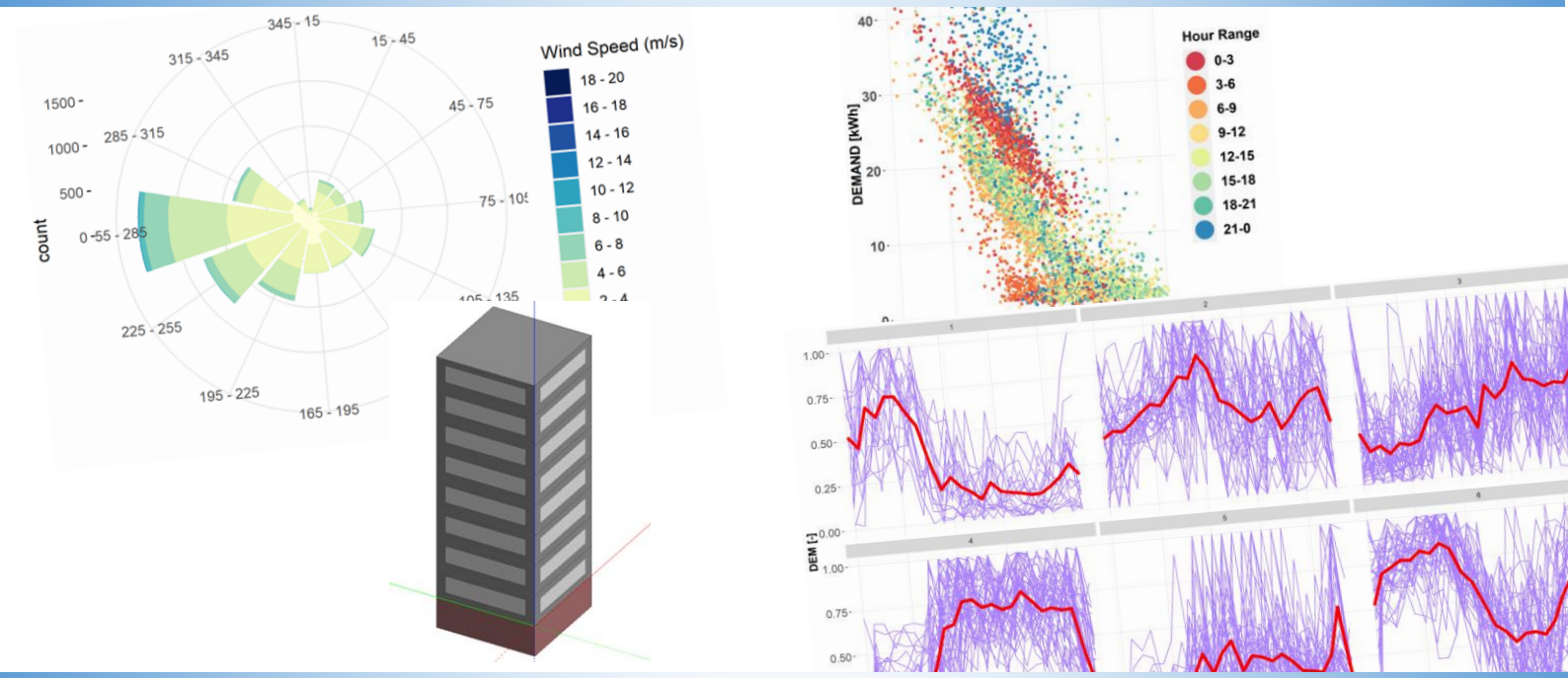
Analysis on the Application of Machine-Learning Algorithms for District-Heating Networks' Characterization & Management



Mikel Lumbreras Mugaguren

Directores: Koldobika Martín Escudero y Gonzalo Diarce Belloso

Universidad del País Vasco Euskal Herriko Unibertsitatea

2023

Analysis on the Application of Machine-Learning Algorithms for District-Heating Networks' Characterization & Management

Mikel Lumbreras Mugaguren

Dissertation presented at the University of the Basque Country (UPV/EHU) in fulfilment of the requirements for the degree of PhD of Energy Efficiency and Sustainability in Engineering and Architecture

Under the supervision of:

Dr. Koldobika Martin Escudero

Dr. Gonzalo Diarce Belloso

Bilbao, May 2023

ABSTRACT

This PhD thesis deals with the feasibility of the application of machine-learning algorithms for energy characterization in district-heating networks. In particular, the dissertation will be focused on four main applications:

- Energy demand outlier identification and removal.
- Recognition of main energy demand patterns in buildings connected to the network.
- Study of interpretability/classification of the energy patterns.
- Forecasting of the energy demand in daily and hourly resolution.

The interest of the thesis was awoken by the current energy situation in the European Union, where buildings are responsible for more than the 40% of the total energy demand. District-Heating networks, and specifically, modern networks (the so-called fourth and fifth generation district-heating network) have been identified as efficient systems for supplying energy from production plants to the final consumers/buildings. Moreover, due to the grouping of multiple buildings that is allowed in these systems, district heating networks enable the development and implementation of unique algorithms for energy management in the overall system.

New directives from the European Commission obliges to remotely read and save the data from the consumption points, opening a new opportunity for large scale algorithm based on big-data structures. In this context, artificial intelligence and in particular, machine learning algorithms are positioned as a great alternative to energy characterization against the traditionally used energy simulation models. The advantages of artificial intelligence models against traditional methods are the time and cost efficiency, flexibility for training and testing models and the lack of necessity of information about the buildings. It is necessary to mention that although this Thesis is focused on buildings that are connected to DH networks, the scope of these models is applicable to the characterization of any heat-load in buildings.

The development of this thesis faces the steps given for the analysis of data coming from real buildings located in Tartu (Estonia) and connected to the district-heating network of the city and it comprises all the steps for the analysis of the data and development of several machine-learning models. The main body of the dissertation finishes in Chapter IX. This final section evaluates the efficiency of the models developed with data from Tartu´s buildings but applied to a simulated case-study in Bilbao, Spain.

Regarding the results obtained, it can be concluded that this PhD Thesis validates the possibility of using machine-learning algorithms in the context of energy characterization in building and district scale. Therefore, the new method developed for the prediction of demand, combining several machine-learning techniques, overperform the rest of the models and is capable to work efficiently for a wide variety of buildings and locations. Additionally, when this model is applied for energy management of a network, relevant economic savings are obtained, reaching a 10% of savings in the simulated case analyzed in the last chapter of the dissertation.

Finally, even though the scope of the dissertation is limited to these ten chapters, in the close future, two different research lines have been identified. On the one hand, these studies are planned to be extended also to the application of Deep-Learning (Neural Networks) algorithms for similar purposes, so that the efficiency of these models could be sized. On the other hand, Industrialization of the models developed in the dissertation is proposed. This type of works that are limited to laboratory research require a high effort investment and more research to be applied in real applications.

RESUMEN

Esta tesis doctoral estudia la viabilidad de la aplicación de algoritmos de aprendizaje automático para la caracterización energética de los edificios en entornos de redes de calefacción urbana (o en inglés, district-heating). En particular, la disertación se centrará en el análisis de las siguientes cuatro aplicaciones principales:

- La identificación y eliminación de valores atípicos de demanda en los edificios.
- Reconocimiento de los principales patrones de demanda energética en edificios conectados a la red.
- Estudio de interpretabilidad/clasificación de dichos patrones energéticos. Análisis descriptivo de los patrones de la demanda.
- Predicción de la demanda de energía en resolución diaria y horaria.

El interés de la tesis fue despertado por la situación energética actual en la Unión Europea, donde los edificios son responsables de más del 40% del consumo total de energía. Las redes de distrito, y en concreto, las redes modernas han sido identificadas como sistemas eficientes para el suministro de energía desde las plantas de producción hasta los consumidores finales/edificios debido a su economía de escala. Además, debido a la agrupación de edificios en una misma red, permitirán el desarrollo e implementación de algoritmos para la gestión de la energía en el sistema completo.

Las nuevas directivas de la Comisión Europea obligan a monitorizar y enviar de forma remota los datos de los puntos de consumo, abriendo una nueva oportunidad de implementar algoritmos a gran escala basado en estructuras de big-data. En este contexto, la inteligencia artificial y en particular los algoritmos de aprendizaje automático se posicionan como una gran alternativa para la caracterización energética frente a los modelos de simulación energética tradicionalmente utilizados. Los modelos basados en datos se han aplicado desde la década de 1980 para la caracterización de datos de energía de baja frecuencia. Sin embargo, la aplicación de algoritmos de aprendizaje automático abrió nuevas líneas de investigación porque muestran varias

ventajas respecto a los métodos más tradicionales como son, la eficiencia de tiempo y coste computacional, flexibilidad para entrenar y validar diversos modelos y la no necesidad de información acerca del edificio. Es necesario mencionar que, si bien esta Tesis está enfocada a edificios que están conectados a redes de DH, el alcance de estos modelos es aplicable a la caracterización de cualquier carga térmica en edificios.

El desarrollo de esta tesis se enfrenta a los pasos dados para el análisis de datos procedentes de edificios reales ubicados en Tartu (Estonia) y conectados a la red de calefacción urbana de la ciudad. La parte principal de la Tesis finaliza con el Capítulo IX. Esta sección evalúa la eficiencia de los modelos desarrollados con datos de los edificios de Tartu, pero aplicados a un caso de estudio simulado en Bilbao, España.

En cuanto a los resultados obtenidos, se puede concluir que esta Tesis Doctoral valida la posibilidad de utilizar algoritmos de aprendizaje automático en el contexto de la caracterización energética a escala de edificios y barrios. Por lo tanto, el nuevo método desarrollado para la predicción de la demanda, que combina varias técnicas de aprendizaje automático, supera al resto de los modelos y es capaz de funcionar de manera eficiente para una amplia variedad de edificios y ubicaciones. Adicionalmente, cuando se aplica este modelo para la gestión energética de una red, se obtienen importantes ahorros económicos, llegando a un 10% de ahorro en el caso simulado analizado en el último capítulo de la tesis.

Finalmente, aunque el alcance de la tesis se limita a estos diez capítulos, en un futuro próximo se han identificado dos líneas de investigación diferentes. Por un lado, se prevé que estos estudios se extiendan también a la aplicación de algoritmos de Deep-Learning (Neural Networks) para fines similares, de manera que se pueda dimensionar la eficiencia de estos modelos. Por otro lado, se propone la industrialización de los modelos desarrollados en la tesis. Este tipo de trabajos que se limitan a la investigación de laboratorio requieren una inversión de alto esfuerzo y más investigación para ser aplicada en aplicaciones reales.

LABURPENA

Doktorego-tesi honek hiri-berokuntzako sareen inguruneetan eraikinen karakterizazio energetikorako (edo ingelesez, district-heating) ikasketa automatikoko algoritmoak aplikatzearen bideragarritasuna aztertzen du. Zehazki, honako lau aplikazio nagusi hauen analisian oinarrituko da hitzaldia:

- Eraikinetako eskariaren balio atipikoak identifikatzea eta ezabatzea.
- Sarera konektatutako eraikinetako energia-eskariaren eredu nagusiak ezagutzea.
- Patroi energetiko horien interpretagarritasuna/sailkapena aztertzea. Eskariaren patroien analisi deskribatzailea.
- Energia-eskaria eguneroko eta orduko bereizmenean iragartzea.

Europar Batasuneko egungo egoera energetikoak piztu zuen tesiaren interesa, eraikinak baitira guztizko energia-kontsumoaren % 40 baino gehiagoren erantzuleak. Barruti-sareak, eta, zehazki, sare modernoak, energia-hornidurarako sistema eraginkor gisa identifikatu dira, ekoizpen-instalazioetatik azken kontsumitzaileetara/eraikinetara, eskala-ekonomia dela-eta. Gainera, eraikinak sare berean biltzen direnez, sistema osoan energia kudeatzeko algoritmoak garatzea eta ezartzea ahalbidetuko dute.

Europako Batzordearen zuzentarau berriek kontsumo-puntuetako datuak urrutitik monitorizatzera eta bidaltzera behartzen dute, big-data egituretan oinarritutako eskala handiko algoritmoak ezartzeko aukera berri bat irekiz. Testuinguru horretan, adimen artifiziala eta, bereziki, ikasketa automatikoko algoritmoak aukera handia dira energia-karakterizaziorako, tradizionalki erabili izan diren simulazio energetikoko ereduen aldean. Datuetan oinarritutako ereduak 1980ko hamarkadatik aplikatu dira behe-maiztasuneko energia-datuen karakterizaziorako. Hala ere, ikasketa automatikoko algoritmoen aplikazioak ikerketa-lerro berriak ireki zituen, metodo tradizionalenen aldean hainbat abantaila erakusten dituztelako, hala nola denboraren eraginkortasuna eta kostu konputazionala, entrenatzeko eta hainbat eredu baliozkotzeko malgutasuna

eta eraikinari buruzko informaziorik behar ez izatea. Aipatu beharra dago tesi hau DH-ko sareetara konektatuta dauden eraikinetara bideratuta dagoen arren, eredu horien irismena eraikinetako edozein karga termikoren karakterizazioari aplika dakiokeela.

Tesi honen garapenak aurre egin behar die Tartun (Estonia) dauden eta hiriko berokuntza-sareari lotuta dauden benetako eraikinetatik datozen datuak aztertzeko emandako pausoei. Tesiaren zati nagusia IX. kapituluarekin amaitzen da. Atal honek Tarturen eraikinen datuekin garatutako ereduen eraginkortasuna ebaluatzen du, baina Bilboko (Espainia) azterketa simulatuko kasu bati aplikatuta.

Lortutako emaitzei dagokienez, ondoriozta daiteke doktore-tesi honek baliozkotu egiten duela ikaskuntza automatikoko algoritmoak erabiltzeko aukera, eraikinen eta auzoen eskalako karakterizazio energetikoaren testuinguruan. Beraz, eskaria iragartzeko garatutako metodo berriak, ikasketa automatikoko hainbat teknika konbinatzen dituenak, gainerako ereduak gainditzen ditu eta eraikin eta kokapen mota askotarako modu eraginkorrean funtzionatzeko gai da. Gainera, sare baten energia-kudeaketarako eredu hori aplikatzen denean, aurrezpen ekonomiko handiak lortzen dira, eta % 10eko aurrezkia lortzen da tesiaren azken kapituluan aztertutako kasu simulatuan.

Azkenik, tesiaren irismena hamar kapitulu horietara mugatzen bada ere, etorkizun hurbilean bi ikerketa-ildo desberdin identifikatu dira. Alde batetik, azterlan horiek antzeko helburuetarako Deep-Learning (Neural Networks) algoritmoak aplikatzera ere hedatzea aurreikusten da, eredu horien eraginkortasuna dimentsionatu ahal izateko. Bestalde, tesian garatutako ereduen industrializazioa proposatzen da. Laborategiko ikerketara mugatzen diren lan horiek ahalegin handiko inbertsioa eta ikerketa gehiago eskatzen dute benetako aplikazioetan aplikatzeko.

AGRADECIMIENTOS-AKNOWLEDGEMENTS

El camino hasta este punto no ha sido fácil, pero… ¿quién dijo que fuera fácil?

Desde el momento en el que decidí entrar a la Escuela de Ingeniería de Bilbao supe que, o lo daba todo o no conseguiría los resultados que he conseguido. Desde entonces, la mayor virtud que he desarrollado es la disciplina y es lo que me ha llevado hasta aquí. Esta tesis doctoral es el culmen a más de diez años (sí, ya son diez… como pasa el tiempo) de estudios universitarios y constante aprendizaje en distintos ámbitos y retos laborales que se me han ido planteando.

Aunque resulte prácticamente imposible resumirlo todo en esta sección, me gustaría utilizar estas líneas para acordarme de todas las personas que han estado cerca durante todo este proceso y ofrecerles un breve agradecimiento que no hace justicia al verdadero apoyo que han sido para mí.

En primer lugar, me gustaría agradecer a todas las personas que han estado conmigo en el ámbito personal. Primeramente, a mi aita y a mi ama, por “aguantarme” durante todos estos años y, es que, como ya he comentado no ha sido un proceso sencillo (ni corto). Siempre han estado ahí, aun cuando la alta carga de trabajo ha hecho que a veces esté un poco más inaguantables. Mención especial también a mi Amama y a mi tío Jesus, que no han podido ver cómo acababa todo esto. Cómo no agradecer a mis amigos más cercanos, que de una forma u otra me han hecho este camino mucho más fácil, pudiendo desconectar con esas salidas en bici y otros mil viajes que hemos hecho durante estos años. Y es que, tan importante es saber desconectar en ciertos momentos para poder avanzar más fuerte.

Volviendo al ámbito más cercano a la tesis, el primer agradecimiento tiene que ir los directores de tesis: Koldo y Gonzalo. Por aportar soluciones y propuestas de gran valor a mi trabajo y, sobre todo, por guiarme y centrar mi gran cantidad de ideas que en un principio se encontraban ordenadas de forma caótica en mi cabeza. También me

gustaría acordarme de mis (ex)compañeros de trabajo en mi época de investigador en la universidad, donde realmente comenzó a gestarse esta disertación.

Por otro lado, y yendo al verdadero origen de esta disertación, agradecer a Roberto por su gran influencia en mí en los primeros años de andadura en el mundo laboral y por contagiarme esa actitud investigadora. Realmente fue quién me transmitió esa pasión por los datos y todo el potencial que estos albergan. Acordarme también al resto de compañeros de TECNALIA que trabajaron conmigo cuando yo no era más que un pez en el océano. Agradecer también a los trabajadores de GREN Eesti (Estonia) por facilitarme los tan valiosos datos de demanda de sus edificios que finalmente han derivado en una tesis doctoral. También me gustaría dar las gracias a mi actual empresa, MAINSTRAT, y al gran equipo humano que está detrás de esta. Y es que, siempre han apoyado cualquier iniciativa que fuera en pro de esta tesis doctoral.

A todos ellos va dedicado este trabajo.

TABLE OF CONTENTS

Abbreviations & Nomenclature

Abbreviations

4GDH	Fourth Generation District-Heating
5GDH	Fifth Generation District-Heating
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ARM	Association Rules Mining
CART	Classification & Regression Trees
CHP	Combined Heat & Power
CVI	Clustering Validation Index
DBSCAN	Density Based Clustering
DH	District-Heating
DHW	Domestic Hot Water
DL	Deep-Learning
DS	DataSet
DT	Decision-Trees
DTW	Dynamic Time Warping
EC	European Commission
EU	European Union
GIS	Georeferenced Information System
HDD	Heating Degree Days
HOB	Heating Only Boilers

HP	Heat Pump
HVAC	Heating, Ventilation & Air-Conditioning
IEA	International Energy Agency
K-NN	K-Nearest Neighbors
MAPE	Mean Absolute Percentage Error
MEA	Mean Absolute Error
ML	Machine-Learning
MVLR	Multi-Variable Linear Regression
NB	Naïve Bayes
NLP	Natural Language Processing
O&M	Operation and Maintenance
RES	Renewable Energy Sources
RF	Random Forest
RMSE	Root Mean Square Error
SARIMA	Seasonal Autoregressive Integrated Moving Average
SH	Space-Heating
SOM	Self-Organizing Maps
ST	Solar Thermal
SVM	Support Vector Machine
SVR	Support Vector Regressor
ULT	Ultra-Low Temperature
XGB	Extreme Gradient Boosting

Nomenclature

A	Surface	[m^2]
α	Regression Coefficient	[-]
β	Regression Coefficient for outdoor temperature	[1/K]
cp	Cost complexity parameter	[-]
C_P	Specific heat	[KJ/Kg·K]
d	distance	[m]
D	Diameter	[m]
Eps	Radius in DBSCAN	[-]
f	Fuel consumption per day	[KJ/day]
$F_{thermal}$	Partial load efficiency reduction factor	[%]
G_T	Global Solar Irradiance	[W/m^2]
IQR	Interquartile	[var]
J	Cost function	[var]
K	Number of clusters	[-]
L	Length	[m]
L(x,y)	Least square´s function	[var]
m	Volumetric Flow	[m^3/s]
MAPE	Mean Absolute Percentage Error	[%]
MinPts	Number of observations inside a cluster	[-]
MinSplit	Minimum number of observations in a node	[-]
n	number of observations	[-]
η	Efficiency	[%]

P	Probability	[-]
PF	Probability Factor	[%]
ρ	Fluid Density	[kg/m^3]
q	Quartile	[-]
q_t	Actual hourly heat load	[kWh]
q_{norm_t}	Normalized hourly load	[kWh]
$q_{\max t}$	Daily minimum heat load	[kWh]
$q_{min\,t}$	Daily maximum heat load	[kWh]
$\bar{q}$	Daily mean heat load	[kW]
Q	Energy Load	[kW]
Q_{REF}	Reference Heat Load in Q-T algorithm	[kWh]
R^2	Coefficient of Determination	[-]
RMSE	Root Mean Square Error	[-]
sd	Standard Deviation	[var]
T	Temperature	[°C]
T_{OUT}	Outdoor Temperature	[°C]
U	Thermal transmittance	[W/m^2·K]
μ	Cluster center & mean value	[-]
W_D	Wind Direction	[-]
W_S	Wind Speed	[m/s]
w	weight	[-]
X	Independent Variable	[var]
Y	Dependent Variable	[var]

YEC	Yearly Energy Consumption Deviation	[%]

Figure List

Table List

Chapter I

Preamble & Structure

Abstract

This first chapter is used for introducing the framework of the thesis and for analyzing which are the main objectives of the dissertation. It will show which were the motivations behind all the studies covered along the document and how this work matches the general research lines of the research group. Additionally, this chapter could be used as the guide for the rest of the chapters since the general structure of the chapters will be covered.

Resumen

Este primer capítulo se utiliza para introducir el marco de la tesis y para analizar cuáles son los objetivos principales de esta disertación. Se mostrarán cuáles fueron las motivaciones detrás de todos los estudios cubiertos a lo largo del documento y cómo este trabajo se ajusta a las líneas de investigación generales del grupo de investigación ENEDI. Además, este capítulo podrá usarse como guía para el resto de los capítulos, ya que se cubrirá la estructura general de los diferentes capítulos que forman el conjunto de la disertación.

Chapter I Preamble & Structure

This first chapter of the thesis will serve as a guide or reference for all the readers of this document, since the objective of this section is to explain how this document is ordered and to set the basis of some of the concepts that will be developed during all the dissertation. All in all, this chapter presents the following sections:

- **Problem statement** that motivates the development of this thesis.
- **Main Structure** of the document. It will present the main body of the document and how the chapters are presented.
- **Framework of the thesis**.

1. Problem Statement

This thesis responds to the necessity of accurate characterization of building energy load since accurate prediction of energy load is one of the effective means to reduce building energy consumption, since it helps achieve better control of power system and improve energy utilization. Accurately characterizing the energy load of buildings can provide benchmarks for energy management of building systems and show the energy-saving potential of buildings.

Traditional heat-load meters allowed to read energy reading with daily or lower frequency. However, the current smart energy meters for heat-load measuring enable to remotely read this variable with hourly and even higher frequency. This high accuracy readings open a new opportunity for accurate data-driven models and in addition to the increasingly use of artificial intelligence in different fields, a novel research area is studied combining energy characterization and machine-learning models.

We will mainly focus on characterizing heat-load energy in buildings that are connected to a DH network, but the models that are developed and analyzed could be oriented for their application in any building.

The document will present different research topics on how machine-learning algorithms can be applied for the characterization of energy demand in buildings. Therefore, we consider relevant for the comprehension of the studies to set the definitions for the following concepts: Energy (or Heating) Demand, Energy (or heat) Load and Energy Consumption in buildings.

