterisation should assume that the methodology was only tested with electricity consumption. The inclusion of other final energy fuel types should slightly vary the data-driven modelling approach presented in this research and require another validation procedure with actual data.

Bibliography

[1] IEA, UCL, and BPIE. 2020 global status report for buildings and construction. Towards a zero-emissions, efficient and resilient buildings and construction sector. Tech. rep. United Nations - Global Alliance for Buildings and Construction.

[2] Energy Technology Perspectives 2020 – Analysis. en-GB.

[3] World Energy Outlook 2020 – Analysis. en-GB.

[4] Steve Mohr. Projection of world fossil fuel production with supply and demand interactions. en. PhD thesis. University of Newcastle, 2010.

[5] Antonio García-Olivares et al. A global renewable mix with proven technologies and common materials. en. In: *Energy Policy.* Modeling Transport (Energy) Demand and Policies 41 (Feb. 2012), ISSN: 0301-4215.

[6] Nima Norouzi, Maryam Fani, and Zahra Karami Ziarani. The fall of oil Age:A scenario planning approach over the last peak oil of human history by 2040. en. In: *Journal of Petroleum Science and Engineering* 188 (May 2020), ISSN: 0920-4105.

[7] Stated Policies Scenario – World Energy Model – Analysis. en-GB.

[8] Sustainable Development Scenario – World Energy Model – Analysis. en-GB.

[9] Net Zero by 2050 – Analysis. en-GB.

[10] Judit Kockat, Sheikh Zuhaib, and Oliver Rapf. A methodology for tracking decarbonisation action and impact of the buildings and construction sector globally. Tech. rep. United Nations - Global Alliance for Buildings and Construction, Dec. 2020.

[11] Delivering the European Green Deal. en. Text.

[12] Energy efficiency directive. en. Text. Aug. 2019.

[13] An Introduction to Statistical Learning. en-US.

[14] Luis Pérez-Lombard, José Ortiz, and Christine Pout. A review on buildings energy consumption information. en. In: *Energy and Buildings* 40.3 (Jan. 2008), ISSN: 0378-7788.

[15] Energy balance sheets - 2013 data - 2015 edition. en-GB.

[16] Kaile Zhou and Shanlin Yang. Understanding household energy consumption behavior: The contribution of energy big data analytics. en. In: *Renewable and Sustainable Energy Reviews* 56 (Apr. 2016), ISSN: 1364-0321.

[17] Steve Sorrell. Reducing energy demand: A review of issues, challenges and approaches. en. In: *Renewable and Sustainable Energy Reviews* 47 (July 2015), ISSN: 1364-0321.

[18] Karine Pollier, Lea Gynther, and Bruno Lapillonne. Energy Efficiency Trends and Policies in the Household and Tertiary Sectors. en. In: *ODYSSEE-MURE project* (),

[19] Linda Steg, Lieke Dreijerink, and Wokje Abrahamse. Factors influencing the acceptability of energy policies: A test of VBN theory. en. In: *Journal of Environmental Psychology* 25.4 (Dec. 2005), ISSN: 0272-4944.

[20] Benjamin K. Sovacool. What are we doing here? Analyzing fifteen years of energy scholarship and proposing a social science research agenda. en. In: *Energy Research & Social Science* 1 (Mar. 2014), ISSN: 2214-6296.

[21] William Prindle and Scott Finlinson. Chapter 11 - How Organizations Can Drive Behavior-Based Energy Efficiency. en. In: *Energy, Sustainability and the Environment*. Ed. by Fereidoon P. Sioshansi. Boston: Butterworth-Heinemann, Jan. 2011, pp. 305–335. ISBN: 978-0-12-385136-9.

[22] Gerald T. Gardner and Paul C. Stern. Environmental problems and human behavior. Environmental problems and human behavior. Pages: xiv, 369. Needham Heights, MA, US: Allyn & Bacon, 1996. ISBN: 978-0-205-15605-4.