- **Heating Load:** It refers to the amount of energy a space with occupancy and plug loads needs to receive to maintain its temperature at the required level (no air changes or re-circulation or anything else; mostly energy transfer except for infiltration).
- **Heating Demand:** is the amount of energy an HVAC system will provide to condition that space or to condition outdoor air before supplying it to the space.
- **Heating Consumption:** is the amount of energy that will be consumed by that HVAC system to satisfy the heating demand.

In our cases, the data from the buildings was received in hourly basis and consequently the heating-load (kW) and heating demand (kWh) presents the same value, even though the theoretical content behind these two concepts are the same. This is why, these two variables are used indistinctly along the document. As we do not have enough information about the efficiency of the HVAC systems inside the buildings, we are not going to use consumption variable, although this value would be completely dependent of the other two.

2. Main Structure

The chapters in the dissertation are structured as follows.

The dissertation can be divided into four main parts. The first part embraces the first three chapters (Chapter II and Chapter III) and includes the introduction and the general state of the art of the dissertation. The second and main part of the document embraces from Chapter IV to Chapter VIII and groups the studied carried out using real data from the DH in Tartu. The following chapter (Chapter IX) transfers the knowledge obtained

with real data of Tartu to a simulated case in Bilbao and the final part consists of the main conclusions of the dissertation (Chapter X).

The following figure (Fig. I-1) summarizes the structure of the document and will serve as a reference point to be consulted during the reading of the document.

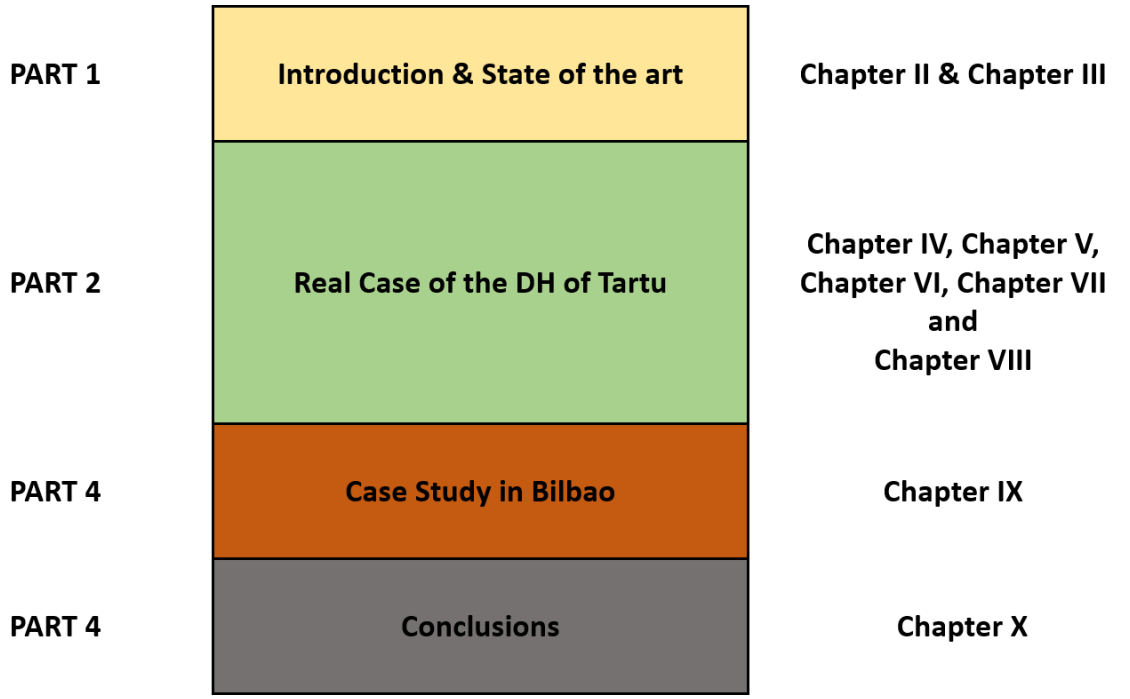


Fig. I-1. General Structure of the document

The first chapter (Chapter II) aimed to introduce the main topics covered within the thesis. In this chapter, the author has tried to introduce the technology of DH networks and the evolution of the technology suffered from the beginnings until today. Moreover, it gives an overview of what artificial intelligence is and specially explains machine-learning concept. This first chapter has tried to set the basics for the rest of the chapters. Whereas this section (Chapter I) introduce the context of the thesis, in the following chapter (Chapter II), an in-depth analysis of the literature available in the topics covered by the thesis is presented and it includes the most important gaps that this thesis is trying to fulfill.

Moreover, in a second section including Chapter IV and Chapter V, an introduction to the data used within this work is presented. In particular, Chapter IV aims to explain the

different data sources used and gives the context of the case-study. This chapter shows the different buildings analyzed and gives an overview of the heat demand profiles of these buildings. Furthermore, Chapter V summarizes the data pre-processing techniques used and presents a general analysis of the main variables, including the self-called Q-T algorithm for demand characterization based on basic linear regression As previously mentioned, Q-T algorithm was developed also by the author based on the nature of the data and using multi-variable linear regressions. It brings forward the work developed in collaboration with R. Garay-Martinez, B. Arregi and people from GREN Eesti [1] and published in ENERGY Journal [2]. A complete version of this article can be found In Appendix section (Chapter XI).

In a third section of the thesis, the different steps that embrace the method for energy management in a district are covered. Thus, Chapter VI, Chapter VII and Chapter VIII aim to present the main tasks of the method and present the different algorithms used and evaluated for each of the steps. First, Chapter VI explore the use of different clustering techniques for the identification of heating patterns. This chapter shows all the analysis done for resulting in the article presented in collaboration with R. Garay-Martinez and B. Arregi in JOURNAL OF BUILDING ENGINEERING [3]. Also, the complete version of this article is attached to the Appendix section.

Then, Chapter VII will include the modelling of the typical heating energy daily profiles by means of the supervised classification models and finally Chapter VIII show the energy forecasting models developed. As a result of this whole section, the paper is expected to be published in ENERGY Journal. A complete version of this article can be found In Appendix section.

Finally, a summary of the main conclusions of the dissertation are shown in Chapter X. In this final chapter, a brief overview of the achievements made is presented and it is followed by the contributions and future research objectives.

3. Framework of the thesis

The research developed in this thesis can be considered as part of the research lines carried out by ENEDI Group. Enedi Group (or Grupo de Investigación ENergética en la EDIficación de la UPV/EHU) [4] is a research group that was established in 2005 as a result of the Agreement signed between the Basque Government Department of Housing and the UPV/EHU by virtue of which this Research Group is in charge of managing and developing the Thermal Area of the Building Quality Control Laboratory (LCCE) of the Basque Government.

The next figure (Fig. I-2) shows the three research lines of the group: EnediPHYS (building physics), EnediSYST (installations) and EnediTES (thermal energy storage) [5]. Furthermore, these research lines, depending on the study to be carried out, can be classified by the scale (material/component, thermal zone, building and district/city) or by the type of study (characterization of properties, monitoring and data processing, simulation and numerical modelling). This classification of the research lines is shown in Fig. I-2.

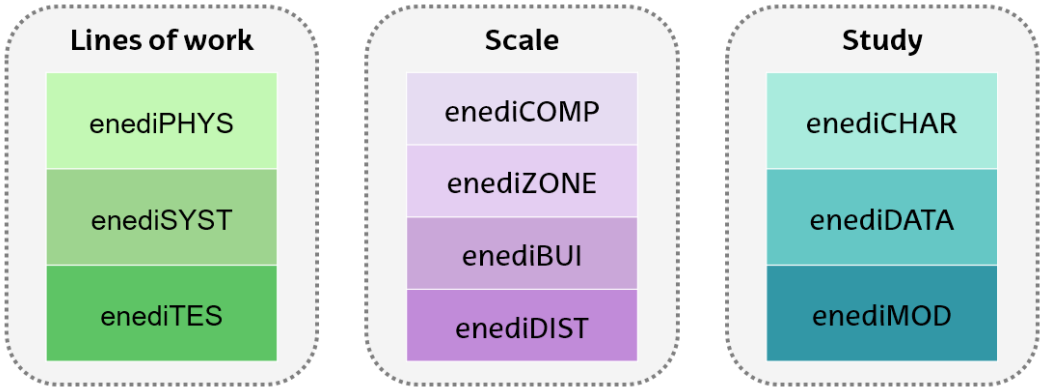


Fig. I-2. Research lines of ENEDI group

This thesis presents a building´ energetic characterization work using monitored data. The monitoring of the data is not done by the author since the real buildings are located in Tartu (Estonia). The work is focused on the characterization of the demand in building only using real data, using different machine-learning models. Therefore, this thesis is integrated in the **EnediPHYS** line, varying between building and district scale

(**enediBUILD** and **enediDIST**) and in **enediDATA** field of study. The research activities within EnediPHYS are the following:

- Thermal characterization of materials: optimization of existing products and development of innovative products.
- Steady-state and dynamic thermal characterization of constructive solutions both under controlled conditions and under real outdoor conditions.
- **Thermal characterization of dwellings by means of in situ measurements and parametric modeling.**
- **Development of thermal models using parameter identification techniques.**
- Modeling the transport and storage of moisture in walls
- Influence of moisture in buildings
- Hygrothermal properties of building materials
- Behavior of wet granular media

Besides, the first steps of this thesis were developed within RELaTED project [6]. RELaTED (or REnewable Low TEmperature District) provided an innovative concept of decentralized Ultra-Low Temperature (ULT) network solution that can pave the way for expanding and modernizing existing DH networks as well as introducing and establishing district heating in emerging EU markets. This project, funded under the European Union's Horizon 2020 research and innovation programme, has developed a robust ultra-low temperature concept, which allow for the incorporation of low-grade heat sources with minimal constraints. Also, ULT DH reduces operational costs due to fewer heat losses, better energy performance of heat generation plants and extensive use of de-carbonized energy sources at low marginal costs. The RELaTED ULT DH concept has been demonstrated in four complementary operation environments (new and existing DH, locations, climatic conditions, dimension…) in Denmark, Estonia, Serbia and Spain.

Therefore, all the data used in this dissertation was provided by the DH operator in Tartu (Estonia). The DH operator in this area is GREN Eesti [1] and they provide 1-hour frequency heating demand data (among other variables) from more than 40 buildings

connected to their DH network. The different types of provided data and their monitoring will be explained and discussed in Chapter IV. At this point, I would like to thank GREN Eesti (again) for providing the data from the substations for academic purposes.

Chapter II

Introduction

Abstract

This first chapter introduces the main concepts covered along the dissertation and explains the motivations behind the development of this PhD Thesis. It will give an overview of the current situation of energy consumption in buildings and how district-heating networks could improve the efficiency of energy production and supply. This chapter will also outline the "trendy" concept, artificial intelligence, and will explain how this technology could be applied for the energy management purposes.

Resumen

Este primer capítulo introduce los principales conceptos tratados a lo largo de la tesis y explica cuáles son las motivaciones detrás de iniciar esta Tesis doctoral. El capítulo ofrecerá una visión general de la situación actual del consumo de energía en los edificios y cómo las redes de calefacción urbana podrían mejorar la eficiencia de la producción y el suministro de esta energía. Este capítulo también describirá el concepto ampliamente utilizado hoy en día, inteligencia artificial, y cómo esta tecnología podría aplicarse realmente para fines de gestión de la energía.

Chapter II Introduction

Energy management is becoming more and more important in today´s scenario and as well as it gains importance, it also becomes a more complex task. Energy consumption is suffering a transition to electrification of the consumption and the very used fossil fuel are being replaced by renewable energy sources or RES. During the last two decades primary energy has grown by 49% and CO_2 emissions by 43%, with an average annual increase of 2% and 1.8%, respectively [7].

In addition, the COVID-19 pandemic suffered in 2020 and the current instability caused by geopolitical issues around the world increases even more the energetic instability. On the one hand, the COVID-19 pandemic decreased (and practically stopped) the global activity with an associated energy consumption reduction, reducing the price of the energy sources due to a lower demand. On the other hand, the geopolitical problems reduce the availability of the energy sources (specially, natural gas and other fossil fuels) imported to the European Union (EU). Thus, from these events on, energy management have become even a more important issue for all the countries in the EU and worldwide. The increasingly evident global warming is accelerating the transition from traditional fossil fuels to RES and consequently, the direction of the energy scenarios of tomorrow are very difficult to predict.

The unique conclusion that is valid to all the phenomena surrounding energy management is that for maintaining a sustainable energy situation it is necessary to reduce energy consumption as much as possible in a wide range of contexts and apply energy efficiency measures to optimize all the energy transformations. Every energy unit must be optimized for its final consumption. Buildings account for more than 40% of the total energy consumption [8], so it becomes a very relevant part of the global energy scenario. This thesis will be focused on the efficiency of energy management in buildings.

1. Energy Demand in Buildings

Buildings are responsible for a large share of the total energy demand worldwide. The latest references on this topic estimate the total energy demand in buildings to reach 40% of the total energy demand in the EU [8], whereas in developed countries this energy share is reduced to a percentage between 20 to 40%. Moreover, according to International Energy Agency (IEA), buildings are responsible for producing more than 30% of CO_2 emissions [9]. Thus, and answering to the big part of the "problem" that are the buildings´ energy demand, the European Commission (EC) is focusing on increasing energy efficiency in buildings by means of directives [10] and [11].

The energy demand in buildings includes energy consumption for heating purposes (the so-called heating load), for cooling purposes (the so-called cooling load) and the electric consumption for all the devices in the building. This dissertation will be focused only on the heating load, which is as well divided into the next two demands: (i) Demand for Space-Heating (SH) and (ii) Demand for Domestic Hot Water (DHW). SH demand comes from the necessity to maintain a certain thermal comfort inside the building and heating the air in the dwelling. On the other hand, DHW demand covers the energy used for heating the water for different applications and devices. The heat load for space heating (SH) is highly correlated with external climate, but relevant transitory effects are generated with building usage and scheduling of Heating, Ventilation & Air-Conditioning (HVAC) systems. Furthermore, the demand for DHW load is principally correlated with the building usage (i.e., scheduling of showers) and the energy demand patterns of the occupants. Energy demand patterns are daily loads or a fraction of the daily demand profile that are repeated over time.

According to the IEA, the energy share of the components that form the total energy demand in buildings vary depending on the final use or building type [9]. The following table (Table II-1) shows the distribution of the energy components for residential buildings and offices in different countries.

Table II-1. Energy demand by end uses in residential and office buildings.

Residential [%]	Spain	USA	UK	Offices [%]	Spain	USA	UK
SH	42	53	62	SH	52	48	55
DHW	26	14	22	DHW	0	10	4
Others[1]	32	18	17	Others	48	42	41

Moreover, the share of the different types of energy demands is not the unique point in which the buildings with different uses differ. The peak demand, the valley demand and the general shape of the energy profiles are different in a residential building and an education building, for example. Demand profiles are also different in buildings with the same final use depending on the user's behavior inside the building. A residential building occupied by people that work will not present the same energy demand profile than the same residential building occupied by retired people. This does not only refer to the total energy consumed but also to the profile of the heat demand and their corresponding energy demand schedule. In any case, when the energy consumption of multiple buildings is grouped and ruled by a unique energy management cluster, individual demand patterns are masked and synergies between the different energy demands can be obtained. The following section introduces the concept of DH networks, which allow to group several buildings in one or few distribution lines, so that could be supplied by a low number of energy production plants.

2. District-Heating Networks

District-heating (DH) networks are usually centralized thermal systems that distribute heat (and cold in some cases) from a production level to different customers, enabling the connection of multiple buildings in the same energy distribution grid. Owing to the high efficiency levels of these networks, DH networks will play a very relevant role in the task of increasing energy efficiency by the optimization of every energy unit produced

[1] Others include lighting and appliances and other cooling loads.

[12]. DH comprises a network of pipes connecting buildings in a neighborhood or a whole city, so that they can be served from centralized plants or a number of distributed heat production plants. These networks enable the connection of multiple heat sources as production units [13]. This technology is currently responsible for covering around 13% of the total thermal energy demand in the European Union and this value is supposed to increase in the following years [14].

In order to understand the context of this dissertation and the activities carried out for this purpose, the following paragraphs will cover a wide range of topics around DH networks and will set the basis for all the analysis and research done in this dissertation. First, Section 2.1 will present a brief summary of the different DH generations and their main characteristics. Then, Section 2.2 shows how the energy management works in these systems, followed by the cost-distribution schemes in Section 2.3.

2.1. Evolution of District-Heating Networks

In the easiest DH scheme configuration (Fig. II-1) the energy is produced (or transformed) in a large production plant (Combined Heat & Power or CHP as shown in Fig. II-1) and distributed throughout the transmission lines up to the DH substation. In order to balance the system against intra/inter-day oscillations, usually there are intermediate thermal storage installations through the transmission line. Then, the substation transforms the energy from the transmission line of the network to the end users in the grid. Thus, substations are responsible for bringing the energy from the primary side of the network (from production to substation) and secondary side of the network (from substation to the building). Fig. II-1 presents a schematical illustration of how a DH networks is distributed.

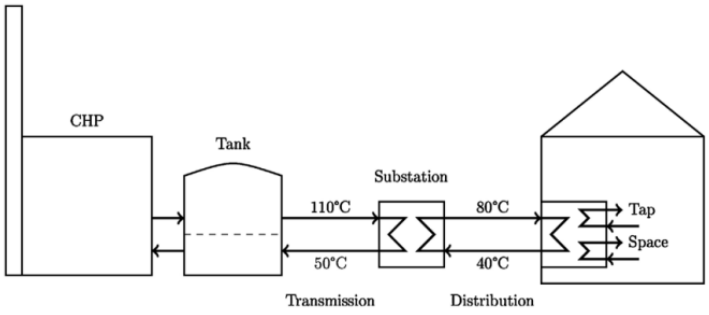


Fig. II-1. Basic scheme of DH Network with indicative temperatures.

Regardless of the DH generation, the efficiency of DH networks mainly depends on the three following factors:

- ***Energy Source***: It is the first step in every DH network and the inlet energy to the network. It could be based on common fossil fuels or alternative sources, such as solar thermal heat or waste heat stream. In this step, the first efficiency metric is the one that evaluated the energy transformation from the heating plant to the network and transferred the energy to the heat carrier fluid of the network.
- ***Energy Conversion Efficiency***: Specially used for the energy conversion and exchange in the heat storage tanks and the DH substations. In these devices, part of the energy will be lost due to the heat exchange efficiency.
- ***System Configuration***: DH networks usually comprises very long pipelines, so that the configuration of all the devices is crucial for reducing the energy loss due to the distribution heat losses.

Even though literature defines five DH generation, this study will classify these networks into high-temperature DH and low-temperature DH. High temperature networks comprise first, second and third generation networks, whereas low temperature networks comprise fourth and fifth generation. The characteristics of each generation are shown in the following lines.

2.1.1. High-Temperature Networks: First, second and third DH Network Generations

The first DH networks started to commercialize in Lockport (USA) and New York (USA) cities in 1880s and 1890s, respectively. The so-called first DH generation were placed

between 1880s and 1930s and their main characteristic is that water stream was used as fluid carrier in the system. The use of water steam (and not water) was motivated due to the lack of electric motors for supplying liquid water. The extreme conditions for supplying steam (high temperatures and high pressure) required large pipelines and the heat losses in the line were very relevant. Thus, in second generation DH generation, the steam was substituted by pressurized water, reducing the supply temperature up to 120°C approximately. However, the insulation of the elements (distribution lines and heat exchangers) was still inexistent and that was the main reason for the development of the third-generation networks in 1980s. These new generation networks include pre-insulated pipelines and start using heat meters for monitoring the energy flows in the grid. In this third-generation network started the development of the data-driven model for energy management. This topic will be deeply covered in next chapter (Chapter III) Thus, Table II-2 presents a brief summary of the most relevant characteristics of these networks.

Table II-2. Main characteristics of first, second and third generation DH networks.

DH Generation	1St Generation	2nd Generation	3rd Generation
Years	1880-1930	1930-1980	1980-2020
Main energy production Plants	Coal	Coal boilers and CHP (Coal)	CHP, Biomass, large ST, etc.
Supply Temperatures	300°C (20 bar)	110-120°C	80-100°C
Heat Carrier	Steam	Pressurized Water	Water

2.1.2. Low-Temperature Networks: Fourth Generation and Next Generations´ DH networks

It is still unclear which is the difference between the fourth and fifth-generation DH networks and depending on the reference, different definitions and boundary conditions are defined for each of the type. In some references, such as [13] or [15],

there is no difference between these two generations since the so-called fourth generation was supposed to cover DH networks up to 2050. However, in other references ([16] or [17]), certain differences are emphasized to distinguish these two generations. This dissertation is focused on developing data-driven models for different purposes, so that the models developed and explained within this work could be applied in both of them.

Regarding the fourth generation DH (or 4GDH hereinafter). This concept is an evolution of the third-generation networks and was motivated by an increased focus on energy efficiency, smart integrated energy systems, and the utilization of locally available RES, such as solar facades or waste heat from factories. Even though it is not part of this dissertation´ main objective, the author studied the connection of façade integration solar thermal (ST) energy to these types of networks and the paper resulting from that study is attached in Chapter XI [18].

One of the most relevant features of these networks is that the supply temperature in the transmission line was reduced to temperature levels as close to the actual temperature demand. Therefore, the maximum supply temperature in these systems round the 60-70°C or lower. The lower supply temperature lowers DH grid losses and enables the economically feasible integration of even more waste heat sources than in third generations', such as excess heat from data centers and supermarkets ([19],[20]). The temperature levels of 4GDH are normally sufficiently high to cover SH demands directly without using devices for temperature boosting through, e.g., heat pumps at the end-users [21].

Therefore, the 4GDH networks may fulfill the following five abilities [17]:

(i) The ability to supply existing, renovated, and new buildings with low-temperature DH for space heating and domestic hot water.

(ii) The ability to distribute heat in DH networks with low grid losses.

(iii) Reuse heat from low-temperature waste sources and integrate RES, such as ST and geothermal heat.

(iv) Be an integrated part of smart energy systems and thereby helping to solve the task of integrating fluctuating renewable energy sources and proving energy conservation into the smart energy system.

(v) Ensure suitable planning, cost and incentive structures in relation to the operation as well as to strategic investments related to the transformation into future sustainable energy systems.

Regarding the fifth-generation DH networks (or 5GDH hereinafter), the concept was firstly introduced in 2015 by the FLEXYNETS project from the H2020 program [22]. These networks are described as networks operating at near-ground temperatures (20-45°C) using a bidirectional exchange of heat and cold between connected buildings, facilitated by seasonal storage. 5GDH requires heat pumps at the connected buildings in order to reach the proper temperature for domestic hot water for avoiding the risk of Legionella. Thus, this type of networks includes the possibility to supply district-cooling at the same time due to the low temperature of the double-loop network. Similar to the analysis for 4GDH and according to [17], the main vectors of this concept are the following:

(i) Take advantage of the synergies of combining heating and cooling in areas of mixed purpose buildings.

(ii) Minimize the barrier of utilizing local waste heat sources and minimize upfront investment cost for the utility company, though the required initial investment at the end-users will be higher.

(iii) Enable less restrictive organic growth of the system, as central heat supply is not as critical since new additional end-users will both add and use heat from the network.

To sum up with the analysis of both modern DH network types, the main difference between 4GDH and 5GDH consists in the ability to supply both heat and cooling load in the same grid, making rather difficult for 5GDH to implement smart energy management system. Besides, this dissertation is only focused on heating loads and consequently, all the works presented along this document are more appropriate to be implemented in

4GDH. Thus, from here on, we will focus only on the analysis of fourth generation DH networks.

2.2. Energy Management of a District-Heating Network

DH network management is not an easy issue. The demand in a district will vary throughout time and the network will have to adapt and control the energy provided at every moment. Basically, there are two ways to control the energy input.

- **Varying the flow/supply temperature** with constant flow rate in the transmission lines.
- **Varying the flow rate** of the heat transfer fluid and with a constant flow/supply temperature.

Usually heat production in real DH networks is only based on the temperature prediction for the following hours. It also depends on the heat production system and the inertia and flexibility of each generation plant to increase and decrease the instant heat production. For example, the ease to increase the demand in a medium-size gas boiler is not the same than the one that needs a large CHP system in which the turbine requires some time to reach a steady and secure status. It is neither the same for intermittent heat production systems such as ST production. Another variable to be considered in DH energy management is the size of the network. Thus, the energy generated in the network requires some time to reach all the buildings connected to the heating grid. A large distance from the production point(s) to the buildings increases the heat losses in the distribution pipeline and increases the time required for the hot water to reach the buildings (substations).

To sum up, the management of a real network includes the analysis of many variables in the system and depends on the specific network to be managed. Before starting with the application, the following section will summarize the main components of the cost structure of DH systems.

2.3. Cost-Distribution Models in District-Heating Networks

The main cost in a DH network is energy production cost. Fixed costs and variable costs are the two main components a DH plant incurs when producing heat and are also the primary inputs for the two main DH pricing structures used for DHs cost-plus pricing model and the marginal-cost pricing model - underscoring their importance to a DH cost analysis [23].

- ***Cost-plus Pricing Model***: It is used in regulated DH markets. The main reason is that DH companies are not allowed to adjust their heating prices below the market price. As a result, DH companies must rely on premiums for their profits, the amount of which is determined by the costs incurred by DH companies.
- ***Marginal-Cost Pricing Model***: It is alternatively used in deregulated DH markets. Essentially, DH companies in a deregulated market compete by pricing the production of heat less than the equilibrium market price, therefore increasing the DH company's market share and profit. The price that is set under the market price is known as the marginal cost, which in the case of DH systems, is the cost of one more unit of heat through DH.

Additionally, this chapter seeks to provide a general overview of the main components for each DH heat production plant, highlighting any special considerations that could affect the total cost related with each facility. Therefore, the objectives will be to address the following for each DH plant:

A. Identify fixed costs (FC) attributed to heat generation.
B. Identify variable costs (VC) attributed to heat generation.
C. Discuss any special features that should be considered regarding the cost of DH plant Most of the cost data for the subchapters.

Based on desk research, the catalogue of common DH technologies provides the most comprehensive overview of the costs corresponding with each technology that is used in low temperature (fourth and fifth generation) DH networks. Furthermore, the

catalogue of technologies is updated on a regular, which ensures that the cost data used in this report is the most up to date. Lastly, the catalogue is meant for international and Danish audiences, and the data provided gives a generalized analysis of energy systems for economic scenario models; in this sense, the information can be seen as a baseline or standard representation of the energy systems and economic indicators.

In this context, it is important to define what cost data was included by the Danish Energy Agency for investment costs, fixed costs, and variable costs.

- Investment costs: Investment costs include engineering, procurement, construction, infrastructure, and connection costs [24].
- Fixed Operation and Maintenance (O&M) costs: Includes all costs independent of how many hours the plant is operated i.e., administration, operational staff, payments for O&M service agreements, network or system charges, property tax, insurance, and reinvestments to extend the lifetime of the plants [24].
- Variable O&M costs: Include consumption of auxiliary materials (water, lubricants, fuel additives), treatment and disposal of residuals, spare parts and output related repair and maintenance (however not costs covered by guarantees and insurances). **Fuel costs are not included** [24].

Investment costs are presented with fixed O&M costs because the assets procured from the initial investment – heat production unit(s), infrastructure, etc. – depreciate over time. Thus, companies will create a depreciation expense schedule for asset investments with values falling over time. The depreciation costs therefore fall under a fixed O&M cost. The goal is to have a general overview of the costs and factors to consider for making the business case for a mixture of different DH plants. The hope for this report is to provide a starting point for policy makers to assess the viability of different DH systems in their city or region. It is important to remark that the production costs identified must be considered as a general reference, as each plant, DH network and country has its own characteristics and, therefore, specific production costs. For

example, electricity tariffs and the cost for energy may differ at any given time from country to country.

Therefore, Table II-3 presents a brief summary of the approximate costs of different heating plants: CHP, Heating Only Boilers (HOB), ST systems and large heat pumps (HP).

Table II-3. Brief Summary of generation costs in DH networks.

	CHP	HOB	ST Systems	Large HP
Generation Capacity	1-500 MW	<100 MW	<100MW	<20MW
Investment Costs	0.59-3.3 M€/MWe	0.59 M€/MW	150-500 €/m^2	M€/MW
Fixed O&M	10^4-10^5 €/MW_e·Year	7-40E4 €/MW·Year	Very Low	20000 €/MW·Year
Variable O&M	5 €/MWh	1.1-2.7€/MWh	€/MWh	4-6€/MWh

3. Artificial Intelligence: Algorithms used in this dissertation.

Artificial Intelligence or AI is a very wide and deep concept at the same time, as the following definition by IBM states [25]:

"It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable."

We will try to write the definition of the AI with our own words:

"Artificial Intelligence is the science of developing computer programs that try to reach their goals by thinking the same way the humans do."

Several applications for this science are currently under development:

- Natural Language Processing commonly referred as NLP.
- Artificial Vision and other robotic stuffs
- Any kind of predictions: from a disease prediction to any numeric variable
- And a large etc.

It is not part of the objectives of this work to deeply introduce to AI. Conversely, we found interesting to introduce the basics of this science and clarify some aspects of its terminology, which is still confusing for the wide public.

AI, in turn, includes Machine-Learning (ML) and Deep Learning (DL) algorithms, among other methods. Even though deep learning and machine learning tend to be used interchangeably, it is worth noting the nuances between the two. Both are subfields of artificial intelligence, but deep learning is actually a subfield of machine learning. The hierarchies of these concepts can be observed in Fig. II-2.

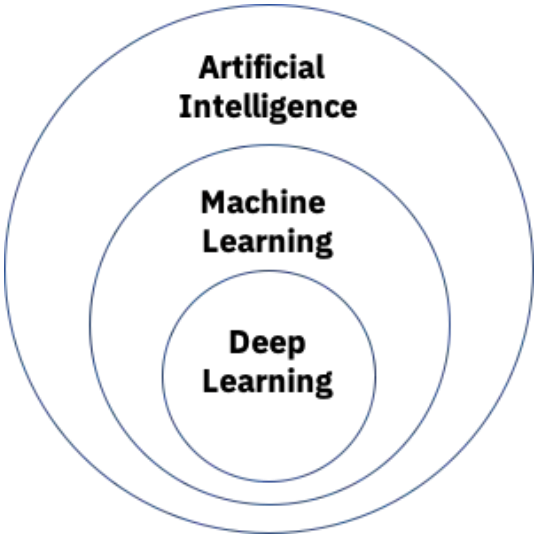


Fig. II-2. Hierarchy definition in AI, Machine-learning and Deep-Learning.

The main difference between machine learning and deep learning is the way the algorithms learn from data. In general, ML requires more human intervention than DL, since the algorithms that are developed depend exclusively on the data model that is obtained and require an understanding of the data and objectives. On the other hand,

in the concept of deep learning (mainly artificial neural networks), no human processing (at least, not at the level of machine learning) of the data is required. In deep learning, it is only required to design the form/structure (input layers, output layers, etc.) of the model without data preprocessing. The model itself is the one that conforms to the specific case study. Therefore, ML models offer greater flexibility when designing AI algorithms, while DL allows us to use more data and a greater scope. Thus, among the black-box models inside AI, ML are more transparent and easier to tune, and they are more suitable for energy topics.

ML or automatic learning is a branch of artificial intelligence that was born in the 1950s. Within ML algorithms there is an extensive classification, where **supervised learning** and **unsupervised learning** algorithms stand out. There are other algorithms that would fall somewhere between these two branches: semi-supervised algorithms, reinforcement learning algorithms, etc. The main characteristics of these algorithms are shown in Table II-4.

Table II-4. Main characteristics of ML algorithm types

Supervised Learning	Unsupervised Learning	Semi-supervised Learning	Reinforcement Learning
Data Scientist provide input, output, and feedback to build model	Use deep learning to arrive at conclusions and patterns through unlabeled training data	Builds a model through a mix of labeled and unlabeled data, a set of categories suggestions and exampled labels	Self-interpreting but based on a system of rewards and punishments learned through trial and error, seeking maximum reward
Linear regressions, support vector machines, decision-trees, etc.	Principal Component Analysis, clustering, a priori etc.	Self-training, Label propagation, etc.	Q-learning, model-based value estimation, etc.
Used for Classification and regression problems	Used for labelling data	Idem than supervised	Application-based: estimate parameters, reduce consumption, etc.

Furthermore, Deep-learning, semi-supervised learning and reinforcement learning are discarded in this dissertation due to the nature of these algorithms. Energy, as a concept, is a measurable variable and its value is closely related to external variables, such as climatic and calendar variables. Moreover, the correlation between these variables is commonly known by the developers. For instance, it is known that when the outdoor temperature is very low, the heat demand for SH will increase and when the sun is shining, this energy demand will decrease. Therefore, it is not essential to tune deep-learning models to understand the correlation between the variables because they are (roughly) known before stating the problem. The applicability of semi-supervised learning is limited to the cases in which only some observations are labelled. This is not the case for this dissertation, as it will be presented in Chapter IV where the data used in this study is presented. In the case of reinforcement learning, it is more application-oriented and the approach to these models is different. In this work we are not trying to make a machine model oriented to any application, so that the objective of this dissertation does not require the use of reinforcement learning.

Therefore, this chapter is focused on introducing the supervised and unsupervised learning concepts. To do so, we will list the main algorithms that can be found in recent literature. The insights of all the algorithms will be provided along the dissertation in the corresponding chapter (from Chapter IV to Chapter VIII). The Fig. II-3 illustrates the main differences among these two types of ML algorithms.

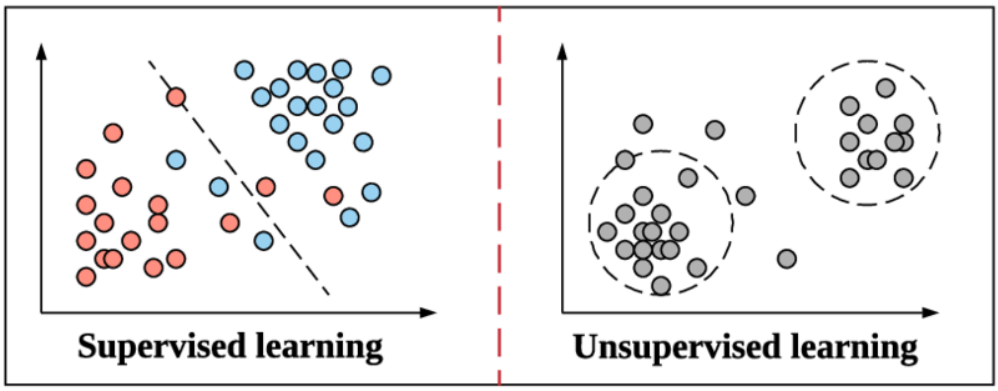


Fig. II-3. Supervised and unsupervised learning concepts

Whereas supervised learning models are built under labelled data (left side in Fig. II-3) for different purposes, unsupervised models are used to discover hidden patterns in unlabeled data. In other words, in the case of supervised learning algorithms, the attributes and classes of the data are used to train a model and gain insight from the data. However, in unsupervised learning, only the attributes are available and there is no knowledge about the class/label of the data, so the objective of this second type of algorithm is to provide knowledge about the data structure.

As a consequence of the differences, each type of algorithm will be used for different purposes. On the one hand, supervised learning uses already labelled data to train a model and can be used for both, classification, and regression problems. The main difference between these two applications is that classification is used for categorical variables (for example, day of the week) and the regression problems concern only numeric variables (energy demand, for example). On the other hand, unsupervised learning is used for dimensionality reduction, association problems and labelling the data.

The supervised algorithms used in this work are the following:

- Linear Regression: Explained in Chapter V.
- Decision-Trees: Explained in Chapter V.
- Support Vector Machines: Explained in Chapter VIII.

- Random Forest: Explained in Chapter VIII.
- Extreme Gradient Boosting: Explained in Chapter VIII.

The unsupervised algorithms covered in this work:

- K-means: Explained in Chapter VI.
- Density Based Clustering: Explained in Chapter VI.
- Dynamic Time-Warping: Explained in Chapter VI.
- Fuzzy C-means Clustering: Explained in Chapter VI.

4. Using Data & Machine-Learning in District-Heating Networks

One of the main characteristics from the third generation DH network on was the use of monitored data to manage energy, as it was presented in Table II-2. This trend increased in fourth and fifth generation DH networks. The most usually used devices for the monitoring of the thermal energy in buildings are heat meters, which are usually located near the end-user (building). They allow the thermal energy demand of each consumer from the heat network to be measured ([26] & [27]). Modern devices allow the hourly or sub-hourly gathering of energy and additional operational variables, including continuous communication with the DH utility. These devices are being widely implemented across the EU, mandated by Directive 2018/2002 [11], which deals with the disaggregation of the final energy use by customers and the obligation to implement remote reading functionalities. Therefore, all meters will be remotely readable by January 2027.

The remote access of such data leads to different energy management systems of heat production in DH networks, such as [28] and [29], based on frequent readings of smart heat meters at consumer level. These systems usually perform short-term forecasting in the range of some hours or days. Another application for demand monitoring applied in DH environment is the urban planning and the design of new and optimized DH networks. Similar than in reference [18], the author of this dissertation with other co-authors, we developed a novel method for the design of a new network using demand

data and LiDAR data with Georeferenced Information System or GIS [30]. The cover of the article resulting from that study is also attached in Chapter XI.

To sum up, modern networks and the use of devices for the remote monitoring enable the application of data driven models, such as artificial intelligence and, specially, machine-learning models. This field, even tough is not new, has recently gained significant relevance due to the use of large datasets and digitalization of data storage systems and can noticeably contribute to energy efficiency in buildings. The state-of-the-art on this topic is presented in Chapter III, which shows the most important works on the topic and identifies the literature gaps covered by this dissertation.

Chapter III

State of the Art

Abstract

This chapter presents the analysis of the state of the art. This literature review will cover from the first data-based models to the most advanced and modern machine learning models used in the framework of the Thesis. A large number of references to other works carried out in the field of electricity demand and how these models could (or not) be adapted to the thermal demand in buildings are presented. The sections that cover the studies on the thermal demand are divided into models oriented to the identification of patterns and to the prediction of the same. The chapter ends with the gaps identified and covered by the development of this Thesis.

Resumen

En este capítulo se presenta el análisis del estado del arte. Esta revisión de la literatura cubrirá desde los primeros modelos basados en datos hasta los modelos de aprendizaje automático más avanzados y modernos utilizados en el marco de la Tesis. Se presenta un amplio número de referencias a otros trabajos realizados el ámbito de la demanda eléctrica y cómo estos modelos podrían (o no) adaptarse a la demanda térmica en los edificios. Las secciones que cubren los estudios sobre la demanda térmica se dividen en modelos orientados la identificación de patrones y a la predicción de la misma. El capítulo finaliza con los vacíos identificados y cubiertos por el desarrollo de esta Tesis.

Chapter III State of the Art

The analysis of energy using machine-learning models (ML models) and data-driven models in general is not new. Thus, the analysis of the most relevant references in the current literature will enable to understand which are the main gaps that this thesis covers and the advantages that the developed models bring compared with current status of the literature. The chapter starts with the presentation of the origin of data-driven models applied to different energetic applications and the rest of the chapter will bring a review through the most relevant references.

Note that the state of the art presented in this section of the dissertation can be considered to be a general analysis, while specific literature review is presented in each of the chapters of the document.

When talking about energy characterization and building modelling, it is important to define the three most relevant type of models, depending on the way they are built.

- **White-Box Models**: In these models one can clearly explain how they behave, since they are models whose inner logic, workings and programming steps are transparent and therefore its decision-making process is interpretable. Usually, these models are significantly easy to explain and interpret and in contrast, they present low accuracy and less predictive capacity than the other two types of models since they require to re-simulate the buildings in case something changed in the system. Anyway, there has been research on white-box model forecasting based on such tools as EnergyPlus [31] or TRNSYS [32], including their calibration against meter data. However, these methods are difficult to be implemented at district scale, as the DH utility does not usually have access to all the required information to develop such models (architectural data, use patterns, etc.) and the model development and calibration process is considered

to be time and resource intensive. Thus, this approach is not considered to be reasonable on a district or city scale.

- **Grey Box Models:** These models are formulated through differential equations that combine metered data with prior physical knowledge. Grey-box models integrate prior physical knowledge and are typically formulated as state-space models through a set of stochastic linear differential equations, either in discrete or continuous time. Grey box models require a deep understanding of all relevant phenomena in a building that impact instantaneous or cumulated values of the load. Due to the complexity of these models, many grey-box models in the literature have been formulated for individual components of the building, such as walls or windows [33]. Thus, it is challenging to fit suitable grey-box models for multi-element systems such as buildings (and even more complex for whole districts), because the interaction between the different elements and parameters is frequently unknown or too complex to be explicitly formulated. Some examples of grey-box models in buildings are provided next. Madsen et al. developed a model based on discrete-time building performance [34]. Andersen et al. described the time modelling of the heat dynamics using stochastic differential equations [35]. Similarly, Bacher et al. applied grey-box modelling for different applications regarding heat dynamics of a building, such as, control of indoor climate and energy demand forecasting [36].
- **Black-Box Models:** Finally, these models are purely based on data and can be trained to infer relations between inputs and outputs using statistical techniques with no physical interpretation. A wide variety of black box models are available in the literature, from the simplest energy signatures op to complex multi-step machine learning models using the most modern algorithms. The most important advantages that present these models is the low computation cost and the high predictive capacity.

This thesis will be focused only on the development of black-box models using different ML techniques. Consequently, the following paragraphs show the evolution of this type of models.

1. First Data-Driven Models for Energy Characterization: Energy Signatures.

The first data-driven methods for thermal energy characterization are found in the 1980s [37]. The first evolution from these models derived into the simplest **energy signatures**. Energy signature models are widely applied data-driven models that express the heating energy use as a function of weather variables. In a study presented by M. Fels in 1986, a data-driven methodology was developed for energy demand modelling using only the variable of heating degree days (HDD). Heating degree days or HDD are a measure of how cold the temperature was on a given day or during a period of days. This method used a unique variable as predictor to define the slope of the gas demand curve and the model was specially used for estimating the total potential savings in buildings. Fig. III-1 (accompanied by Eq. (2)) illustrates how the model estimated the gas demand in buildings, where the function was divided by a reference temperature.

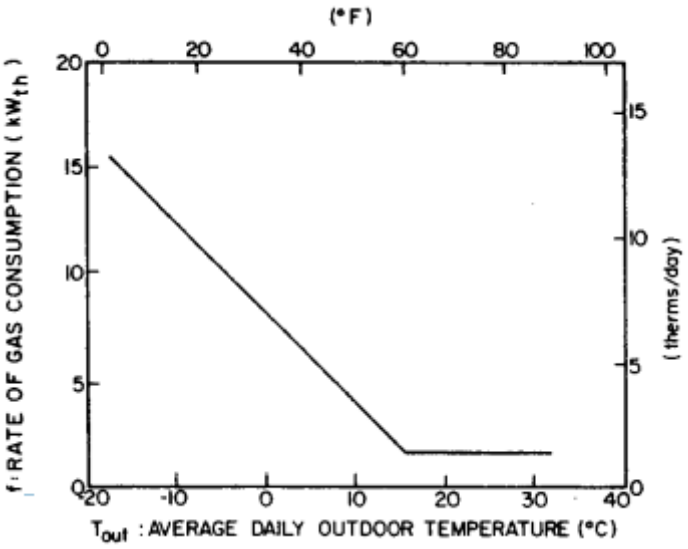


Fig. III-1. PRISM method to characterize gas demand in buildings. Source: [37]

$$f = \alpha + \beta \cdot (\tau - T_{OUT}) \qquad \text{Eq. (2)}$$

where; f is the fuel consumption per day, τ is the reference temperature and T_{OUT} is the mean daily temperature.

Some years later and with the objective of generalizing the previous PRISM method, the ASHRAE changepoint method was presented in 2002 by J. Kelly Kissock et al [38]. This report is formed by more than 180 pages explaining the method but, summarizing, this method generalized the PRISM method in order to characterize heating, cooling and a combination of both energy demand. Fig. III-2 illustrates the different types of regression models proposed in [38], from the two-parameters modelling to five-parameters models.

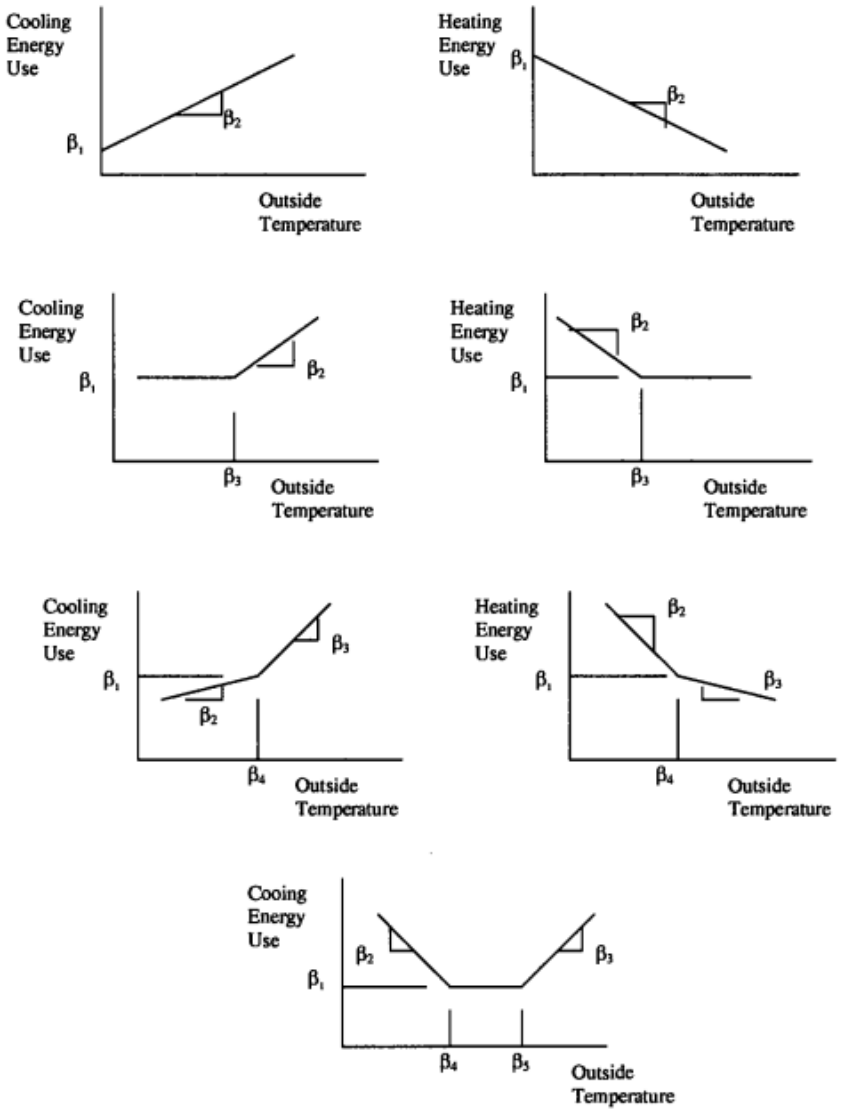


Fig. III-2. ASHRAE Changepoint models. Source: [38]

As an example of the regression proposed, Eq. 3 presents the five-parameter regression equation:

$$Y = \beta_1 + \beta_2 \cdot (X_1 - \beta_4)^- + \beta_3 \cdot (X_1 - \beta_5)^+ \qquad \text{Eq. (3)}$$

where β_1 is the constant term, β_2 is the left slope, β_3 is the right slope, β_4 is the left change point and β_5 is the right change point. As it was mentioned in [38], this five-parameter models are appropriate for modelling energy consumption data that includes both heating and cooling.