[23] Tim Jackson. Motivating Sustainable Consumption: A Review of Evidence on Consumer Behaviour and Behavioural Change : a Report to the Sustainable Development Research Network. en. Google-Books-ID: 7CHDMgAACAAJ. Centre for Environmental Strategy, University of Surrey, 2005.

[24] Linda Steg. Promoting household energy conservation. In: *Energy Policy* 36.12 (2008). Publisher: Elsevier, ISSN: 0301-4215.

[25] Roy Thomas Fielding. Architectural styles and the design of network-based software architectures. AAI9980887 ISBN-10: 0599871180. phd. University of California, Irvine, 2000.

[26] Katarina Grolinger et al. Energy Forecasting for Event Venues: Big Data and Prediction Accuracy. en. In: *Energy and Buildings* 112 (Jan. 2016), ISSN: 0378-7788.

[27] Bing Dong, Cheng Cao, and Siew Eang Lee. Applying support vector machines to predict building energy consumption in tropical region. en. In: *Energy and Buildings* 37.5 (May 2005), ISSN: 0378-7788.

[28] F. J. Ardakani and M. M. Ardehali. Long-term electrical energy consumption forecasting for developing and developed economies based on different optimized models and historical data types. en. In: *Energy* 65 (Feb. 2014), ISSN: 0360-5442.

[29] Rishee K. Jain et al. Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. en. In: *Applied Energy* 123 (June 2014), ISSN: 0306-2619.

[30] L. Suganthi and Anand A. Samuel. Energy models for demand forecasting—A review. en. In: *Renewable and Sustainable Energy Reviews* 16.2 (Feb. 2012), ISSN: 1364-0321.

[31] Geoffrey K. F. Tso and Kelvin K. W. Yau. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. en. In: *Energy* 32.9 (Sept. 2007), ISSN: 0360-5442.

[32] Cheng Fan, Fu Xiao, and Shengwei Wang. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. en. In: *Applied Energy* 127 (Aug. 2014), ISSN: 0306-2619.

[33] Wessam El-Baz and Peter Tzscheutschler. Short-term smart learning electrical load prediction algorithm for home energy management systems. en. In: *Applied Energy* 147 (June 2015), ISSN: 0306-2619.

[34] Thomas Berthou et al. Development and validation of a gray box model to predict thermal behavior of occupied office buildings. en. In: *Energy and Buildings* 74 (May 2014), ISSN: 0378-7788.

[35] D. H. Vu, K. M. Muttaqi, and A. P. Agalgaonkar. A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables. en. In: *Applied Energy* 140 (Feb. 2015), ISSN: 0306-2619.

[36] A. S. Ahmad et al. A review on applications of ANN and SVM for building electrical energy consumption forecasting. en. In: *Renewable and Sustainable Energy Reviews* 33 (May 2014), ISSN: 1364-0321.

[37] J. Cortés et al. Energy Consumption Prediction by Using an Integrated Multidimensional Modeling Approach and Data Mining Techniques with Big Data. In: *ER Workshops*. 2014.

[38] Cheng Fan, Fu Xiao, and Chengchu Yan. A framework for knowledge discovery in massive building automation data and its application in building diagnostics. en. In: *Automation in Construction* 50 (Feb. 2015), ISSN: 0926-5805.

[39] Alfonso Capozzoli, Fiorella Lauro, and Imran Khan. Fault detection analysis using data mining techniques for a cluster of smart office buildings. en. In: *Expert Systems with Applications* 42.9 (June 2015), ISSN: 0957-4174.

[40] Franklin L. Quilumba et al. Using Smart Meter Data to Improve the Accuracy of Intraday Load Forecasting Considering Customer Behavior Similarities. In: *IEEE Transactions on Smart Grid* 6.2 (Mar. 2015). Conference Name: IEEE Transactions on Smart Grid, ISSN: 1949-3061.