The last method that we find in these types of simple models is the Holt methods and its complement Holt-Winters method [39]. Other extensions/complements from the first methods can be also found, such as the four-parameter exponential smoothing developed by Ferbar et al. in 2016 [40]. These methods changed the view from the original energy signature models and started to focus on developing time-series modelling. It was found that these methods were appropriate for the long-term heat load forecasting because they were able to follow seasonal patterns that it was not possible until then.

Thus, energy signatures were the first data-driven models used for the characterization for both heating and cooling energy. Most of these studies concluded that the most dominant predictor within weather variables is outdoor temperature [41]. However, energy signature models have been demonstrated to be valid only for low resolution heat load predictions, such as weekly or monthly accumulated energy forecasting [42].

2. Artificial Intelligence for Electricity and Applicability to Heating Energy

Due to the greater number of electricity meters that have been installed over the years (compared to heat meters), most of the studies regarding energy characterization are applied to electric energy. Specially, we are referring to high frequency monitoring meters that enable the development of data-driven models for several applications in electricity. Therefore, it is important to summarize the most relevant references on electric demand characterization.

Data driven models, based on different machine learning methods focused on electricity demand, have been widely used in recent years (from [43] to [44]). Despite the different nature of thermal and electric demand, some of the methods developed for electric demand are interesting to be studied.

Regarding unsupervised ML techniques, Tureczek et al. studied the conclusions from more than 30 papers about the applicability of clustering techniques to electricity demand profiles in [43]. That study resulted in a large number of references on that

topic. The review applied 30 search phrases with relevance to smart meters, initially encompassing 71 papers containing potential studies regarding electricity demand classification using smart meter data. These 71 papers were thoroughly screened for purpose, data, method and results until a final list of 34 relevant papers concerning electricity demand classification using smart meter data. These studies are classified by the application and the method used: classification, forecasting, dimension reduction, etc.

Analyzing particular studies amongst the selected 34 papers, McLoughlin et al. presented a study about electricity use patterns within the residential sector in Ireland, based on different clustering processes [45]. This study characterized diurnal, intra-daily, seasonal and between customer electricity use. For this analysis, the paper investigates three of the most widely used unsupervised clustering methods: K-means, K-medoid and Self Organizing Maps (SOM) [46]. This last method (SOM) is included within deep-learning algorithm. Moreover, in a study carried out by Madeira Do Carmo et al., data from more than 4500 smart meters were used to conclude that individual electricity loads should be differentiated by use categories (residential, industrial, etc.), weekday and weekend, and summer and winter [47]. Lopez et al. in 2011 proposed the very used K-means algorithm for the segmentation of electricity demand [48]. Then, in another study carried out by Albert et al., Hidden Markov Model was proposed to infer occupancy states from demand time series data [49]. They shew that temporal patterns in the user's demand data can predict with good accuracy certain user characteristics. In another interesting article published by Ozawa et al. in 2016, they investigated the correlation between households´ lifestyle and the total electricity demand [50]. Two different methods were developed for determining the typical lifestyle in those buildings and, in both cases, it was concluded that most households consume less electricity when following a regular routine. Furthermore, the use Demand Side Response in smart grid context was discussed in [51]. They created simulated patterns of load curves, used these patterns to train and validate Artificial Neural Networks (or ANNs) and used this ANN to classify new data using these defined patterns.

There are also several works and references focused on forecasting electric demand in buildings. In a study carried out by Lindberg et al. developed a review analysis where most relevant trends in electric load forecasting were identified [52]. This work identified the most relevant works in this field, presenting the most updated methodologies and the challenges that these methodologies were facing. The most important challenge identified is the need for new approaches that can forecast long-term electric load with hourly or sub-hourly time resolution, with chronological hours (preferably over weeks or months). It is searched that these approaches could include the effects of energy-efficiency measures and new technologies for electric heating/cooling, EVs, local storage and flexibility resources. In another work in the same research line, Andersen et al. in 2013 concluded that climatic conditions highly affect final electricity demand in dwellings and applied this knowledge for the long-term forecasting of hourly electricity load [53].

Following with other methods, [54] and [55] developed different frameworks for prediction of peak electricity load and fault prediction, respectively. In the framework presented by Youngchan et al. in 2020, the daily peak electric demand is estimated combining a method to estimate peak temperatures and demand side management and they obtained accurate results [54]. Finally, the method used by Hu et al in [55] concluded that the combination of support vector machine and a model for signal decomposition achieved very accurate results for predicting the fault prediction.

From all the previous references studied regarding data driven models for electric demand, it is concluded that several applications using ML models have been already developed in this field: fault detection, households´ lifestyle identification or electricity demand forecasting. However, although the existing literature developed with data from electric loads [56] can be partially applicable to heat loads, it presents specificities through several effects [44], such as outdoor temperatures and activities taking place in the building. Electric and thermal load are different since the variables affecting the energy demand are different. Thus, some of the predictors used for electricity characterization are not applicable in thermal load and vice versa. In the end, the

methods used for building the ML models are different and consequently, this literature review cannot be directly used in our application. The following two sub-chapters will introduce to the most relevant studied in heating loads in buildings and some of them in DH context.

3. Review on Heating-Load Pattern Identification

The literature on heat load patterns is more limited than those on electric characterization. This is due to the access to high-quality and high-frequency data on heating loads to be pretty recent. Before starting the analysis of the state of the art on this topic, we find interesting to define what a heat demand pattern is. Energy demand patterns are daily loads or a fraction of the daily demand profile that are repeated over time [57]. These energy demand patterns may be caused by a repetitive demand action by the users of the buildings or the energy management strategies by the DH operator, and they may be repeated over different days within a heating season. A correct understanding of the energy demand patterns and the causes will help in the characterization process of the heating demand in the building [58]. Unsupervised learning algorithms have been successfully applied to identifying usage patterns commonly used in electricity load analysis ([59]–[62]); however, their use in heat-related applications has been limited so far.

Starting with the found references, a statistical approach to heating energy demand patterns of buildings connected to a DH was presented by Ma et al. (2014), based on such simplified variables as time and building types [63]. In the study, a Gaussian mixture model was presented for heat load prediction with an average absolute deviation of 4-8% (Mean Absolute Error or MAE). In the same year, a fault detection algorithm was proposed by Gadd et al. (2015) based on the identification of two heat-load patterns corresponding to DHW and SH demand [64]. This method clustered the customer profiles into different groups, extracting their representative patterns, and they detected unusual customers whose profiles significantly deviate from the rest of their groups.

Moreover, Gianniou et al. (2018) performed a clustering work over district heating data. It successfully identified a set of daily building heat load profiles, with specific patterns for weekdays and weekends [65]. The likeliness of pattern changes in a building based on calendar due to changes in the space heating baseload magnitude was set with a monthly resolution. Tureczek et al. (2019) and Calikus et al. (2019) included clustering methods to study the energy demand patterns in [66] and [67], respectively. Tureczek et al. (2019) demonstrated that unsupervised clustering can be applied to heat load data by analyzing data from 49 district heating substations and showing the autocorrelation existence between the clusters identified [66]. Moreover, decision-trees were used by Calikus et al. (2019) for mining the different demand patterns in a unique office building located in New York [67].

Johra et al. (2020) also performed clustering over district heating data, resulting in the profiling of 1665 households to 4 profiles [68]. This work was performed independently for all 4 seasons in a year, and the correlation between the clusters assigned for each of the 4 seasons was studied. In both cases, the data presented a quite stable baseload, mainly with one clear peak in the morning, which somehow limited the handling of more varied building usage. In addition, the clustering process was performed jointly for all the daily profiles in all the buildings, which hindered the possibility of adapting the cluster identification processes to the specificity of each building. What is more, the way to use the identified patterns in forecasting applications was not defined, which would anyhow be limited to the lack of any explicit relation to climate and calendar. Finally, Liu et al. (2021) presented an application for anomaly detection of building energy demand based on the knowledge obtained with unsupervised learning techniques [69]. The knowledge developed in these studies focused on electricity demand is partially applicable also to heat loads.

An overview of all these references and the main features of these works are provided in Table III-1.

Table III-1. Overview of the main features in references concerning heat-load patterns.

Location	Buildings	Data Frequency	Algorithm	Reference
Shandong Province, China	Multiple buildings	Hourly	Gaussian mixture model (GMM)	[63] (Ma et al. 2014)
Helsingborg & Ängelholm (Sweden)	82 & 53	Hourly	Statistical Analysis	[64] (Gadd et al. 2015)
Aarhus City, Denmark	8293 single family household	Hourly	K-means	[65] (Gianniou et al. 2018)
Aarhus City, Denmark	53 substations	Hourly (Only 1-month)	K-means	[66] (Tureczek et al. 2019)
Helsingborg and Ängelholm (Sweden)	2239 buildings	Hourly	K-shape	[67] (Calikus et al. 2019)
Denmark	1665 dwellings	Hourly	K-means	[68] (Johra et al. 2020)
Chongqing (China)	3 Office buildings	Hourly	DBSCAN + K-means	[69] (Liu et al. 2021)

Most of the studies are located in Denmark and similar countries due to the extended use of DH networks. Additional references on this topic can be found in [70], [71].

4. Review on Heating-Load Forecasting

The other main application of ML models is short and long-term forecasting of heat load in dwellings. Built on energy signatures, more advanced ML models have been developed for heat energy demand forecasting.

In contrast to the advanced situation of electricity demand analysis, forecasting methods applied to heat loads are relatively new, and this research field is yet to be

consolidated. To the best knowledge of the author, initial works in this field ([72] and [73]) were developed in the early 2000s. In reference [72], a simple model was developed for forecasting demand in a DH network using outdoor temperature and social behavior. Besides, [73] presented an energy signature model for modelling different variables for the operation of a DH network.

Starting with the more advanced ML models for heat load prediction, Grosswindhager et al. developed a model for online short forecasting of thermal loads in DH networks using Seasonal Autoregressive Integrated Moving Average (SARIMA) models [74]. SARIMA models are variations for linear regression, and they use both seasonal variables and the outputs obtained from timesteps before. In this study, a mean absolute percentage error (MAPE) of around 5% was obtained for one-day-ahead forecasts. In another work published by paper Paudel et al. in 2017, a ML model using support vector machine algorithm was developed [75]. The main objective of this model was to forecast energy in Low Energy Buildings (LEB). The aim of the analysis was to demonstrate that using only part of the training data the model was able to achieve better prediction results. The forecasting results achieved reached an R^2 value of 0.93 and RMSE value up to 7.1. Following with similar works in 2017, Dahl et al. presented a simple autoregressive forecast model with weather prediction input for DH network load prediction [76]. The prediction accuracy was also measured by MAPE metric, and they achieved MAPE values between 5.4% and 5.7%.

Although some of the following recent references include deep-learning models, we consider interesting to present them in this analysis of the state of the art. In a study by Sandberg et al. a forecasting model using a neural network but only for a commercial building connected to a DH was developed [77]. The obtained R^2 for this prediction was 0.968 for a whole year in hourly basis and a MAPE value of 3.2%. Following with deep learning models and neural networks, Lei et al. studied the efficiency of these models for energy forecasting of more than 100 civil public buildings [78]. These neural networks had a MAPE value between 4% and 5%. In general, deep learning has accurate results for the prediction of hourly heat load, however, the computational cost of this

models is very high in comparison with simpler ML models. Another approach was developed by [79] in which the heat load predictions were carried out for the whole DH network, and thus individual effects of each building are avoided. This study compared various ML models and concluded that the most accurate was a gaussian process regression with MAPE below 3% for the accumulated energy in one year for the whole DH network. Furthermore, Sauer et al. presented a forecasting optimizer using extreme gradient boosting (XGB) algorithm and applied to simulated heat and cool loads in buildings [80]. The prediction results obtained in the analysis achieve R^2 values above 0.90. In [81] several ML algorithms were developed using neural networks for load forecasting in several buildings and the MAPE varied from 28,81% down to 8,98% in the optimal variant of the model. To finish with the state of the art on heat load forecasting, Zhao et al. in 2022 developed an optimization of convolutional neural network for short-term forecasting in residential buildings and he MAPE was reduced by 12.33% compared with a traditional network [82]. The MAPE values in this work ranged between 0.6% and 1.1%.

This being so, several references are found in this context and specially in recent years. Similar than it has been presented for pattern recognition, an overview of all these references and the main features of these works are provided in Table III-2.

Table III-2. Overview of the main features in references concerning heat-load forecasting.

Location	Data Frequency	Algorithm	Error Metric	Reference
Tannheim (Austria)	Half-Hourly (30')	SARIMA	MAPE ~5% (24 hours ahead)	[74] (Grosswindhager et al. 2011)
France	LEB Standards (Not real Data)	Support Vector Machine (SVM)	R^2 = 0.93; RMSE = 7.1 (kWh)	[75] (Paudel et al. 2017)
Aarhus (Denamrk)	Hourly	Autoregressive forecast model (ARX)	MAPE between 5.4%-5.7%	[76] (Dahl et al 2017)
Sweden	Hourly	Autoregressive neural network	R^2 = 0.96 & MAPE = 3.2% (1 Commercial building)	[77] (Sandberg et al. 2021)
Dalian (China), Public buildings	Hourly	Neural Networks (Deep, back-propagation, etc.)	MAPE ~5%	[78] (Lei et al-2021)
Ljubljana (Slovenia)	Hourly	Gaussian process regression	MANE ~3%[2]	[79] (Potočnik et al. 2021)
Simulated Data	Not Defined	eXtreme Gradient Boosting (XGB)	R^2 around 0.98	[80] (Sauer et al. 2022)
Cottbus (Germany)	Hourly	Artificial neural networks	MAPE = [8,98%,28,81%]	[81] (Sakkas et al. 2022)
Xi'an, Shaanxi Province (China)	Hourly	Convolution Neural networks (CNN)	MAPE = [0.6%,1.1%]	[80] (Zhao et al. 2022)

5. Gaps Identified

Based on the literature review carried out in the previous paragraphs, the gaps identified and consequently the challenges faced in this thesis are listed below:

- **Wide range of applicability**: the multi variable models presented in this thesis aim to be valid for any type of building, regardless of the heating profile or final

[2] MANE = Mean Absolute Normalized Error

use, since the building stock connected to a DH network is usually made up of all kinds of building types. Thus, the models are also valid for any type of climatic conditions and location of the network.

- **High temporal resolution predictions**: the present model is applied to hourly and daily data, thus meeting the necessity of high temporal resolution models. Most of the models are focused on hourly resolution data. This resolution is considered as high-frequency data for heat-load due to the heating load patterns in buildings.
- **Low-Calculation/Computation Cost:** In line with the first point, the models are suitable for the characterization of entire district, and they have to be valid for hourly energy management. Thus, the simulations must be fast enough to be able to control energy production. The developed models are based on relatively simple equations so that they maintain certain relation with real effects on buildings.
- **High Simplicity and accuracy**: Of course, the obtained accuracy must be high. The objective is to achieve similar predictions´ accuracy than more complex deep-learning models.
- **Use of Heat demand Pattern as predictors**: No other studies found in references used heat demand patterns as input variables for the prediction models.
- **Based on Real Data:** As it will be presented in Chapter IV, all the simulations and models are trained and tested against real data. This type of data usually includes higher deviation, and it is usually more difficult to obtain high accuracy. However, the models are tested in real conditions.
- **Application to manage a simulated district:** All the calculations are based on data from real buildings and to conclude the thesis in Chapter IV, the models are scaled to manage the energy production of an entire district and transferred to other location´ buildings.

6. Objectives

The main goal of this thesis is to explore the usability of Machine-Learning (ML) algorithms for energy characterization in a building-scale and evaluate the efficiency of these black-box models for the energy management in a district-scale. The objective of the thesis is to characterize the demand (SH + DHW) in building connected to DH networks. The thesis is purely based on data and as every data-based project, it will include:

- Data access and pre-processing activities.
- Different ML models development and evaluation.
- Application of the optimal model for a simulation of a district-heating network energy management.

This goal is reached by the development of a multi-step method in which different algorithms and models are built and evaluated with the purpose of characterizing the thermal demand in buildings connected to DH networks.

Apart from the abovementioned main objective of the dissertation, several technical or secondary objectives are expected to be achieved:

- To give an overview of the current status of district-heating networks and focusing on modern grids: fourth and fifth generation. This study analyses the energy generation systems, distribution temperatures and other technical characteristics of heating grids.
- To analyze heat demand in buildings using real data. This analysis will try to discover the main demand patterns in buildings for different uses: residential buildings, educational buildings, commercial buildings and offices. It enables to understand the occupational behavior of the dwellings as well as the space-heating techniques.

- To develop ML models using unsupervised techniques for the evaluation of heat demand patterns in different buildings. The dissertation will try to obtain a method to evaluate the efficiency of these unsupervised techniques.
- To develop ML models using supervised techniques for different purposes: classification of typical heating profiles or hourly demand forecasting in building and district-scale, among others.
- To compare the efficiency of the developed models against other models for the same purposes that are nowadays used or that can be found in a literature review.
- To apply the forecasting model for the energy management of a simulated district-heating and compare the results against a baseline.

Chapter IV

Data Presentation & Tartu´s Case-Study

Abstract

After the analysis of the state of the art, this chapter starts with the introduction to the data that will be used throughout all the dissertation. This chapter introduces the data sources used and presents a general description of the district-heating network where all the buildings are connected. Two main data-sources will be used along the dissertation: (i) climatic data-source and (ii) demand data from the buildings. It will finally analyze the correlation between these two data-sources using correlation coefficients.

Resumen

Tras el análisis del estado del arte, este capítulo comienza con una introducción general de los datos que se utilizarán a lo largo de toda la tesis. Este capítulo analiza las fuentes de datos utilizadas y presenta una descripción general de la red de calefacción urbana donde están conectados todos los edificios. Se utilizarán únicamente dos fuentes de datos principales a lo largo de toda la tesis: (i) una fuente de datos climáticos y (ii) los datos de demanda de los edificios. Finalmente, se analizará la correlación entre estas dos fuentes de datos utilizando coeficientes de correlación típicamente utilizados en análisis estadísticos.

Chapter IV Data Presentation & Tartu´s Case-Study

1. Introduction & Objectives of this Chapter

Today, data is becoming more and more important in all kind of activities, and it is often said that the one who owns data also owns power. This dissertation concentrates on developing ML models that use real energy demand data so that all the dissertation is based on data. Before starting with ML aspects, we have considered necessary to present the nature of the data and how the used data sources are associated.

Therefore, the main objective of this chapter is presenting the characteristics of the DH network used as a case-study, in which the buildings are connected and where all the following sections of the dissertation are located.

We are presenting the two main data sources used as input for our models:

- **Climatic data, obtain from an available** weather station.
- **Heat Load data** from the buildings connected to the DH network.

Besides, this chapter aims to fulfill the following additional objectives:

- Analyze the characteristics of the DH Network of this Case-Study.
- Present the nature of the data, analyzing frequency, data-sources, data acquisition, etc.
- Describe the characteristics of the buildings under study.
- Analyze the potentially existing correlation between the heat-load and climatic variables.

2. Summary of the DH in Tartu (Estonia)

The DH adopted as a case-study is located in Tartu (Estonia). With over 90.000 inhabitants, Tartu is the second largest city of the country. Tartu is served by a district heating system privately owned and operated by GREN [1]. The most used technology for energy production is CHP, which yearly delivers around 500 GWh to over 1500 consumers/buildings in the city. From this energy, 94 % is obtained from biomass and peat. In Fig. IV-1, the lay-out of the most important heat production facilities is shown, accompanied by the fuel source used and the nominal capacity (in MW) of each plant.

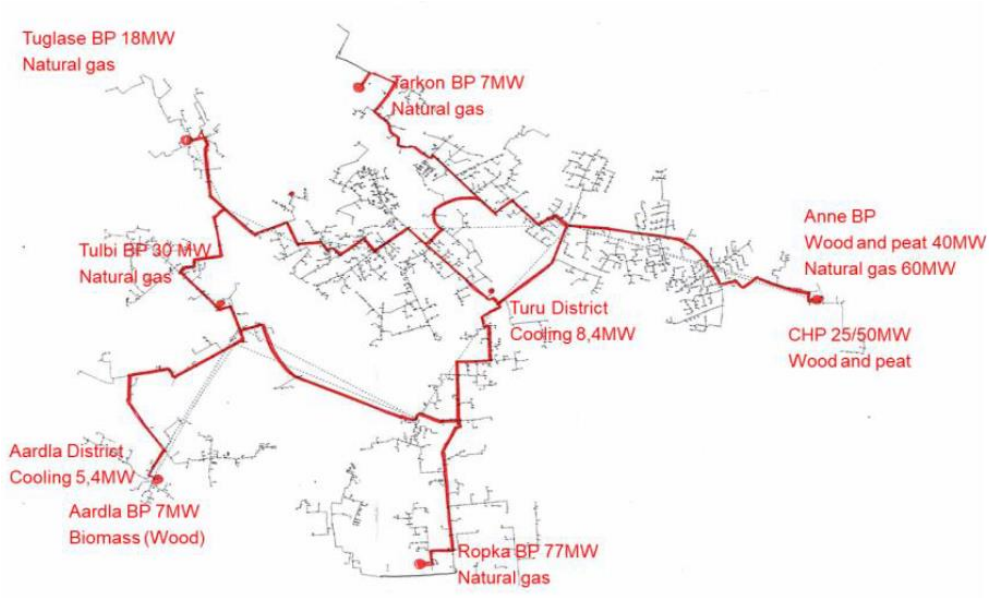


Fig. IV-1. DH Production-Scheme of Tartu´s Network

The main consumers of this network are collective housing (49%), industry and commercial buildings (33%) and individual housing (18%). The grid receives 40-60 new connections per year. While it is important to describe the entire network, this dissertation will focus on a specific part of the network (subnetwork) consisting of 43 buildings connected in the Tarkon-Tuglase district in the northern part of the city. Each building is identified by a code (or ID number), completely independent from its real address to avoid any identification problem and preserve the privacy of the users. A statistical analysis of the data of these buildings is presented in the following section,

All the data is collected for the complete year of 2019.

3. Data from Weather Station

Used climatic data was obtained from a weather station located and managed by the Physics Institute of the University of Tartu [83]. It collects data with a 15-min frequency and the station is located at a maximum distance of 5 kilometers from the buildings under study (the distance can be less, depending on the location of each building). Fig. IV-2 show the location of the weather station and the distance to the train station of Tarkon-Tuglase. The real address of the buildings is avoided for privacy concerns.

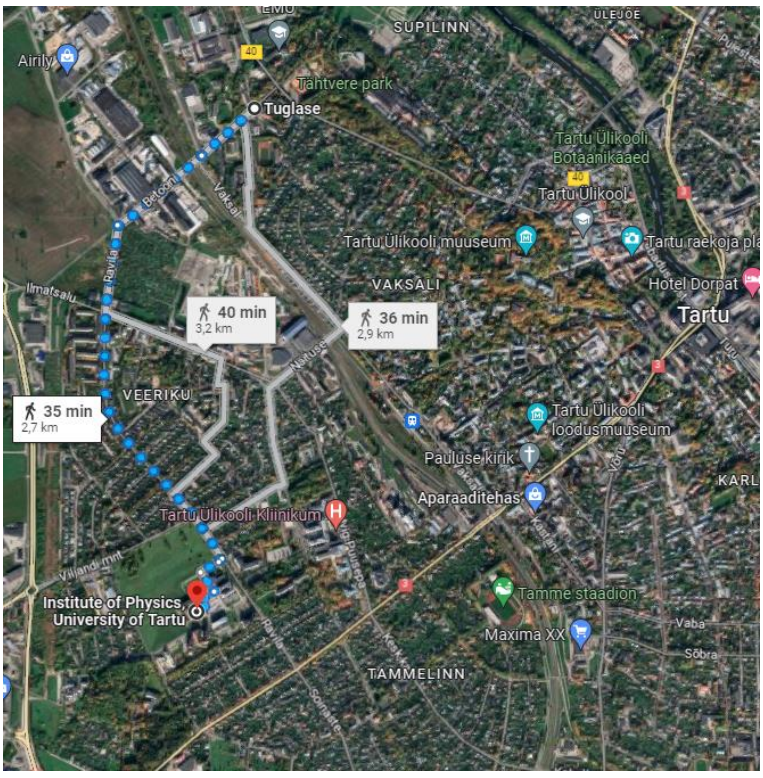


Fig. IV-2. Location of the Physics Institute of the University of Tartu and Tarkon-Tuglase. Source: Google Maps

For extracting the data, an online query provided by the Physics Institute of the University of Tartu is used. For extracting the data, we only required to introduce the starting and final date and the climatic variables that were considered to be relevant for the study.

According to the Köppen-Geiger classification [84], Tartu's climate is classified as D_{fb}, characterized by a very low outdoor temperature. Minimum outdoor temperature in winter can reach −20 °C, and all monthly averages fall below 20 °C. The climatic variables analyzed in this dissertation are the detailed next. Note that relative humidity has not been considered in the model, due to its little relevance in the absence of cooling demand.

- Outdoor temperature (T_{OUT}) in [°C].
- Global Solar Irradiance on a horizontal plane (G_T) in [W·h/m^2]
- Wind Speed (W_S) in [m/s]
- Wind Direction (W_D) (from 0 to 360°)

The pictures below show a statistical distribution of these four climatic variables under study. First, the outdoor temperature and solar irradiance are shown in Fig. IV-3 and Fig. IV-4, respectively.

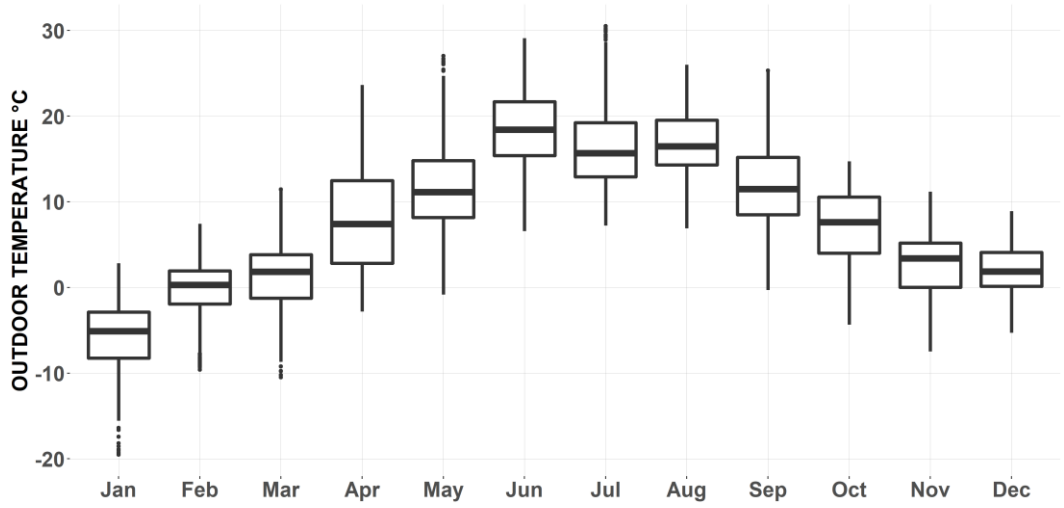


Fig. IV-3. Yearly outdoor temperature or T_{OUT} in °C in Tartu (Estonia). Data for 2019.

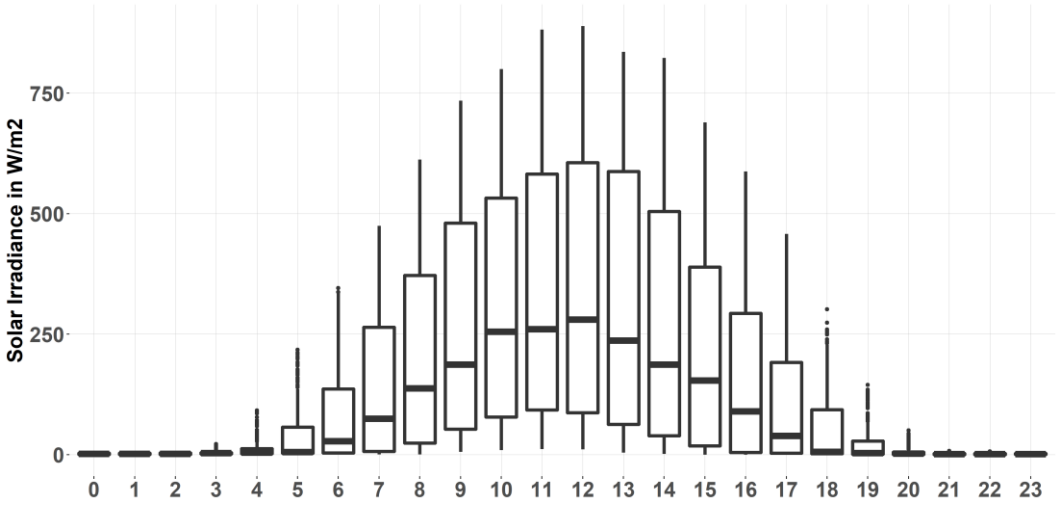


Fig. IV-4. Yearly global solar irradiance on the horizontal plane in Tartu (Estonia). Data for 2019.

The temperature distribution, T_{OUT}, presented in Fig. IV-3 shows that the minimum temperatures almost reach -20°C in the coldest months of the year, especially in January. The summer is quite hot, with maximum outdoor temperatures around +30°C in June and July. Almost all the months of the year present hour intervals with outdoor temperatures below 0°C. The only months with positive temperatures during all hours are June, July and August, coinciding with the summer period in Tartu[3].

Regarding the global solar irradiance, illustrated in Fig. IV-4, the hourly maximum solar energy received is below 1000 W/m^2 and the general profile of the daily solar irradiance follows a Gaussian distribution. The median value of this energy at 12am slightly surpasses 250 W/m^2.

Additionally, Fig. IV-5 and Fig. IV-6 characterize the wind speed (W_S) and wind direction (W_D), respectively. As observed in Fig. IV-5, the distribution of wind speed frequency follows a Weibull distribution [85].

[3] The Appendix section presents a data-based method for determining the beginning and the end of the summer period in function of the delivered heat.

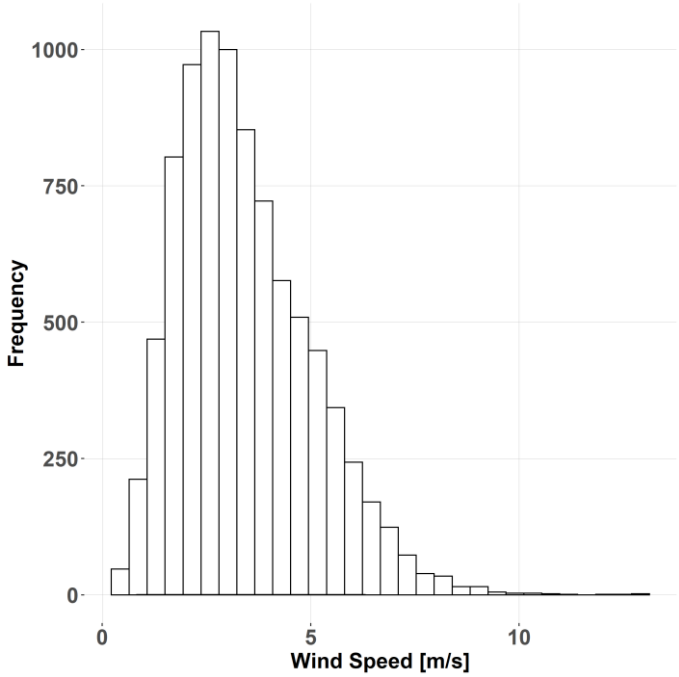


Fig. IV-5. Yearly wind Speed histogram in Tartu (Estonia). Data for 2019.

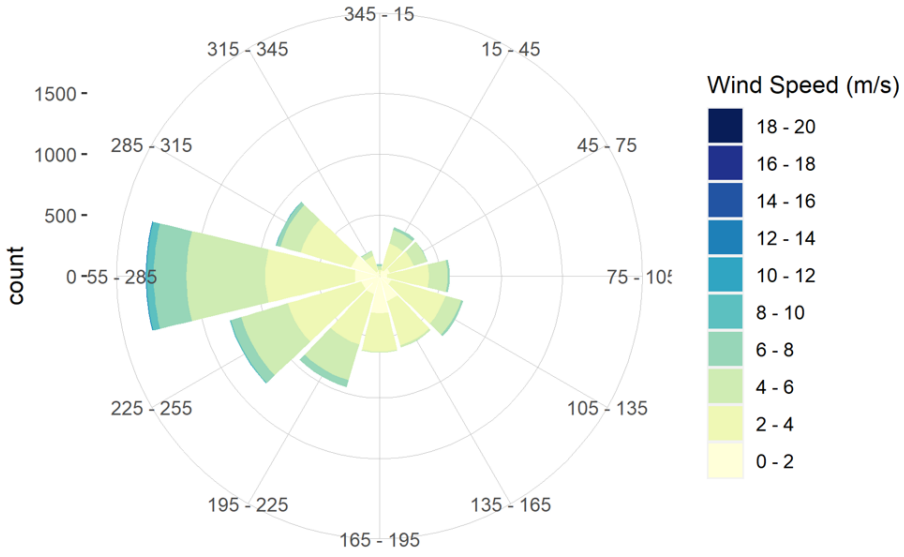


Fig. IV-6. Yearly wind direction Wind Rose in Tartu (Estonia). Data for 2019.

In addition to these weather data, it is essential to analyze the heat demand patterns from the buildings in the same location. It is performed in the following section.

4. Data from DH Substation. Buildings´ demand

The heating load profile consists of data from 43 substations of the DH network in Tartu (Estonia) (see The hourly energy use is calculated as the measured reading in that hour minus the measured value in the previous hour. Among the substations under study, different types of thermal zones can be found in terms of the final use or demand profile of the building. In this sense, residential apartments (also referred as private house), offices, educational buildings and commercial buildings are included.

This dissertation is focused on 43 different buildings connected to the DH in Tartu and the most relevant characteristics of these building are shown in Table IV-1. Each building is identified by a code (ID number), completely independent from the real address of the location to avoid any type of identification problems and to preserve the privacy of the users. Some buildings have a NO in the "DHW (Y/N)" column although that does not mean that there is no DHW demand in the building but that they might use other energy sources but DH to fulfill these requirements.

Table IV-1), all located in the sub-network of Tarkon. Each substation contains a smart meter that measures different variables in the system with an hourly frequency and sends it remotely to GREN, the DH operator. The energy meter installed in the buildings is the Multical® 603 from Karsmtrup [86]. The uncertainty of this device remains below 5% in all the measured variables, which is better than that specified in the European directive for this purpose (EN-1434-1:2015 [87]). Heat energy demand is saved as a cumulated variable and read hourly. The measures of each substation correspond to the features of one building. Fig. IV-7 presents the monitoring scheme of the heat meters used. As it can be observed, the smart heat meter measures additional operational variables in the system, such as the supply and return temperatures of the primary and secondary sides of the substation (from T1 to T5 in Fig. IV-7) and volumetric flow (m in Fig. IV-7). Nevertheless, this dissertation will on be focused on the total heat demand on the buildings, which comprises the sum of SH and DHW demand. Of course, these two

energy-demands are correlated with the rest of the operational variables in the system (temperature difference and volumetric flow) and consequently, cannot be used for the characterization of the total energy demand.

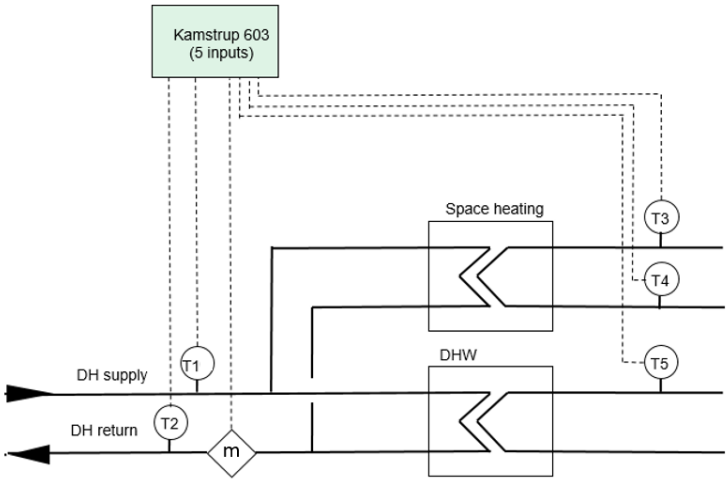


Fig. IV-7. Location and lay out of the smart energy meters in the DH in Tartu

The hourly energy use is calculated as the measured reading in that hour minus the measured value in the previous hour. Among the substations under study, different types of thermal zones can be found in terms of the final use or demand profile of the building. In this sense, residential apartments (also referred as private house), offices, educational buildings and commercial buildings are included.

This dissertation is focused on 43 different buildings connected to the DH in Tartu and the most relevant characteristics of these building are shown in Table IV-1. Each building is identified by a code (ID number), completely independent from the real address of the location to avoid any type of identification problems and to preserve the privacy of the users. Some buildings have a NO in the "DHW (Y/N)" column although that does not mean that there is no DHW demand in the building but that they might use other energy sources but DH to fulfill these requirements.

Table IV-1. Summary of the buildings connected to the DH in Tarkon-Tuglase (Estonia)

Building ID	Type of Use	DHW (Y/N)	Building ID	Type of Use	DHW (Y/N)
10045	Residential (Apartments)	Y	11163	School	N
10051	Residential	N	11164	School	Y
10258	Residential	N	11165	School	Y
10259	Residential	Y	11166	School	Y
10266	Residential	N	11195	Commercial Building	Y
10280	Residential	N	11491	School	Y
10298	Private house (residential)	Y	11494	Residential	Y
10512	Residential	N	11522	Residential	Y
10686	Residential	Y	11582	Residential	Y
10696	Residential	Y	11604	Residential	Y
10718	Residential	Y	11676	Residential	Y
10725	Residential	Y	11708	Residential	Y
10777	Residential	N	11718	Office	Y
10888	Residential	Y	11741	Residential	Y
10922	Residential	Y	11765	Residential	Y
10949	Kindergarten	Y	11780	Residential	Y
11008	Private house (residential)	Y	11794	Private house (residential)	Y
11009	Private house (residential)	Y	11795	Private house (residential)	Y
11014	Private house (residential)	Y	11836	Residential	Y
11015	Private house (residential)	Y	11860	Private house (residential)	Y
11044	Residential	N	11882	Residential	Y
11064	Residential	Y			

There is no additional information available about the external energy sources neither about the constructive characteristics such as: number of floors, windows-to-wall ratios or other useful information for characterizing the energy demand inside the buildings. Thus, the main objective of the work is to characterize the demand only using the climatic data.

As there is data for 43 buildings, the following figures will show some general information about the statistics of demand variable and the demand of some particular buildings. The images for the rest of the figures are annexed in the Appendix Section (Chapter XI).

Fig. IV-8 shows a boxplot with the hourly total demand of the 43 buildings under study. These boxplots enable to visualize the minimum, maximum, median and quartiles of the demand. The standard deviation of the data is represented by the height of the box in each of the buildings. A greater difference between quartiles means that the data is more diffuse and that the point dispersion is larger.

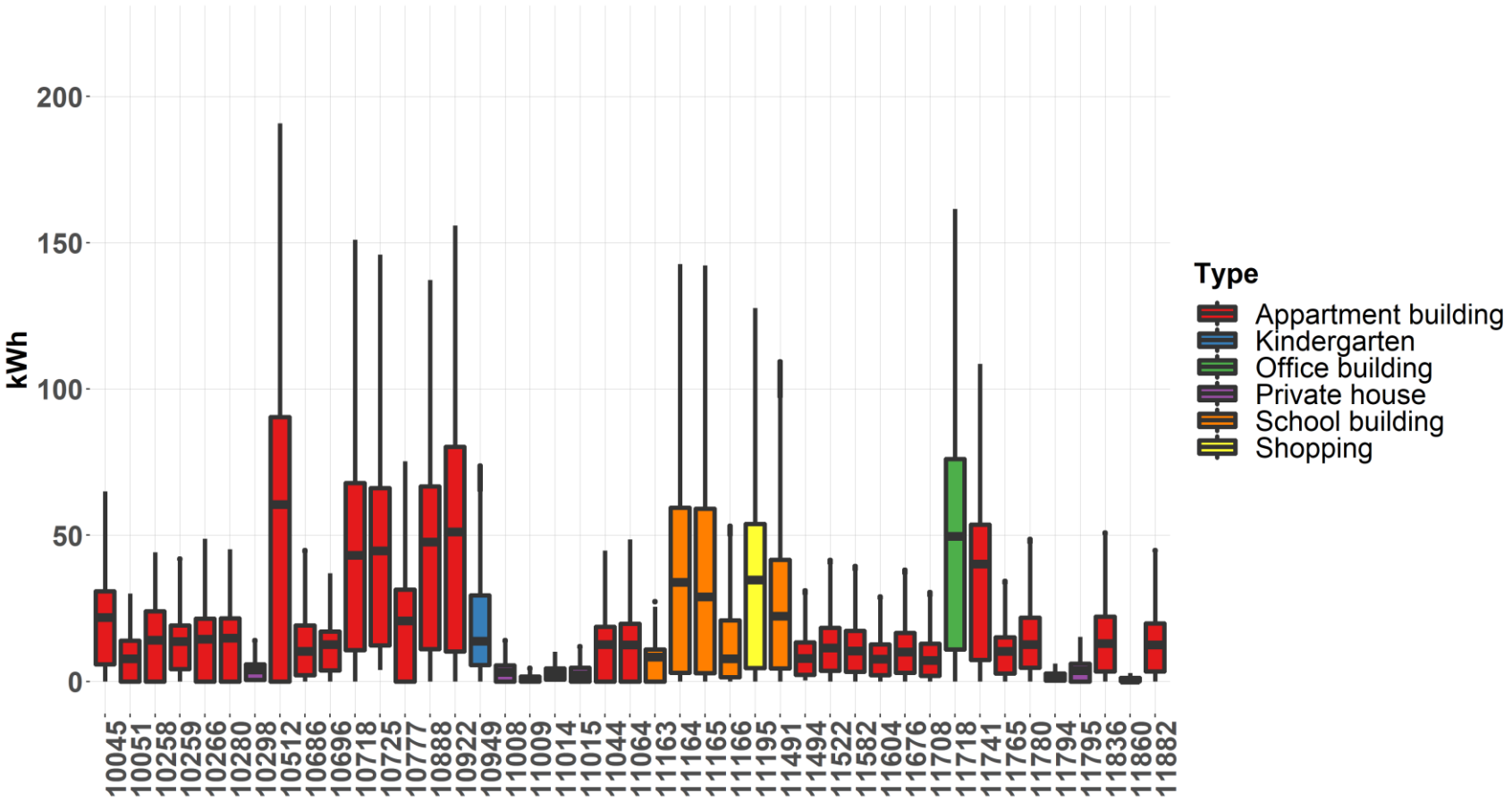
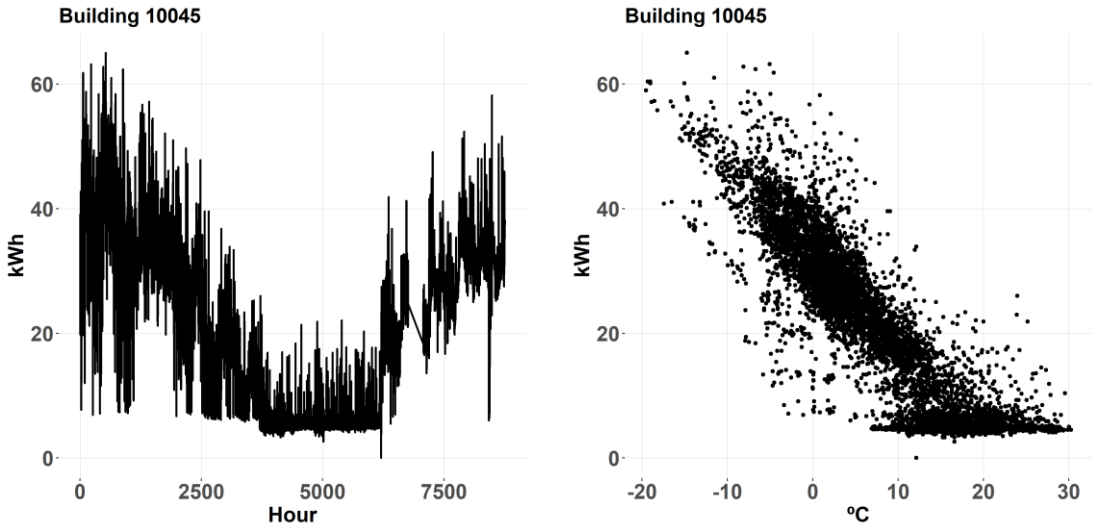


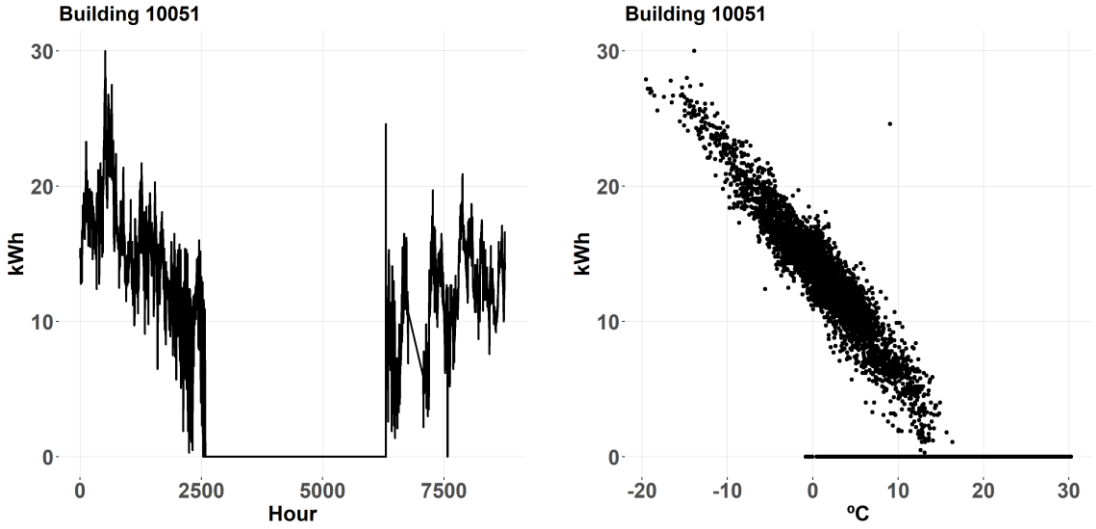
Fig. IV-8. Heating Demand statistics in the 43 buildings in Tarkon-Tuglase

Additionally, the following figures (Fig. IV-9, Fig. IV-10, Fig. IV-11 and Fig. IV-12) present the heating profile of certain individual buildings. These building are examples of each of the typologies studied: Building 10045 and Building 10051 are residential apartments with and without DHW, respectively. Additionally, Building 10949 and Building 11718 are offices and a kindergarten, respectively, both with DHW demand covered by the DH network. As commented above, only some representative examples are herein shown; the rest of the buildings are detailed in the Appendix 2 (Page 315). The pictures labelled with (a) present the hourly total heating demand hourly profile while those marked with (b) represent the relation between demand and the outdoor temperature, T_{OUT}.