[41] Gianfranco Chicco. Overview and performance assessment of the clustering methods for electrical load pattern grouping. en. In: *Energy*. 8th World Energy System Conference, WESC 2010 42.1 (June 2012), ISSN: 0360-5442.

[42] Daisuke Takaishi et al. Toward Energy Efficient Big Data Gathering in Densely Distributed Sensor Networks. In: *IEEE Transactions on Emerging Topics in Computing* 2.3 (Sept. 2014). Conference Name: IEEE Transactions on Emerging Topics in Computing, ISSN: 2168-6750.

[43] Zheng Yang and Burcin Becerik-Gerber. Modeling personalized occupancy profiles for representing long term patterns by using ambient context. en. In: *Building and Environment* 78 (Aug. 2014), ISSN: 0360-1323.

[44] Xiaoli Li, Chris P. Bowers, and Thorsten Schnier. Classification of Energy Consumption in Buildings With Outlier Detection. In: *IEEE Transactions on Industrial Electronics* 57.11 (Nov. 2010). Conference Name: IEEE Transactions on Industrial Electronics, ISSN: 1557-9948.

[45] IEA Annex 31 Energy Related Environmental Impact of Buildings, 1996-1999.

[46] Zhun Yu et al. A systematic procedure to study the influence of occupant behavior on building energy consumption. en. In: *Energy and Buildings* 43.6 (June 2011), ISSN: 0378-7788.

[47] EVRY. Big data in banking. for marketers - How to derive value from big data.

[48] W. Liu and E.K. Park. Big Data as an e-Health Service. In: *2014 ICNC*. Feb. 2014, pp. 982–988.

[49] Jinsong Wu et al. Big Data Meet Green Challenges: Big Data Toward Green Applications. In: *IEEE Systems Journal* 10.3 (Sept. 2016). Conference Name: IEEE Systems Journal, ISSN: 1937-9234.

[50] Nanpeng Yu et al. Big data analytics in power distribution systems. In: *2015 IEEE PES ISGT*. Feb. 2015, pp. 1–5.

[51] Cisco Visual Networking Index: Forecast and Methodology, 2014-2019 White Paper. en. In: (2015),

[52] Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. In: *Communications of the ACM* 51.1 (Jan. 2008), ISSN: 0001-0782.

[53] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. en. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001). _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9868.00293, ISSN: 1467-9868.

[54] Bruce D. Meyer. Natural and Quasi-Experiments in Economics. In: *Journal of Business & Economic Statistics* 13.2 (1995). Publisher: [American Statistical Association, Taylor & Francis, Ltd.], ISSN: 0735-0015.

[55] Eurostat. Statistics Explained. Energy consumption in households. ISSN 2443-8219. en. Tech. rep. 2021.

[56] European European. Union. Clean energy for all Europeans. Publications Office of the European Union. In: (2019).

[57] Anne Stafford. An exploration of load-shifting potential in real in-situ heat-pump/gas-boiler hybrid systems. In: *Building Services Engineering Research and Technology* 38.4 (2017),

[58] Matteo Rivoire et al. Assessment of Energetic, Economic and Environmental Performance of Ground-Coupled Heat Pumps. en. In: *Energies* 11.8 (Aug. 2018),

[59] Xi Zhang et al. Economic assessment of alternative heat decarbonisation strategies through coordinated operation with electricity system – UK case study. In: *Applied Energy* 222 (July 2018), ISSN: 0306-2619.

[60] Stephen Clegg and Pierluigi Mancarella. Integrated electricity-heat-gas modelling and assessment, with applications to the Great Britain system. Part II: Transmission network analysis and low carbon technology and resilience case studies. In: *Energy* (Feb. 2018). ISSN: 0360-5442.

[61] M. Jarre et al. Opportunities for heat pumps adoption in existing buildings: real-data analysis and numerical simulation. In: *Energy Procedia*. Sustainability in Energy and Buildings 2017: Proceedings of the Ninth KES International Conference, Chania, Greece, 5-7 July 2017 134 (Oct. 2017), ISSN: 1876-6102.