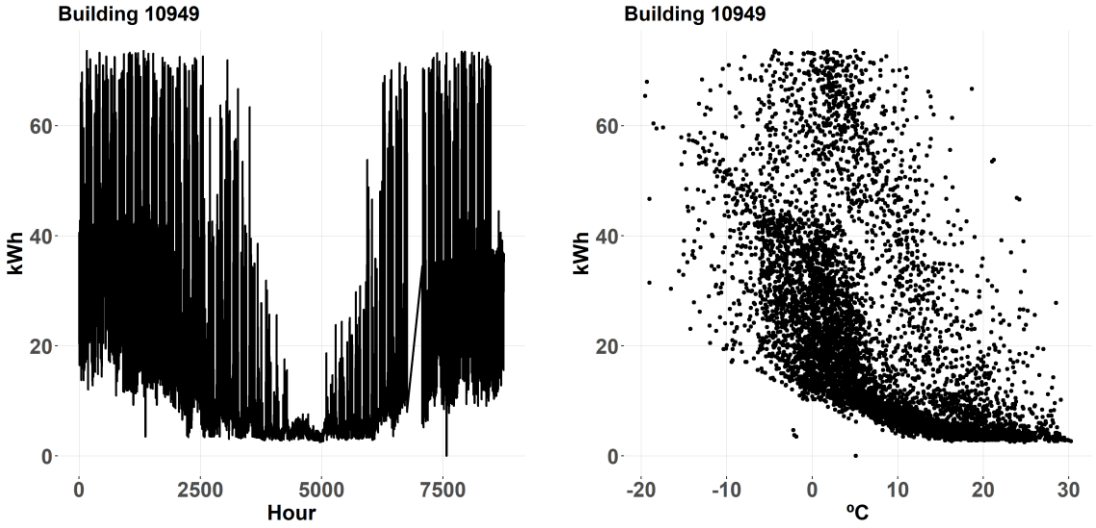
(a) (b)

Fig. IV-9. Heating year profile and Demand vs T_{OUT} for Building 10045 (Residential apartment)



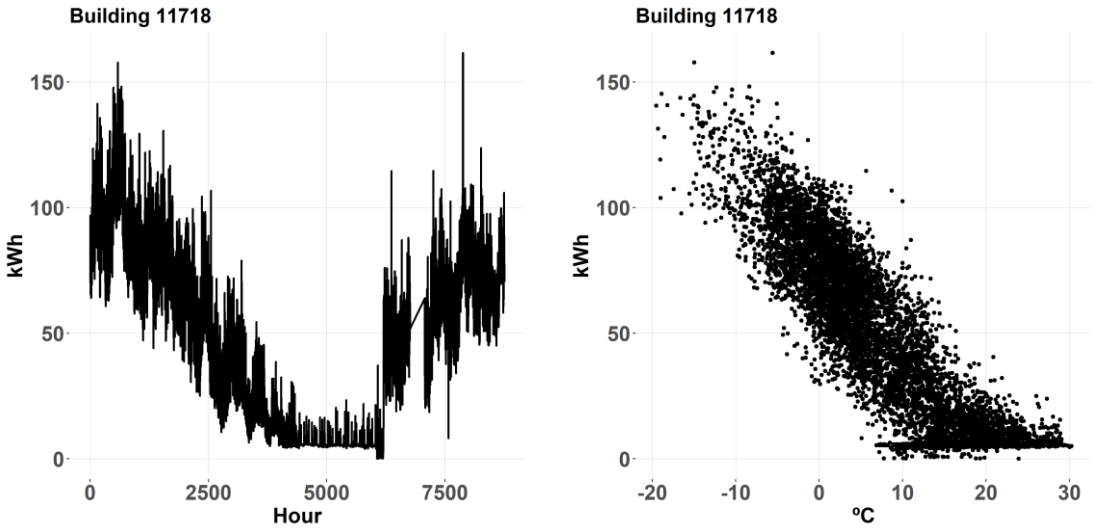
(a) (b)

Fig. IV-10. Heating year profile and Demand vs T_{OUT} for Building 10051 (Residential Apartment)



(a) (b)

Fig. IV-11. Heating year profile and Demand vs T_{OUT} for Building 10949 (Educational Building)



(a) (b)

Fig. IV-12. Heating year profile and Demand vs T_{OUT} for Building 11718 (Office)

To see the rest of the demands, go to Page 315.

Previous figures show the hourly profile of the demand (on the left) and the demand against the outdoor temperature (on the right) of four buildings. The first difference observed is the magnitude of the demand. While the office (Building 11718 Fig. IV-12) presents a maximum demand above 150 kWh, the first residential building (Building 10045 in Fig. IV-9) only reaches a maximum demand slightly above 60 kWh. Other remarkable difference that can be observed n this raw data is that some of the buildings do not present demand with high outdoor temperature, coinciding with the summer period. While Building 10051 presents a lot of observation with no demand, the rest of the buildings in this paragraph show a minimum demand above zero throughout all the year. This is caused by the DHW supply using other heating sources than the DH network. Additionally, in general, the nature of the demand is different in all the buildings. Some of the buildings present relatively low dispersion (Building 10051), while Building 10949 presents, at least, two or three clear trends in high demand zone.

Therefore, it is concluded that the nature of the demand is different in each building so that we need a general method that may adapt to the different “shape” of the demand, finding real relations (or correlations) with external variables that may be affecting the

demand. And we refer as external variables because they are not dependent on the buildings itself. The challenge of this dissertation is characterizing the demand in all the buildings using a unique model for all the buildings and only based on external variables, since there is no information about the characteristics of the buildings.

The following section presents an introduction to correlation analysis between external variables and the demand, which is the basis for the rest of the chapters.

5. Correlation between Data Sources

As a final part of the data presentation, it is important to discuss how the variables used for the study are correlated with the total demand in each building. For that purpose, the Pearson correlation coefficient and pairs of scatter plots are shown for Building 10045 and Building 10051 in Fig. IV-13 and Fig. IV-14, respectively. Pearson correlation shows the linear correlation between two variables, and it range between -1 and 1. A coefficient equal to -1 shows a negative perfect relation between the variables and +1 presents a perfect positive correlation. When this coefficient is near to zero means that these variables are hardly correlated and will not contribute with added value to data-driven models.

These two buildings are shown (Fig. IV-13 and Fig. IV-14) to cover different buildings with and without DHW demand, but the discussion would be very similar for the rest of the buildings.

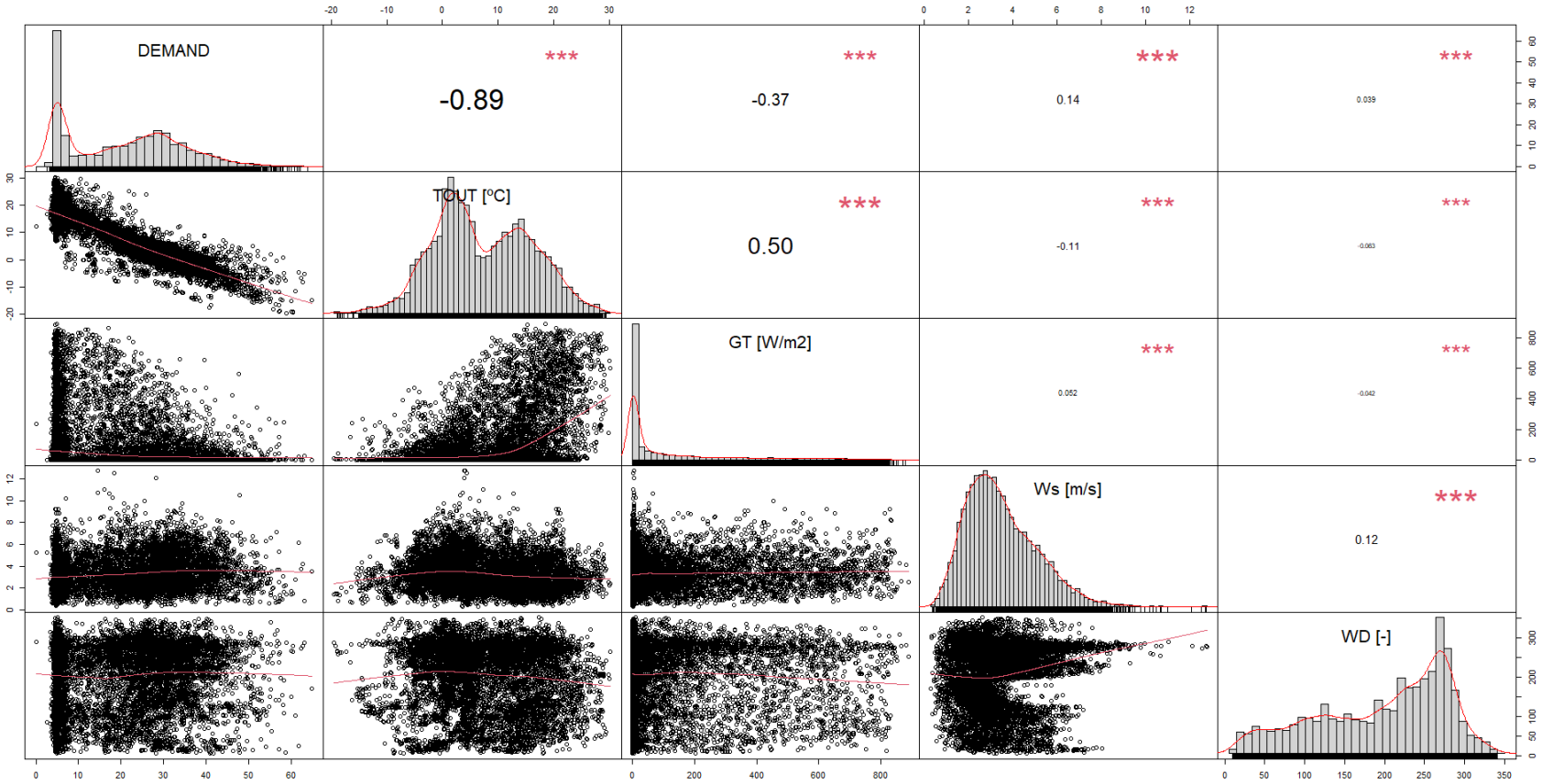


Fig. IV-13. Pearson Correlation between heat-load and climatic variables in Building 10045 (Residential with DHW demand).

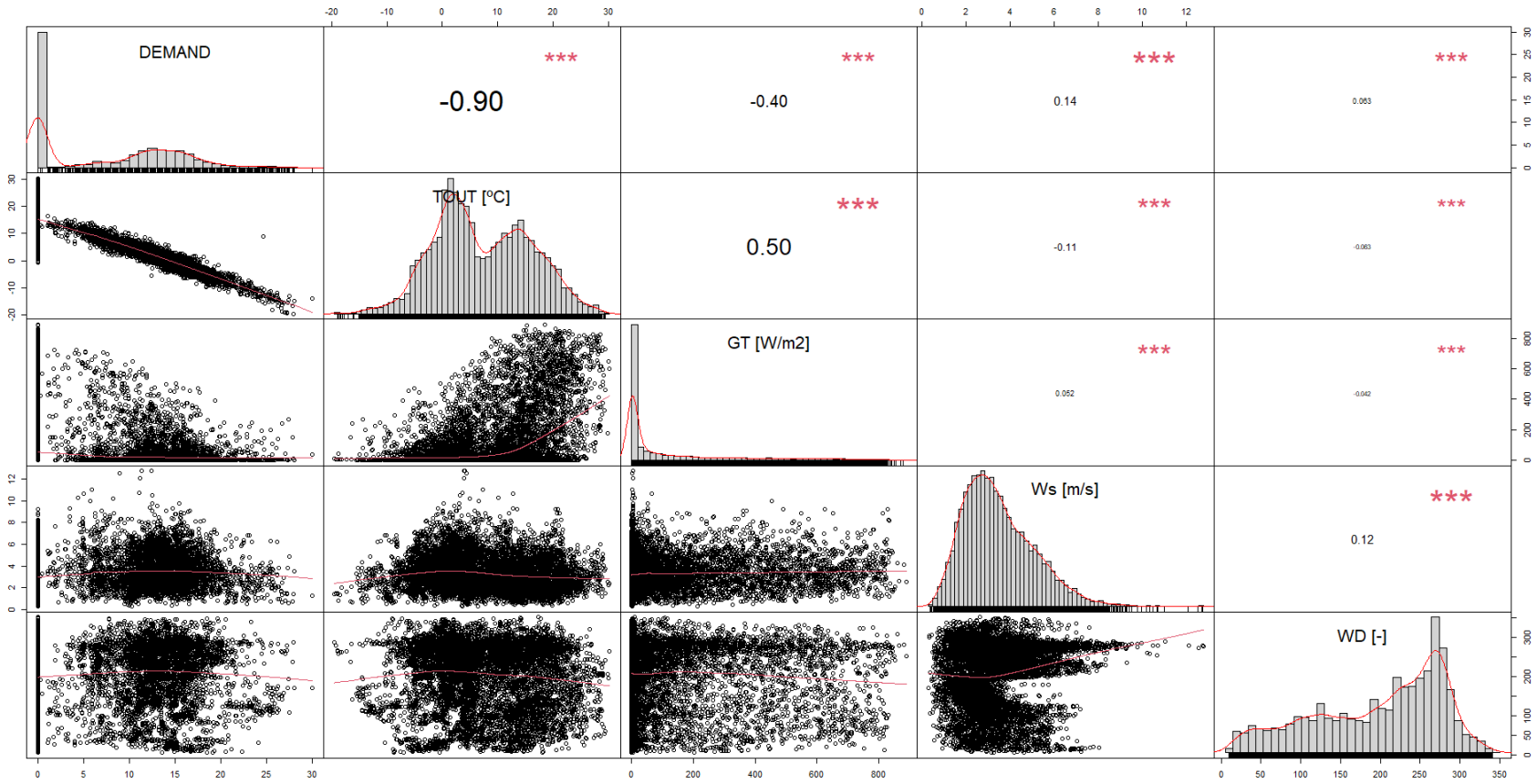


Fig. IV-14. Pearson Correlation between Heat-load and climatic variables in Building 10051 (Residential without DHW demand).

The following tables (Table IV-2 and Table IV-3) enables a better visualization of the coefficients shown in the previous figures.

Table IV-2. Pearson coefficients in Building 10045

Building 10045	DEMAND	T_{OUT}	G_T	W_S	W_D
DEMAND	1	-0.89	-0.37	0.14	0.039
T_{OUT}	-0.89	1	0.50	-0.11	-0.063
G_T	-0.37	0.50	1	0.052	-0.042
W_S	0.14	-0.11	0.052	1	0.12
W_D	0.039	-0.063	-0.042	0.12	1

Table IV-3. Pearson coefficients in Building 10051

Building 10051	DEMAND	T_{OUT}	G_T	W_S	W_D
DEMAND	1	-0.90	-0.40	0.14	0.063
T_{OUT}	-0.90	1	0.50	-0.11	-0.063
G_T	-0.40	0.50	1	0.052	-0.042
W_S	0.14	-0.11	0.052	1	0.12
W_D	0.063	-0.063	-0.042	0.12	1

The conclusions from the correlation analysis are the followings:

- The outdoor temperature and the solar irradiance are the variables with highest correlation to the demand, followed by the wind speed. Finally, the wind direction has almost no correlation with the demand.
- Outdoor temperature and solar irradiance have negative correlation with the demand. Thus, when the outdoor temperature is low (or the solar irradiance), the demand in the buildings increases and vice versa. The opposite trend is observed with the wind speed. When the wind is strong, the building losses by convective effects are higher and consequently, the demand increases.

- The correlation coefficients in Building 10051 are higher than Building 10045. This is caused by the lower dispersion of the demand observations in this building.

Chapter V

Data Analysis & Q-T Algorithm

Abstract

This chapter presents the first analysis of the data received from the smart meters in the substations of the buildings. This chapter analyzes the nature of the data and will present a novel method for outlier removal based on a density-based clustering algorithm. Additionally, and once that the anomalous values are removed from the dataset, it will present a novel method for characterizing the demand, the so-called Q-T algorithm. This chapter will study the efficiency of this method for hourly and daily frequency data.

Resumen

Este capítulo presenta un primer análisis de los datos recibidos de los medidores inteligentes en las subestaciones de los edificios. Este capítulo analiza la naturaleza de los datos y presentará un método novedoso para la eliminación de valores atípicos de la demanda basado en un algoritmo de agrupamiento no-supervisado basado en la densidad de puntos. Adicionalmente, y una vez que se eliminen los valores anómalos del conjunto de datos, se presentará un método novedoso para caracterizar la demanda, el denominado algoritmo Q-T. Este capítulo estudiará la eficiencia de este método para los datos con frecuencia horaria y diaria.

Chapter V Data Analysis & Q-T Algorithm

1. Introduction

As mentioned in previous chapters, predicting (with high accuracy) the energy demand in buildings connected to a DH network is a key factor for developing energy reduction strategies or for detecting anomalies in the networks.

As we commented in the State of Art chapter (Chapter III), regarding heat load forecasting alternatives for buildings, white-box model forecasting (using tools as EnergyPlus [31] or TRNSYS [32]), including their calibration against meter data became an opportunity for buildings and lower scale (elements of the building, for example). However, these methods are not valid at DH scale, as the DH utility does not have the required information to develop such models (architectural data, use patterns, etc.) and the model development and calibration process is considered to be time and resource intensive.

Similarly, grey-box approaches require to simulate complex and time-consuming differential equation to model the demand. These methods could be considered reasonable for building scale. However, each building would require the characterization of a differential equation and this model requires a large amount of information that is usually not available for the DH operators.

In contrast, black-box data-driven models do not require the differential equations that govern building physics to be understood and implemented. Such models are purely based on data and can be trained to infer relations between inputs and outputs using statistical techniques with no physical interpretation. Energy signature models are one of the simplest types of black-box models that express the heating energy use as a function of weather variables. They can provide successful results for low-resolution heat load predictions, such as monthly or seasonal data. In energy signature literature,

outdoor temperature is the most dominant weather variable ([41] or [88]). The usual choice of outdoor temperature as the unique predictor variable can be partially explained by the difficulty to access good historical data of other climatic variables. However, there are also studies, such as [89] where, outdoor temperature, global solar radiation and wind speed were used as the weather parameters for the models. In other studies, such as [90], relative humidity was also included. Results from all these studies concluded that outdoor temperature is the most influential parameter, although it is highly commendable to consider also solar radiation [79].

This chapter will explain, step by step, a self-developed Q-T algorithm which is considered as an evolution of simple energy signature models.

2. Objectives of this Chapter

This chapter is focused on the two following objectives:

- Analyze the relation between the heat-load, climatic variables and calendar variables for the DH in Tartu and assess the dependencies between these variables. This chapter will go one step further in correlation analysis and will explain how some of the variables are related.
- Developing a simple ML model (the so-called Q-T algorithm) based on the dependencies found on this chapter.

Thus, the main objective of this chapter is understanding the data and developing a simple model based on the general knowledge obtained in this study. The proposed model will be based on relatively simple equations and the low calculation/processing cost, and the consideration of any type of final use of the buildings modelled, makes it suitable for deployment on such large scales as full DH networks. Moreover, the model presented in this study aims to be valid for any type of building, regardless of the heating profile or final use, since the building stock connected to a DH network is usually made up of all kinds of building types.

3. Methodology

This section details the general methodology followed in this chapter. In terms of structure, section 3.1 provides an overview of two methods to identify the outliers or anomalies in the raw dataset presented in the previous chapter. The study of the anomalies will be limited to heat-Load variables since it is the variable to be characterized. Then, section 3.2 analyzes the relations between all the variables in the study and stablishes the basics for the model development using simple decision-trees (DT). Section 3.3 defines the equations that rules the model, while section 3.4 shows the error metrics that have been used for evaluation the efficiency of the model. Section 4 will show the results, divided by general results and specific results for some individual cases.

The described methodology is illustrated in Fig. V-1.

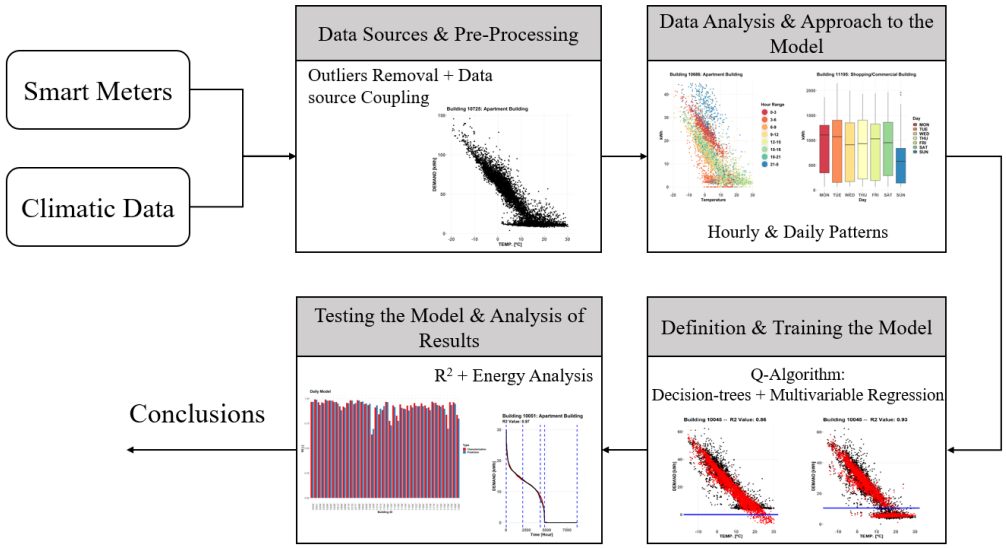


Fig. V-1. General Methodology followed for developing Q-T algorithm.

3.1. Pre-Processing of the Data: Outlier Identification

The first pre-processing activity that is required for preparing any original data set for a further analysis is the outlier removal. Outliers in a dataset are defined as anomalous observations that have been caused by reading errors or by unusual heating demand.

Usually, the reading errors are saved as NULLs or NAs by the installed energy meter and could be directly removed from the original data set. However, unusual observations can be caused by a sudden increase in the demand or by electricity peaks. Regardless of the cause, these observations can disturb the real nature of the demand and this way, decrease the accuracy of the data-driven methods subsequently developed.

Even though outliers and reading errors have been found in the two used data sources (data from substations and data from weather stations, described in Chapter IV), this study is only focused on identifying the outliers of heat-load demand variable (from substations). Then, the weather variables corresponding to the heat-load hour identifies as an outlier will also be removed from the original dataset.

Reading errors are directly removed from the original dataset, reducing the total data points available.

For the identification of outliers, the following two methods are proposed:

- **IQR (Interquartile) Method**
- **Density Based Clustering**

3.1.1. IQR Method

For the identification of the outliers, quartiles of each variable are calculated using a boxplot function in R [91]. R software is a programming language for statistical computing and graphics supported by the R Core Team and the R Foundation for Statistical Computing. For a specific variable such as heat demand, the statistical distribution consists in the identification of three quartiles (Q1, Q2 and Q3). The first quartile, Q1, is defined as the value in which the 25% of the observations are below. The second quartile, Q2, is defined as the median value and finally, the third quartile corresponds with the 75% of the data below this value. The interquartile (IQR) variable is defined as the difference between third and first quartile, as it is observed in Fig. V-2. The values above Q3 + 1.5·IQR and the value below Q1-1.5·IQR are considered as

outliers, coinciding with the values that are the furthest from the median value of the variable. All this process is illustrated in Fig. V-2.

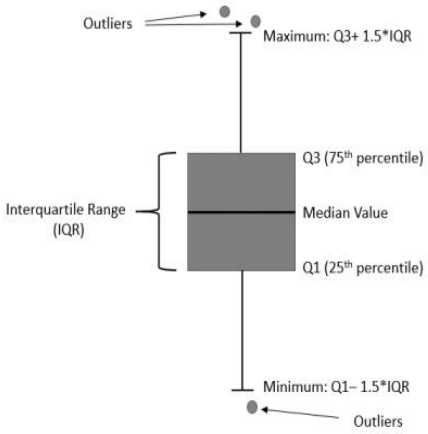


Fig. V-2. IQR Method for outlier removal

This methodology is only based on the statistical distribution of the observations and may remove values that are not real outliers. However, as an initial approach to outliers' removal this method is quite efficient and is usually used in similar studies such as [92] and [93].

3.1.2. Density Based Clustering

The other main methodology studied for outlier removal is based on an unsupervised clustering method named as density-based clustering or DBSCAN. The objective of this algorithm is to identify high-density observations that are closely together and the points that are identified in low-density areas are considered outliers. This algorithm was firstly proposed in 1996 by M. Ester et al. [94] and the implementation of this algorithm in R is made using the library *"dbscan"* [95].

This density-based clustering algorithm has the main advantage that is not required to pre-assign the number of clusters (unlike other clustering algorithms such as K-means) and that the outliers usually coincide with the isolated points in the low-density areas identified with this algorithm. The initialization of DBSCAN is made with the process called hyperparameter tuning, by the determination of the following two parameters:

Eps and *MinPts*. *Eps* represents a radius used from different observations in the algorithm and *MinPts* is used to denote the minimum number of observations within the circular region formed by *Eps* are key parameters in this clustering process. The density-based clustering algorithm starts with an arbitrary point selection and generating the neighborhood based on the *Eps* selected. The number of points within this region are assessed; if the number of points is higher than *MinPts*, this point is labelled as a core-point and the cluster formation starts. Otherwise, the point is labelled as noise. However, this point can be later found within the *Eps* neighborhood of a different point and thus can be part of the cluster. All the points within the region will be part of the cluster if they are also core points and this process continues until the density-connected cluster is completed. The next step is to randomly select another point that has not been selected in previous steps and apply the same procedure. After all the points are processed, the points that are not assigned to any cluster are labelled as noise.

The main difficulty of this algorithm is the optimization of these two parameters. According to the authors of the library "dbscan" [95], the variable *MinPts* is initialized as the dimensionality of the dataset or dimensionality of the dataset plus one. An overly small *Eps* can cause that values that are not outliers are considered as outliers and an overly high *Eps* value cause that the outliers are not identified. For the optimization of this step, calculation of K-nearest neighbors' (k-NN) distances is carried out and the elbow of the ascending ordered distances correspond the optimal *Eps* value. The elbow of that curve is calculated by means of the second derivative of the k-NN distances.

3.2. Initial Data Analysis, Correlation Analysis and Modelling Approach

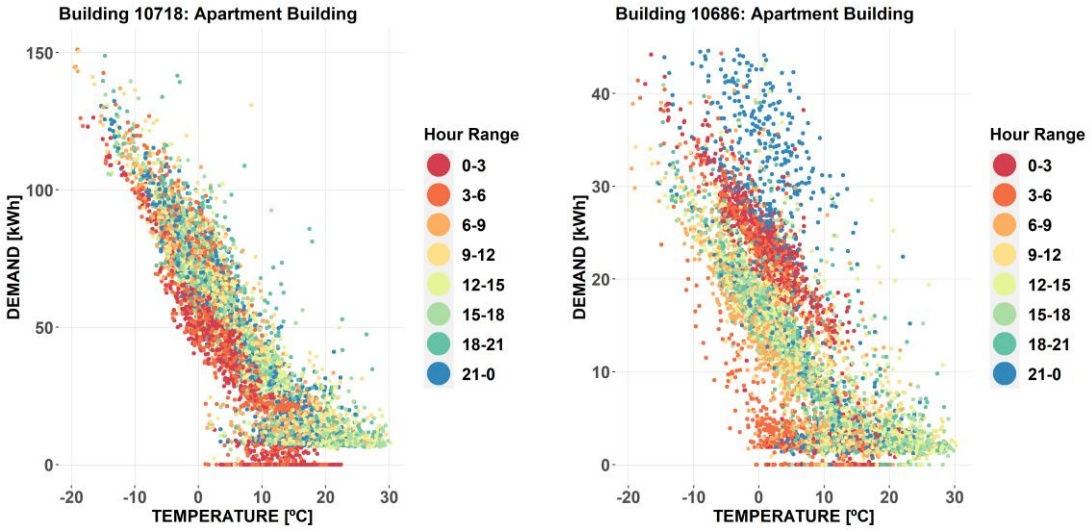
In a first analysis of the heating energy demand, a range of different heat profiles were found among the different buildings under study. These can be attributed to the different final uses of the buildings and the energy demand patterns of the users in their respective dwellings. As it has been presented in Table IV-1, some of the buildings included in the study show energy demand only for SH purposes (e.g., Building 10051,

Building 10512, etc.); whereas other buildings consume energy from the DH network for both SH and DHW production (e.g., Building 10045, Building 10718, etc.). In all cases, the measured heat-load represents the total heat demand of the building (see location of the energy meter in Fig. IV-7). Even though the district is composed by several building typologies, the model presented in the following section aims for a general application to any building, independently of the usage or heat profile of that building. The model and the method to train will be the same in all the buildings under study.

The energy required to satisfy the SH demand is dependent on both the climatic variables and the characteristics of the building (such as geometry and thermal envelope). It can be anticipated that when the outdoor temperature is low or the solar irradiance is low, the demand for space heating demand will be higher and thus, the weather variables and SH demand show a large correlation. However, DHW demand shows little to no dependence on climatic variables and primarily responds to use patterns and seasonal variations (e.g.: a young worker and a retired person are expected to have quite different DHW demand profiles). As an example, the heat demand of Building 10051 is not affected by DHW demand, whereas, in Building 10725, part of the heat demand is dedicated to that purpose. These figures have been shown in Chapter IV (Fig. IV-9 and Fig. IV-10) or Chapter XI (Appendix).

A night setback or a reduction in the demand has been identified in certain buildings, where heat energy demand patterns differ along different hours, independently from the climatic variables of that moment, incorporating a time dependency into the demand. Thus, calendar variables and heating demand variables are correlated, as it was proven with Fig. IV-13 or Fig. IV-14. The night setback can be used by the DH operator to reduce energy production in periods when a low heat load is expected, regardless of the climate conditions. Moreover, the high thermal inertia of the DH network could be used to satisfy the possible heat energy demand at night. In this context, Fig. V-3 shows the heat profile of two buildings under study where a night setback has been identified. In both buildings, a reduction of the heat load is identified between 3AM to 5AM.

The mining of the effect of the night setback in the demand results in the definition of the first level for the decision-tree (LVL3 in Section 3.3) for application in the model.



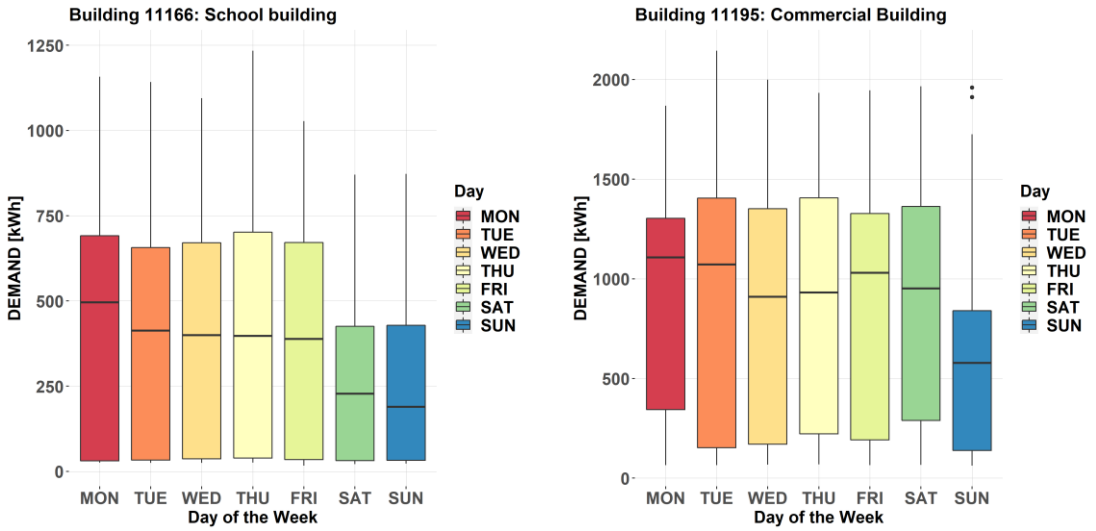
(a) (b)

Fig. V-3. Identification of night setback in Building 10718 (a) and Building 10686 (b)

The usually low activity of the occupants at night gives the possibility to DH operators to carry out demand flexibility strategies in order to reduce energy resources in the system. Even though the size of the buildings under study is unknown, large buildings have high thermal inertia that permits to maintain an indoor temperature within the comfort limits for some hours. These thermal inertias in the system (building and network) enable to control the demand peaks and reduce energy use in the network.

Furthermore, the daily aggregated heating energy demand profiles allow the energy share used for daily DHW to be identified, as shown in the next section. However, in the same way as has been done for the hourly data, additional time-dependent patterns have been identified in the daily aggregated data. In buildings that have no occupation at the weekends (e.g., offices or schools), this phenomenon is more noticeable. Fig. V-4 shows an example of how heat energy demand varies with respect to the day of the week, by means of a boxplot of the quartiles of daily heat energy demand. It can be observed that Building 11166 presents a lower demand on Saturdays and Sundays,

matching the days of no occupancy. In the same manner, in Building 11195, which corresponds to a commercial building, only presents lower demand on Sundays, as this type of building is usually closed on this day. As a general conclusion, heat demand at the weekends is lower than on weekdays in some of the buildings. This is caused by the lower or non-occupancy of the buildings those days. This effect leads to the definition of the second variable of the decision-tree (LVL2 in section 3.3).

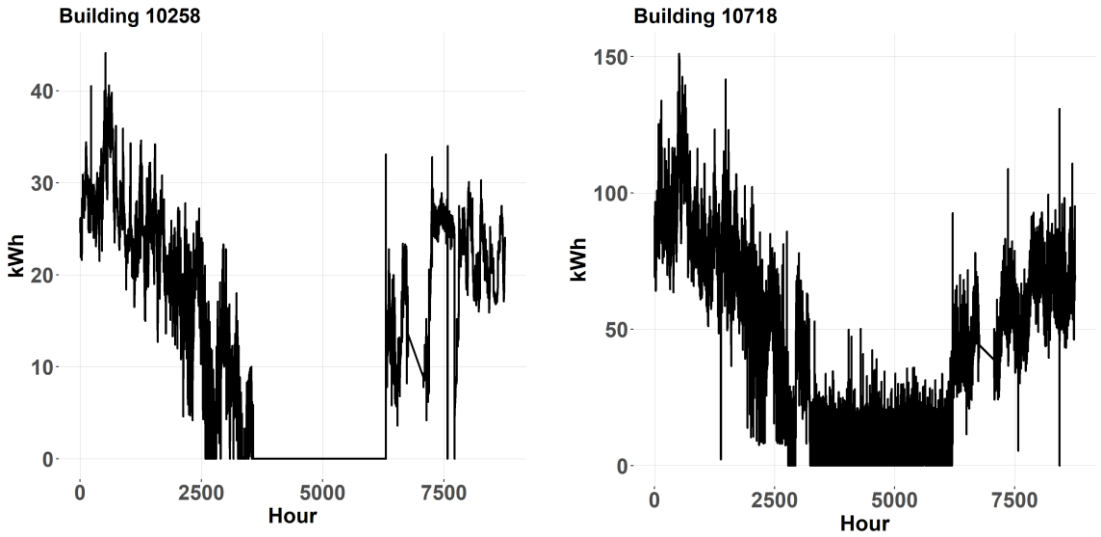


(a) (b)

Fig. V-4. Daily demand patterns in Building 11166 (a) and Building 11195 (b)

Specific seasonal patterns have been also identified, grouped in two main periods: summer & rest of the year (referred as REST). Despite undergoing relatively low external temperatures at some moments of the summer, the monitored heat energy demand does not correspond to the expected values for similar climatic conditions outside this season. This divergence could be motivated by a reduction of the heat load by the DH operator in this period.

As an example of the seasonal heat-load variation, Fig. V-5 presents the yearly demand profiles for two residential buildings (Building 10258 and Building 10718). The demand in the intermediate part of the year, the so-called “SUMMER” is lower than the rest of the year.



(a) (b)

Fig. V-5. Seasonal demand patterns in Building 11166 (a) and Building 11195 (b)

This phenomenon is not identified in all the buildings under study, and, in consequence, a general methodology is necessary for the identification of the summer period performance. It has been observed that the variability of the demand in the summer period is much lower than the variability in other periods of the year. Consequently, the summer heat demand follows a more stable (less varying) profile through time. Indeed, the comparison of the standard deviation between data periods was found to be an accurate method for identifying the relevant summer period for each building. Batches of 15 days were selected, and this methodology was applied to all the buildings under study. As a result, from the application of this method, the start and finish day of the summer are obtained for each of the buildings. These periods differ from building to building and this is way this methodology is applied to all the buildings independently. This list of days will serve as the input variable in the decision-tree (LVL1 in section 3.3).

It can be concluded that energy demand in the analyzed buildings is highly dependent on the weather parameters and also on the specific user-behavior. The latter factor is frequently omitted in modelling tools due to its random nature, which adds a significant complexity to the problem. However, it has been considered here for the sake of accuracy. The developed model is detailed in the following sections.

3.3. Definition of the Model. Q-T Algorithm

As a first approach for the mathematic characterization of the model, it was proposed to split the data in two parts by a specific temperature threshold. The demand data matching an external temperature above that threshold was attributed to periods with no space heating demand (no demand or DHW demand only); whereas the data below that temperature threshold would also entail SH demand. However, unsuccessful results were obtained, since this initial premise was not representative of most of the buildings and a large part of the data was not included in the characterization process by the model. That first approach was named as T-Algorithm, but it was finally discarded for the low accuracy that was capable to obtain.

As a more suitable alternative, we decided to use the heat load as the threshold and the following type of equation is proposed, the baseline equation for the so-called Q-T algorithm:

$$\mathrm{Q_{alg}} \quad \mathrm{Q} = \begin{cases} \alpha_1 \cdot T_{OUT} + \alpha_2 \cdot G_T + \alpha_3 \cdot W_S + \alpha_4 \cdot W_D, & Q < \mathrm{Q_{REF}} \\ \alpha_0, & Q \geq \mathrm{Q_{REF}} \end{cases} \qquad \text{Eq. (4)}$$

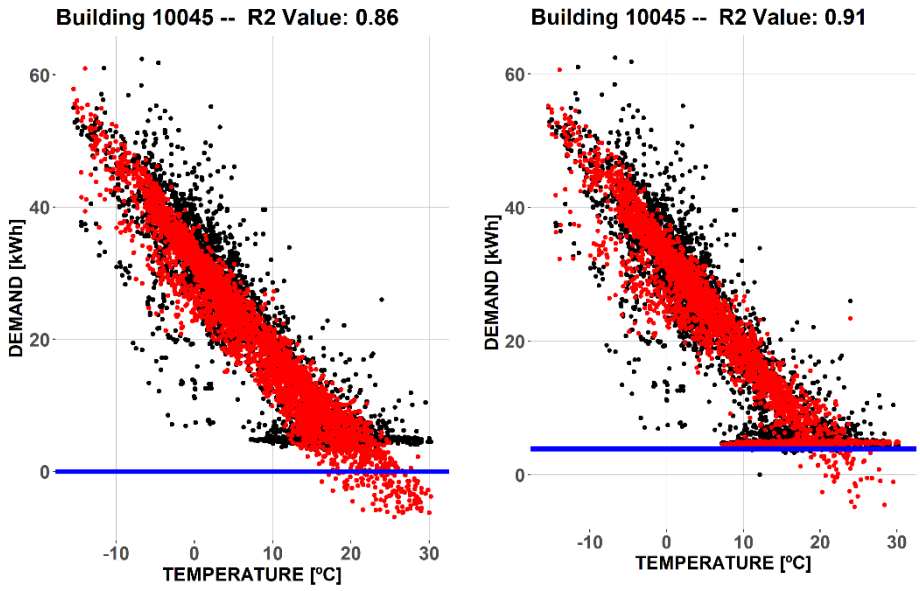
Relative humidity is not included in the model proposed in this work, since the climate in Tartu is very cold and dry, its impact on the heating energy use thus being low.

In this algorithm, a calibration process must be performed using training data to obtain the coefficients needed for the application of the same model to testing data. For this calibration process, the data are split by a reference heat load, Q_{REF}. The data below this reference load would not be weather dependent, whereas the data above this point is assumed to follow a linear correlation with the abovementioned set of climatic variables. The process for the calculation of Q_{REF} is carried out in an iterative manner by using a range of different heat load thresholds to split the data, ranging from a minimum of Q = 0 to a maximum of $0.5 \cdot Q_{MAX}$.

DHW corresponds with the horizontal trend observed in Fig. V-3, where the heat-load maintains relatively constant despite different outdoor temperatures. Observing data,

the instant DHW demand never exceeds 40% of the maximum load in any building. Thus, 50% is taken as the maximum limit for this iterative process. The absolute error of the regression is calculated in each step, so the heat load that minimizes the error in the second part of the equation ($Q \geq Q_{REF}$) determines the Q_{REF} value. This same algorithm logic is applied to both hourly and daily data.

The iterative process proposed for the calibration of the model is replicated for all the buildings under study. As expected, different calibration coefficients are obtained for each of the buildings in the district. Fig. V-6 illustrates 4 steps (the number of iterations for each building are 50) of the iterative calibration process for one of the buildings studied: in this example, the third one (bottom left) would represent the most accurate choice. Together with the figure of the iterative process, the R^2 value obtained in each of the regressions is shown. Note that the Q_{REF} value is not necessarily equal to the base DHW demand. In all the cases the heat demand for DHW is equal or less than Q_{REF}. In other words, the optimal Q_{REF} is the same or higher than the constant part of the demand in Fig. V-6 (third step).



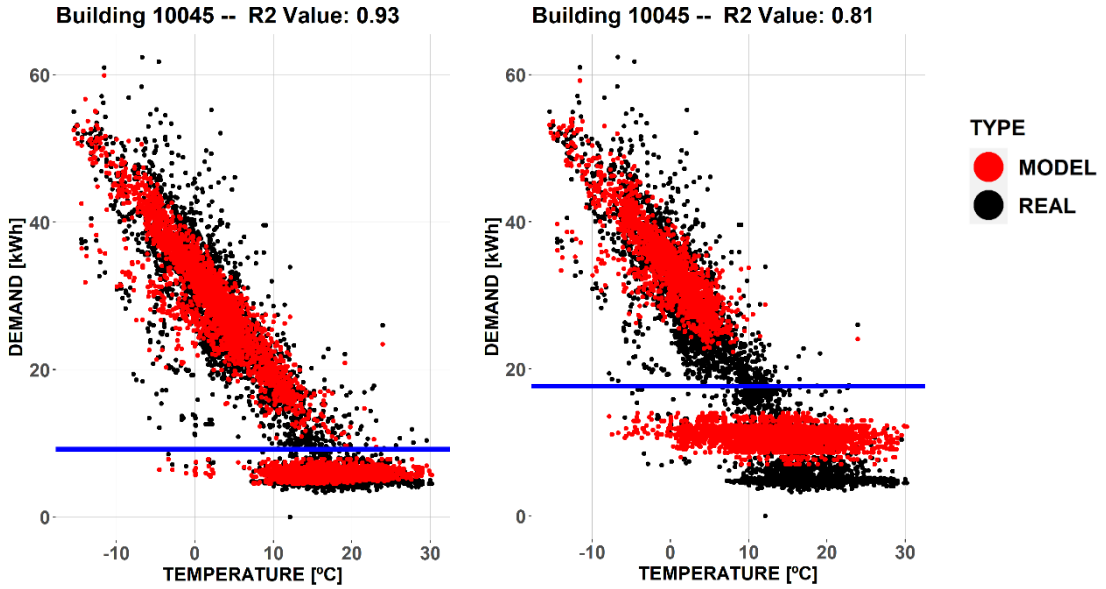


Fig. V-6. Four steps of the calibration iterative process of one building (Building 10045, residential) using hourly data.

As concluded from the previous section, heat demand data is not only weather dependent but also time dependent, following different demand patterns as a function of the hour of the day, day of the week and day of the year. But if Eq. (4) is applied to small batches instead of applying to the whole data, characterization of heat demand is more precise. Therefore, and in order to obtain a more accurate result, decision trees (DT) are proposed to be applied in three different levels, resulting in the so-called Q-T algorithm. Decision trees are non-parametric supervised techniques that predict values of responses by learning decision rules derived from features. For this model, the following three time-variables or features are introduced:

- **LVL1:** Variable season, divided into summer and rest of the year (SUM/REST)
- **LVL2:** Day of the week (MON, TUE, WED…)
- **LVL3:** Hour of the Day (1AM, 2AM, 3AM…)

This supervised classification process enables the characterization of a dynamic problem using stationary equations. The first level of the DT enables the characterization of the possible seasonal variations in the demand, as observed in some of the buildings under study. Besides, daily and hourly levels of the DT allow to introduce the influence of user-

behavior in the demand and identifying the different heat demand patterns shown in Fig. V-3 and Fig. V-4 for some of the buildings.

Therefore, for the hourly model, each hour is classified by the consecutive application of the three levels of DT, whereas for the daily data model, only the two first levels of the DT are used for the corresponding classification. The classification by means of the supervised clustering method results in different equation coefficients for each data subset, increasing both the calculation cost and the accuracy of the proposed model.

The whole process, including the decision trees and the abovementioned iterative process of the Q-T algorithm, is applied to the training data to obtain the parameters that make up the model for each of the buildings. Then, the fitted model is applied over testing data to verify the accuracy of the model.