[62] European. Commission - COM(2014) 356 final. Benchmarking smart metering deployment in the EU-27 with a focus on electricity. In: (2014).

[63] H. Farhangi. The path of the smart grid. In: *IEEE Power and Energy Magazine* 8.1 (Jan. 2010), ISSN: 1540-7977.

[64] Gary Newe. Delivering the Internet of Things. In: *Network Security* 2015.3 (Mar. 2015), ISSN: 1353-4858.

[65] Marco Pritoni et al. Energy efficiency and the misuse of programmable thermostats: The effectiveness of crowdsourcing for understanding household behavior. en. In: *Energy Research & Social Science* 8 (July 2015), ISSN: 22146296.

[66] Therese Peffer et al. How people use thermostats in homes: A review. In: *Building and Environment* 46.12 (Dec. 2011), ISSN: 0360-1323.

[67] LLC Apex Analytics. Energy Trust of Oregon Smart Thermostat Pilot Evaluation. en. In: (Jan. 2016),

[68] Diba Malekpour Koupaei et al. An assessment of opinions and perceptions of smart thermostats using aspect-based sentiment analysis of online reviews. en. In: *Building and Environment* 170 (Mar. 2020), ISSN: 0360-1323.

[69] Brent Huchuk, William O'Brien, and Scott Sanner. A longitudinal study of thermostat behaviors based on climate, seasonal, and energy price considerations using connected thermostat data. en. In: *Building and Environment* 139 (July 2018), ISSN: 0360-1323.

[70] Helen Stopps and Marianne F. Touchie. Managing thermal comfort in contemporary high-rise residential buildings: Using smart thermostats and surveys to identify energy efficiency and comfort opportunities. en. In: *Building and Environment* 173 (Apr. 2020), ISSN: 0360-1323.

[71] Brent Huchuk, Scott Sanner, and William O'Brien. Comparison of machine learning models for occupancy prediction in residential buildings using connected thermostat data. en. In: *Building and Environment* 160 (Aug. 2019), ISSN: 0360-1323.

[72] D Parker, K Sutherland, and D Chasar. Evaluation of the Space Heating and Cooling Energy Savings of Smart Thermostats in a Hot-Humid Climate using Long-term Data. en. In: *ACEEE Summer Study on Energy Efficiency in Buildings* (2016),

[73] Marco Pritoni, Jonathan M. Woolley, and Mark P. Modera. Do occupancy-responsive learning thermostats save energy? A field study in university residence halls. en. In: *Energy and Buildings* 127 (Sept. 2016), ISSN: 03787788.

[74] H Stopps and M F Touchie. Reduction of HVAC system runtime due to occupancy-controlled smart thermostats in contemporary multi-unit residential building suites. en. In: *IOP Conference Series: Materials Science and Engineering* 609 (Oct. 2019), ISSN: 1757-899X.

[75] M.M. Manning et al. The effects of thermostat setback and setup on seasonal energy consumption, surface temperatures, and recovery times at the CCHT twin house research facility. In: vol. 113 PART 1. 2007, pp. 630–641.

[76] Božidar Soldo. Forecasting natural gas consumption. In: *Applied Energy* 92 (Apr. 2012), ISSN: 0306-2619.

[77] Ehsan Tavakoli and Nader Montazerin. Stochastic analysis of natural gas consumption in residential and commercial buildings. In: *Energy and Buildings* 43.9 (Sept. 2011), ISSN: 0378-7788.

[78] Longquan Diao et al. Modeling energy consumption in residential buildings: A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation. In: *Energy and Buildings* 147 (July 2017), ISSN: 0378-7788.

[79] Marek Brabec et al. A nonlinear mixed effects model for the prediction of natural gas consumption by individual customers. In: *International Journal of Forecasting*. Energy Forecasting 24.4 (Oct. 2008), ISSN: 0169-2070.

[80] Božidar Soldo et al. Improving the residential natural gas consumption forecasting models by using solar radiation. In: *Energy and Buildings* 69 (Feb. 2014), ISSN: 0378-7788.

[81] Wancheng Li et al. Stepwise calibration for residential building thermal performance model using hourly heat consumption data. In: *Energy and Buildings* 181 (Dec. 2018), ISSN: 0378-7788.

[82] Junke Wang et al. Predicting home thermal dynamics using a reduced-order model and automated real-time parameter estimation. In: *Energy and Buildings* 198 (Sept. 2019), ISSN: 0378-7788.

[83] Alessandro Aliberti et al. A Non-Linear Autoregressive Model for Indoor Air-Temperature Predictions in Smart Buildings. en. In: *Electronics* 8.9 (Sept. 2019),

[84] Abdulrahman Alanezi, Kevin P. Hallinan, and Rodwan Elhashmi. Using Smart-WiFi Thermostat Data to Improve Prediction of Residential Energy Consumption and Estimation of Savings. en. In: *Energies* 14.1 (Jan. 2021). Number: 1 Publisher: Multidisciplinary Digital Publishing Institute,

[85] Vincenzo Trovato, Antonio De Paola, and Goran Strbac. Distributed Control of Clustered Populations of Thermostatic Loads in Multi-Area Systems: A Mean Field Game Approach. en. In: *Energies* 13.24 (Jan. 2020). Number: 24 Publisher: Multidisciplinary Digital Publishing Institute,

[86] Ahmet Doğan and Mustafa Alçı. Real-time demand response of thermostatic load with active control. en. In: *Electrical Engineering* 100.4 (Dec. 2018), ISSN: 1432-0487.

[87] Henrik Aalborg Nielsen and Henrik Madsen. Modelling the heat consumption in district heating systems using a grey-box approach. en. In: *Energy and Buildings* 38.1 (Jan. 2006), ISSN: 03787788.

[88] Peder Bacher et al. Short-term heat load forecasting for single family houses. en. In: *Energy and Buildings* 65 (Oct. 2013), ISSN: 03787788.

[89] Luca Scrucca. GA: A Package for Genetic Algorithms in R. en. In: *Journal of Statistical Software* 53.1 (Apr. 2013), ISSN: 1548-7660.

[90] G. Mor et al. EMPOWERING, a Smart Big Data Framework for Sustainable Electricity Suppliers. In: *IEEE Access* 6 (2018), ISSN: 2169-3536.

[91] Apple. Dark Sky API. https://darksky.net/dev.

[92] Council of the European Union , European Parliament. Regulation (EU) No 377/2014 of the European Parliament and of the Council of 3 April 2014 establishing the Copernicus Programme and repealing Regulation (EU) No 911/2010 Text with EEA relevance. Apr. 2014.

[93] C. Lo and N. Ansari. Decentralized Controls and Communications for Autonomous Distribution Networks in Smart Grid. In: *IEEE Transactions on Smart Grid* 4.1 (Mar. 2013). Conference Name: IEEE Transactions on Smart Grid, ISSN: 1949-3061.

[94] Peter D. Lund et al. Review of energy system flexibility measures to enable high levels of variable renewable electricity. en. In: *Renewable and Sustainable Energy Reviews* 45 (May 2015), ISSN: 1364-0321.

[95] D. S. Kirschen et al. Flexibility from the demand side. In: *2012 IEEE Power and Energy Society General Meeting.* ISSN: 1944-9925. July 2012, pp. 1–6.

[96] Stefan Weitemeyer et al. Integration of Renewable Energy Sources in future power systems: The role of storage. en. In: *Renewable Energy* 75 (Mar. 2015), ISSN: 0960-1481.

[97] P. Denholm et al. Role of Energy Storage with Renewable Electricity Generation. English. Tech. rep. NREL/TP-6A2-47187. National Renewable Energy Lab. (NREL), Golden, CO (United States), Jan. 2010.

[98] Pierluigi Siano. Demand response and smart grids—A survey. en. In: *Renewable and Sustainable Energy Reviews* 30 (Feb. 2014), ISSN: 1364-0321.

[99] Vicenzo Giordano et al. Smart grid projects in Europe: lessons learned and current developments 2012 update. . 2013.

[100] Zheng Ma, Joy Dalmacio Billanes, and Bo Nørregaard Jørgensen. Aggregation Potentials for Buildings—Business Models of Demand Response and Virtual Power Plants. en. In: *Energies* 10.10 (Oct. 2017). Number: 10 Publisher: Multidisciplinary Digital Publishing Institute,

[101] Xiaodong Cao, Xilei Dai, and Junjie Liu. Building energy-consumption status worldwide and the state-of-the-art technologies for zero-energy buildings during the past decade. en. In: *Energy and Buildings* 128 (Sept. 2016), ISSN: 0378-7788.

[102] Thomas M. Lawrence et al. Ten questions concerning integrating smart buildings into the smart grid. en. In: *Building and Environment* 108 (Nov. 2016), ISSN: 0360-1323.

[103] Gianluca Serale et al. Model Predictive Control (MPC) for Enhancing Building and HVAC System Energy Efficiency: Problem Formulation, Applications and Opportunities. en. In: *Energies* 11.3 (Mar. 2018). Number: 3 Publisher: Multidisciplinary Digital Publishing Institute,

[104] Peter Kohlhepp et al. Large-scale grid integration of residential thermal energy storages as demand-side flexibility resource: A review of international field studies. en. In: *Renewable and Sustainable Energy Reviews* 101 (Mar. 2019), ISSN: 1364-0321.

[105] J. Le Boudec and D. Tomozei. Demand response using service curves. In: *2011 2nd IEEE PES ISGT*. ISSN: 2165-4824. Dec. 2011, pp. 1–8.

[106] Cherrelle Eid et al. Managing electric flexibility from Distributed Energy Resources: A review of incentives for market design. en. In: *Renewable and Sustainable Energy Reviews* 64 (Oct. 2016), ISSN: 13640321.

[107] Alessia Arteconi, Alice Mugnini, and Fabio Polonara. Energy flexible buildings: A methodology for rating the flexibility performance of buildings with electric heating and cooling systems. en. In: *Applied Energy* 251 (Oct. 2019), ISSN: 0306-2619.

[108] Christian Finck et al. Quantifying demand flexibility of power-to-heat and thermal energy storage in the control of building heating systems. en. In: *Applied Energy* 209 (Jan. 2018), ISSN: 0306-2619.

[109] Glenn Reynders et al. Energy flexible buildings: An evaluation of definitions and quantification methodologies applied to thermal storage. In: *Energy and Buildings* 166 (May 2018), ISSN: 0378-7788.

[110] Rami El Geneidy and Bianca Howard. Contracted energy flexibility characteristics of communities: Analysis of a control strategy for demand response. en. In: *Applied Energy* 263 (Apr. 2020), ISSN: 0306-2619.

[111] Adamantios Bampoulas et al. A fundamental unified framework to quantify and characterise energy flexibility of residential buildings with multiple electrical and thermal energy systems. en. In: *Applied Energy* 282 (Jan. 2021), ISSN: 0306-2619.

[112] Rune Grønborg Junker et al. Characterizing the energy flexibility of buildings and districts. In: *Applied Energy* 225 (Sept. 2018), ISSN: 0306-2619.

[113] Rune Grønborg Junker et al. Stochastic nonlinear modelling and application of price-based energy flexibility. en. In: *Applied Energy* 275 (Oct. 2020), ISSN: 0306-2619.

[114] https://www.next-kraftwerke.de/wissen/intraday-handel. de-DE. May 2014.

[115] https://www.next-kraftwerke.de/wissen/regelenergie. Nov. 2010.

[116] https://bit.ly/3xkNuj9. de-DE. Feb. 2020.

[117] https://www.next-kraftwerke.de/wissen/day-ahead-handel. 2014.

[118] https://www.next-kraftwerke.be/en/knowledge-hub/intraday-trading/.

[119] F. Ocker, S. Braun, and C. Will. Design of European balancing power markets. In: *2016 13th International Conference on the European Energy Market (EEM)*. ISSN: 2165-4093. June 2016, pp. 1–6.

[120] Matthias D. Galus. Smart grid roadmap and regulation approaches in Switzerland. en. In: *CIRED - Open Access Proceedings Journal* 2017.1 (Oct. 2017). Publisher: IET Digital Library, ISSN: 2515-0855.

[121] J. Cipriano et al. Evaluation of a multi-stage guided search approach for the calibration of building energy simulation models. In: *Energy and Buildings. ENB-D-14-01064R1* ().

[122] P. Bacher and H. Madsen. Identifying suitable models for the heat dynamics of buildings. In: *Energy and Buildings* 43 (2011),

[123] Benedetto Grillone et al. A review of deterministic and data-driven methods to quantify energy efficiency savings and to predict retrofitting scenarios in buildings. en. In: *Renewable and Sustainable Energy Reviews* 131 (Oct. 2020), ISSN: 1364-0321.

[124] Frédéric Amblard et al. D.3.1-Optimization strategies for the use case scenarios. SIM4BLOCKS H2020 project. Grant agreement n° 695965. Tech. rep. Deliverable D.3.1. SIm4Blocks project, 2018.

[125] Trevor Hastie and Robert Tibshirani. Generalized additive models. en. In: (1990),

[126] Francesco D'Ettorre et al. A set of comprehensive indicators to assess energy flexibility: a case study for residential buildings. en. In: *E3S Web of Conferences* 111 (2019). Publisher: EDP Sciences, ISSN: 2267-1242.

[127] Laura Romero Rodríguez et al. Heuristic optimization of clusters of heat pumps: A simulation and case study of residential frequency reserve. en. In: *Applied Energy* 233-234 (Jan. 2019), ISSN: 0306-2619.

[128] Thomas Schütz et al. Comparison of models for thermal energy storage units and heat pumps in mixed integer linear programming. ECOS 2015-28th International Conference on Efficiency, Cost, Optimization, Simulation and Environmental Impact of Energy Systems. In: Pau, France, July 2015.

[129] Directive (EU) 2018/844 of the European Parliament and of the Council of 30 May 2018 amending Directive 2010/31/EU on the energy performance of buildings and Directive 2012/27/EU on energy efficiency (Text with EEA relevance). en. Code Number: 156. June 2018.

[130] Smart Metering deployment in the European Union | JRC Smart Electricity Systems and Interoperability. https://ses.jrc.ec.europa.eu/smart-metering-deployment-european-union.

[131] Jimeno A. Fonseca and Arno Schlueter. Integrated model for characterization of spatiotemporal building energy consumption patterns in neighborhoods and city districts. en. In: *Applied Energy* 142 (Mar. 2015), ISSN: 0306-2619.

[132] J. Langevin et al. Developing a common approach for classifying building stock energy models. en. In: *Renewable and Sustainable Energy Reviews* 133 (Nov. 2020), ISSN: 1364-0321.

[133] Lukas G. Swan and V. Ismet Ugursal. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. en. In: *Renewable and Sustainable Energy Reviews* 13.8 (Oct. 2009), ISSN: 1364-0321.

[134] Narjes Abbasabadi and Mehdi Ashayeri. Urban energy use modeling methods and tools: A review and an outlook. en. In: *Building and Environment* 161 (Aug. 2019), ISSN: 0360-1323.

[135] Nina Voulis, Martijn Warnier, and Frances M. T. Brazier. Understanding spatio-temporal electricity demand at different urban scales: A data-driven approach. en. In: *Applied Energy* 230 (Nov. 2018), ISSN: 0306-2619.

[136] N. Voulis, M. Warnier, and F. M. T. Brazier. Statistical Data-Driven Regression Method for Urban Electricity Demand Modelling. In: *2018 IEEE EEEIC*. June 2018, pp. 1–6.

[137] Constantine E. Kontokosta and Christopher Tull. A data-driven predictive model of city-scale energy use in buildings. en. In: *Applied Energy* 197 (July 2017), ISSN: 0306-2619.

[138] Marta J. N. Oliveira Panão and Miguel C. Brito. Modelling aggregate hourly electricity consumption based on bottom-up building stock. en. In: *Energy and Buildings* 170 (July 2018), ISSN: 0378-7788.

[139] Magnus Österbring et al. A differentiated description of building-stocks for a georeferenced urban bottom-up building-stock model. en. In: *Energy and Buildings* 120 (May 2016), ISSN: 0378-7788.

[140] Laura Romero Rodríguez et al. Mitigating energy poverty: Potential contributions of combining PV and building thermal mass storage in low-income households. en. In: *Energy Conversion and Management* 173 (Oct. 2018), ISSN: 0196-8904.

[141] João Pedro Gouveia, Pedro Palma, and Sofia G. Simoes. Energy poverty vulnerability index: A multidimensional tool to identify hotspots for local action. en. In: *Energy Reports* 5 (Nov. 2019), ISSN: 2352-4847.

[142] J. Kwac, J. Flora, and R. Rajagopal. Household Energy Consumption Segmentation Using Hourly Data. In: *IEEE Transactions on Smart Grid* 5.1 (Jan. 2014). Conference Name: IEEE Transactions on Smart Grid, ISSN: 1949-3061.

[143] João Pedro Gouveia, Júlia Seixas, and Ana Mestre. Daily electricity consumption profiles from smart meters - Proxies of behavior for space heating and cooling. en. In: *Energy* 141 (Dec. 2017), ISSN: 0360-5442.

[144] Christoffer Rasmussen et al. Method for Scalable and Automatised Thermal Building Performance Documentation and Screening. en. In: *Energies* 13.15 (Jan. 2020). Number: 15 Publisher: Multidisciplinary Digital Publishing Institute,

[145] Zhe Wang et al. Predicting City-Scale Daily Electricity Consumption Using Data-Driven Models. en. In: *Advances in Applied Energy* (Apr. 2021), ISSN: 2666-7924.

[146] Winston Chang et al. shiny: Web Application Framework for R. Jan. 2021.

[147] INSPIRE Data Specification on Buildings – Technical Guidelines. https://inspire.ec.europa.eu/id/document/tg/bu.

[148] Cartografía catastral. http://www.catastro.minhap.es/webinspire/index.html.

[149] Instituto Nacional de Estadística - Estadística experimental. https://www.ine.es/en/experimental/atlas/experimental_atlas_en.htm.

[150] DATADIS. La plataforma de datos de consumo eléctrico. https://datadis.es.

[151] https://bit.ly/3jTwT29.

[152] Codigos Postales de España. https://www.codigospostales.com.

[153] Cartografía secciones censales y callejero de Censo Electoral. https://www.ine.es/prodyser/callejero/.

[154] The official home of the Python Programming Language. https://www.python.org/. en.

[155] Welcome to the QGIS project. https://www.qgis.org/en/site/.

[156] The R Project for Statistical Computing. https://www.r-project.org/.

[157] The most popular database for modern apps. https://www.mongodb.com. en-us.

[158] Jelle Goeman et al. penalized: L1 (Lasso and Fused Lasso) and L2 (Ridge) Penalized Estimation in GLMs and in the Cox Model. July 2018.