3.4. Training and test datasets and metrics employed for result analysis

Finally, training and test datasets are determined. Different demand patterns have been recognized with respect to the season of the year. In order not to exclude these demand patterns, training and testing data are defined containing odd and even days, respectively. The data from odd days have been used to calibrate and train the models; whereas data from the even days have been used to test and verify the model's performance.

The accuracy and efficiency of the model is numerically evaluated by the R squared value or coefficient of determination, R^2. This value represents the proportion of the variance that is predictable using the predictors of the model. The R^2 variable is calculated as follows:

$$R^2 = 1 - \frac{SSE}{SSYY} \qquad \text{Eq. (5)}$$

$$SSE = \sum_{i=1}^{N} (X_i - Y_i)^2 \qquad \text{Eq. (6)}$$

$$SSYY = \sum_{i=1}^{N} \left(X_i - mean(X)\right)^2 \quad \text{Eq. (7)}$$

However, the approach for the evaluation of the accuracy of the model is not only based on the calculation of the R^2 value and its analysis. The practicality of the model resides in the prediction of the heating demand so that the heat generation process can be optimized. The DH operator is responsible for the management of the heat production process in the entire DH network, and in this context, the analysis of the model's accuracy also is evaluated in energy terms. Adopting the R^2 value as the only criterion can favor an overfitted or biased model. For the application assessed in this study, the high thermal inertia of the DH network could assume these fluctuations and, therefore, the analysis focuses on global energy results.

The other metric used evaluates total difference between the predicted demand and the real demand in a complete year. Thus, the total yearly energy demand deviation (YEC) is calculated as follows, where 0% indicates a perfect match between measurement and prediction. This metric is comparable with the abovementioned Mean Absolute Percentage Error (MAPE) mentioned in Chapter III.

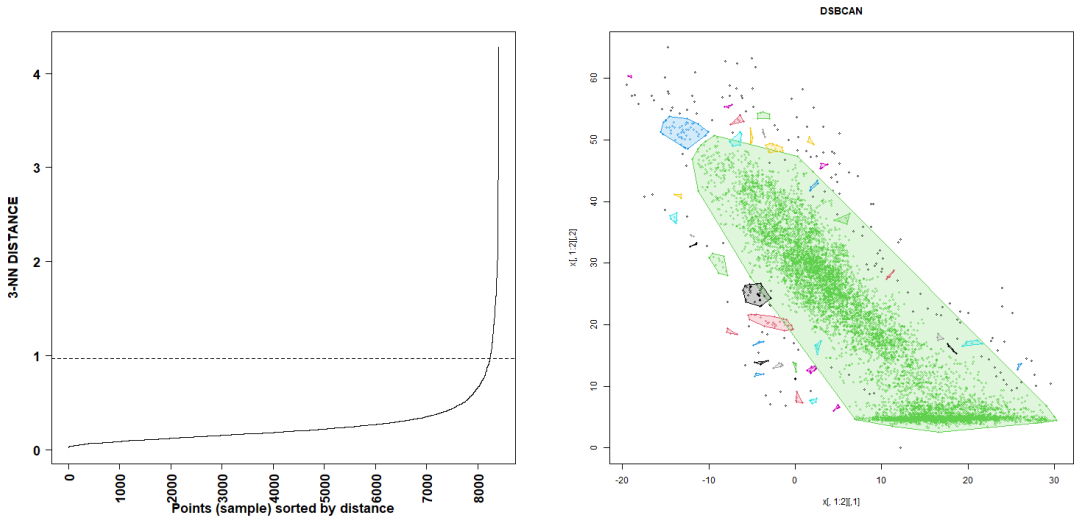
$$\mathrm{YEC} = 100 \cdot \frac{\left|\sum_{i=1}^{N} X_i - \sum_{i=1}^{N} Y_i\right|}{\sum_{i=1}^{N} X_i} \quad \text{Eq. (5)}$$

4. Results

This section is divided into general results for all the buildings and a specific analysis of some of the buildings as representing the whole dataset. First, the results obtained for the outlier identification process are presented, followed by the energy characterization obtained with the self-developed Q-T algorithm.

4.1. Outlier Identification

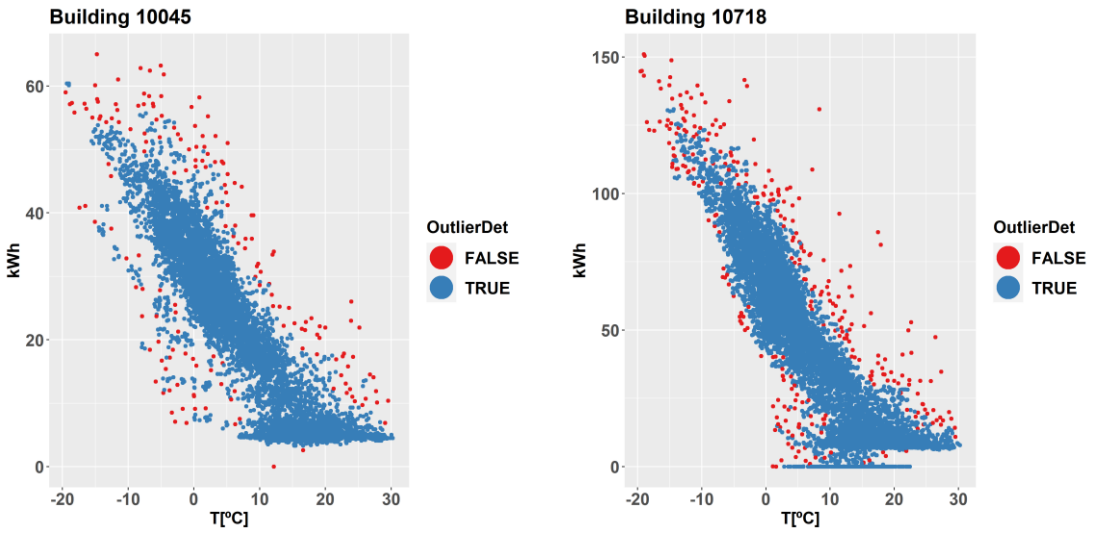
Due to the scattered nature of heat demand data and its correlation with climatic variables, we decided to use density-based clustering for the identification and removal of the possible outliers from the original dataset. Outdoor temperature (T_{OUT}) and solar irradiance (G_T) are found to be the climatic variables with the highest correlation to heat demand (as concluded in Section 5 from Chapter IV). Thus, DBSCAN algorithm is applied to the hourly heat demand against these two variables and, consequently, *MinPts* is initialized as three. The 3-NN distance is calculated, and the *Eps* variable is calculated in the elbow of the sorted curve. As a first approach to outlier detection, first overall results for all the buildings are shown followed by the step-by-step images for one of the buildings. Focusing on this process in one of the buildings, Fig. V-7 shows the sorted 3-NN distance on the left, and the clusters formed in the unsupervised process on the right side of the figure.



(a) (b)

Fig. V-7. (a) 3-NN sorted distance and (b) clusters formed in DBSCAN process in Building 10045

Finally, the output from the algorithm is a Boolean vector with FALSE/TRUE, corresponding FALSE with the outliers' observations. Fig. V-8 presents the result for two of the buildings under study: Building 10045 (left) and Building 10718 (right).



(a) (b)

Fig. V-8. Outliers identified in (a) Building 10045 and (b) Building 10718

Even though only two buildings are shown, the type of outliers identified by this algorithm is similar in all the buildings: individual observations that are far from the data core. These outliers could be: (i) Heat demand values equal to cero (in buildings with domestic hot water demand), (ii) High heat demand values with high outdoor temperatures and extremely high and low demand.

Fig. V-9 summarizes the number of outliers identified in each building in function of the observations in the original data. The same DBSCAN process is carried out for all the buildings under study, independently from the building type or any other building characteristics. Blue points represent the number of observations in the raw data (after removing the NAs) and the red points correspond with the amount of data after the application of density-based clustering and after removing the outliers.

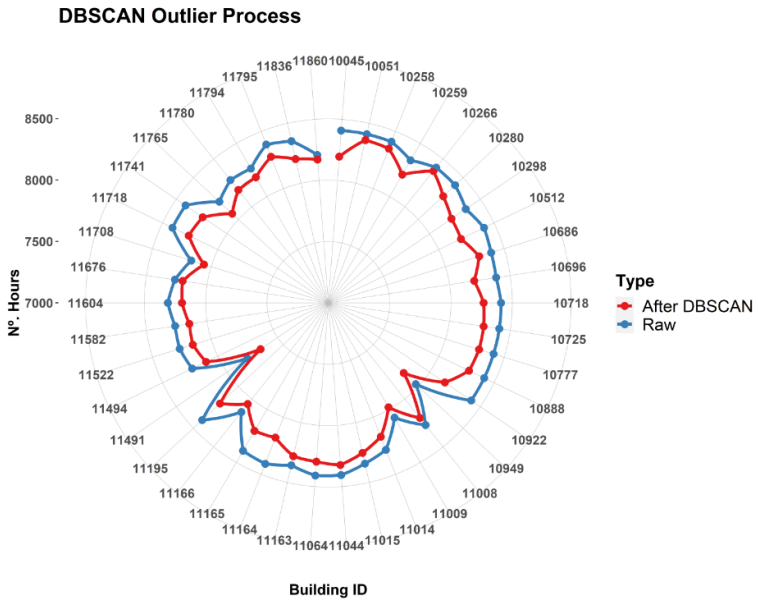
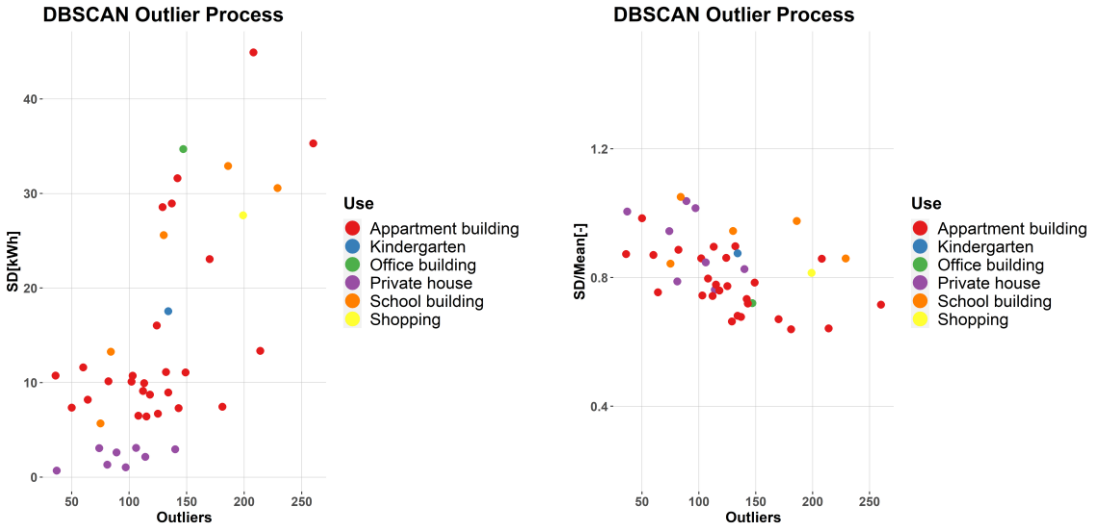


Fig. V-9. Summary of the number of points identified as potential outliers.

Fig. V-9 presents the number of hourly measures before and after the application of density-based clustering in blue and red respectively. Note that the radial axis starts (the central point of the circumference) at 7000 hours and a full year contains 8760 hours (2019). Due to the measuring errors in both data sources (smart meters in the buildings and weather station), the raw data is reduced to around 8400 available yearly reading in the best cases.

The number of outliers range between 37 and 260 hours in different buildings. This difference is consequence of the nature of the data in each case and for analyzing this difference, original data is compared with the number of outliers. The DBSCAN clustering algorithm sets its basis on identifying the outliers as lonely points out of the trend of most of the data. Hence, heat demand data with higher variability is supposed to result on a larger number of outliers. Moreover, the climatic dependence of heat demand (analyzed in Section 5 from Chapter IV) makes this effect more important, since a high SH demand in summer may be considered an outlier or the same with low SH demand in cold days/hours. Standard deviation of heat-load represents the variability degree and consequently, Fig. V-10 presents the number of outliers against the standard

deviation of heat-load and the same variable divided by the mean value of the heat-load.



(a) (b)

Fig. V-10. Number of outliers vs (a) Standard deviation and (b) Standard deviation divided by mean demand.

In Fig. V-10 the correlation between the number of outliers identified by DBSCAN and the standard deviation for all the buildings (grouped by type of use) is observed. On the left side of the figure (Fig. V-10a) it is observed the positive correlation between the standard deviation of the data and the number of outliers identified by the algorithm. When dividing the standard deviation with the mean heat demand of each building (Fig. V-10b), this effect is the opposite. The number of outliers identified by this algorithm increases with the deviation of the data and with high mean heat demand. This is caused because the algorithm usually considers the peak demand as outliers.

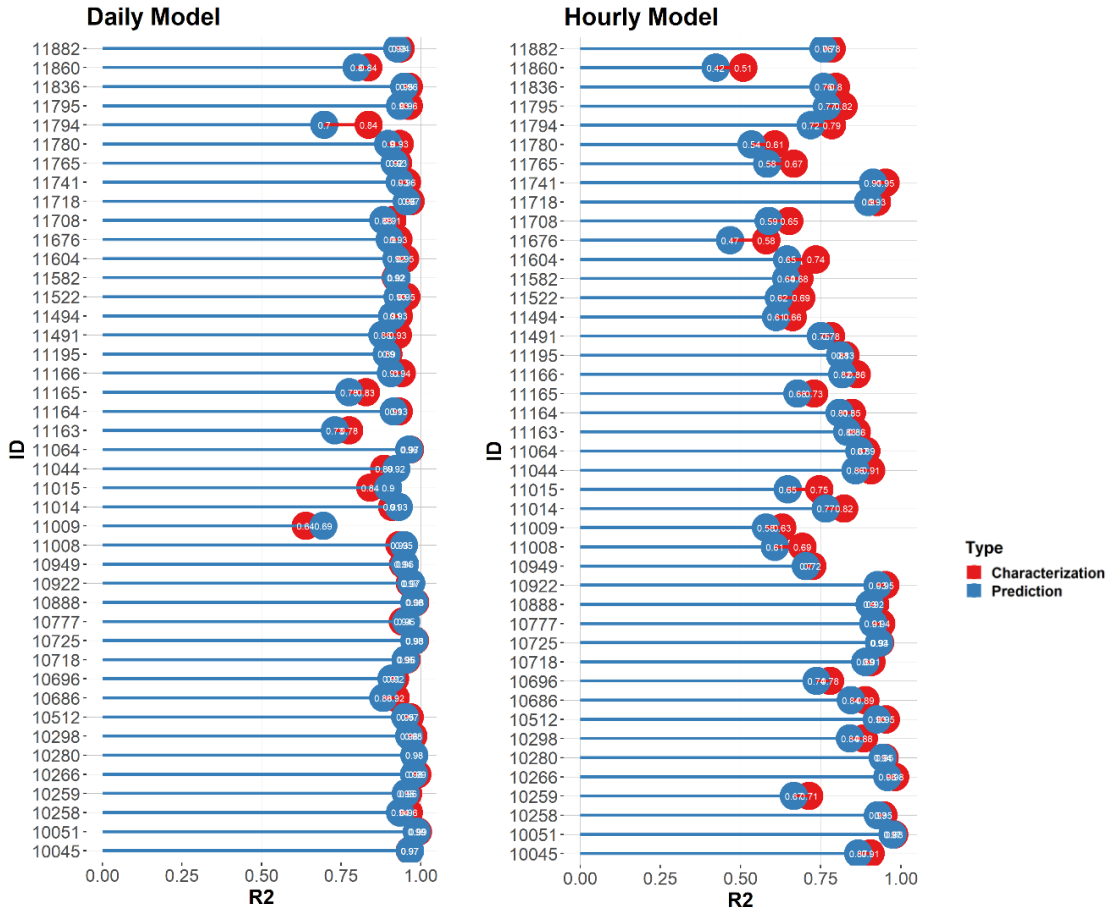
4.2. Q-T Model for Heat-Load Characterization

After the application of DBSCAN algorithm for identification and removal of anomalies, it is time for the analysis and evaluation of the characterization results obtained by the proposed Q-T Algorithm.

4.2.1. General Results

Before starting with the discussion of the results, it has to be remarked that, even if the basis of the model is the same, the results obtained for daily and hourly data will be separately shown and discussed. Besides, it is important to clarify that, when the model is applied to the training data (odd days) again, the results measure the accuracy of the model to characterize the heat load of the building. If the model is applied to testing data (even days), the results measure the accuracy to predict the heat-load.

First, in order to evaluate the accuracy of the model, Fig. V-11 presents the R^2 values obtained from the application of the model to daily and hourly data. In these plots, both results for characterization and prediction of heat loads are shown. Whereas Fig. V-11-a (left figure) presents the results for daily data, Fig. V-11-b shows the results of the application of the Q-T algorithm to hourly data. Note that when Q-T algorithm is applied to daily data only LVL1 and LVL2 of the decision-trees are used.



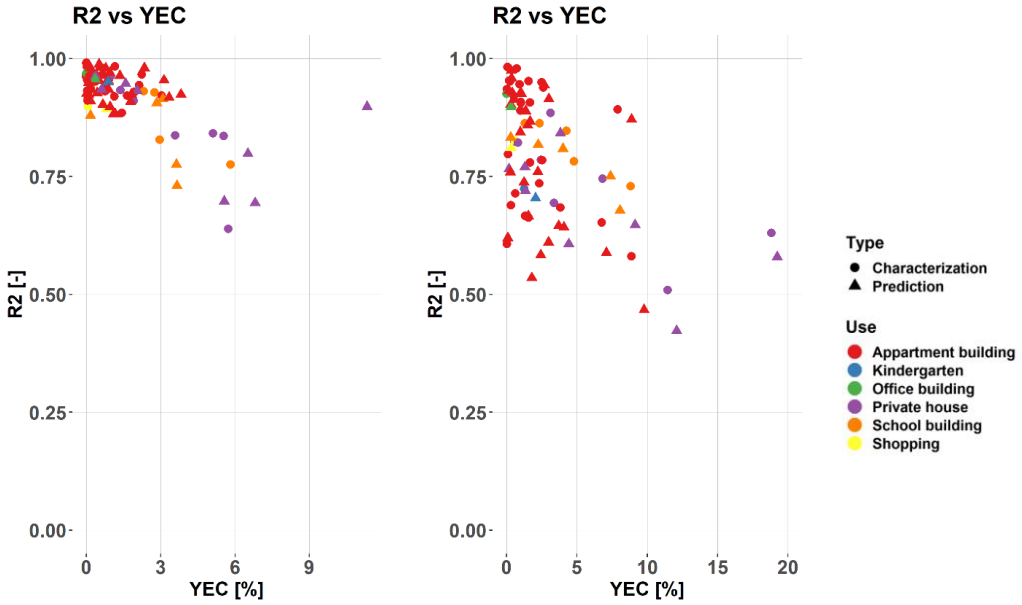
(a) (b)

Fig. V-11. R^2 Values in all the cases from (a) daily model and (b) hourly model

For a daily resolution, the model yields an excellent fit to the monitored data: the minimum value for the R^2 among the studied 43 buildings is 0.69, with the maximum value very close to one (R^2= 0.99). The daily aggregation filters out the hardly predictable intra-daily variations, thereby reducing the inherent uncertainty of demand prediction. In general, R^2 values in characterization of the heat load are higher than the ones for prediction because the data used for tuning the parameters of the model is the one applied for characterization. However, in some of the buildings (e.g., Building 10922 and Building 10949) where the model obtains R^2 values above 0.90, prediction results are even better than those for characterization.

Lower accuracy is obtained for hourly data resolution. The lower correlation and changing variability of the demand patterns of the users in the building reduces the accuracy of the model. Nevertheless, accuracy results with R^2 values above 0.60 are obtained for around 90% of the buildings. The minimum R^2 value (R^2 = 0.47) is obtained in Building 11676 (residential building) and the maximum R^2 (R^2 = 0.97) is reached in Building 10051 (residential building). As it occurs in daily data, the prediction accuracy results to be lower than characterization.

Some of the biggest deviations between model estimations (prediction) and monitored data (real data) correspond to buildings with private dwellings (e.g., Buildings 11795, 11009 & 11860). From the authors´ belief and experience, the implementation of statistical models on this type of buildings can be challenging, especially if they feature manual heat switching systems with an intermittent usage. These activities are hardly predictable for a data interval as low as one hour. For this purpose, Fig. V-12 presents the correlation between R^2 and YEC (defined in Section 3.4). The correlation between a purely statistic variable (R^2) and the variable including energy management is observed (YEC), classified by the final use of the building. This figure is divided into results for daily data (Fig. V-12a) and hourly data (Fig. V-12b).



(a) (b)

Fig. V-12. R^2 vs YEC classified by type of building for (a) daily model and (b) hourly model.

As illustrated in Fig. V-12, a slightly negative correlation is observed between R^2 and YEC values. Thus, lower R^2 values mean that the yearly energy predicted to be used in the building deviates more from the real energy use. This figure confirms that buildings used as private houses (purple) present the lowest accuracy results, both for daily and hourly data. It is remarkable that some of the buildings with relatively low R^2 values show almost no error for YEC. This means that despite that the prediction deviations throughout the year are offset by each other, reaching a perfect result for the annual energy demand (YEC = 0 %) at the end of the year is possible.

As the buildings are connected to a DH network, the proposed model can be used to improve the control of the heat production system in the network. The modelling of individual buildings´ demand enables the characterization of heat load patterns in each dwelling. This methodology provides an individual demand characterization and the demand of the whole district or specific branches could be obtained by the aggregation of the relevant buildings´ demand. Thus, one of the most important advantages of this methodology is that the demand of the network can be adjusted if one building is

disconnected from the network or if a new building is connected to the heating grid. Therefore, the heat production can be continuously optimized by matching the production to the predicted demand.

4.2.2. Heat-Load Characterization. Individual Buildings

A deeper focus has been placed on four buildings. These buildings have been selected for a deeper analysis because they cover a range of different heat load profiles, as requirements for their associated building uses are completely different. In this sense, residential apartments (also referred as private house), offices, educational buildings and commercial buildings are included. The following table (Table V-1) shows the R^2 values obtained in the following four buildings: Building 10051 (residential building), Building 10949 (kindergarten), Building 11164 (school) and Building 11718 (offices).

Table V-1. R^2 values for the buildings selected for a deeper analysis.

	DAILY MODEL		HOURLY MODEL	
	Characterization	**Prediction**	**Characterization**	**Prediction**
Building 10051	0.99	0.99	0.98	0.97
Building 10949	0.95	0.95	0.72	0.70
Building 11164	0.96	0.92	0.85	0.81
Building 11718	0.91	0.96	0.93	0.90

In Fig. V-13, the hourly heat loads of these buildings are presented, comparing the monitored heat loads (black points) to the model estimations (red points). The mentioned Fig. V-13 presents a plot of the heat load against the outdoor temperature for the selected buildings, while Fig. V-14 shows a monotonic plot of their heat loads. In Fig. V-14, the quartiles (0%, 25%, 50%, 75% and 100% percentiles) of the demand are also included as vertical blue lines. The plotted monotonic functions represent the ordered hourly heat profile from maximum (peak) to minimum load. These are valuable for DH operators as they portray a good overview of the heat demand patterns of a building, such as maximum peak load, number of hours at peak load, number of hours

at summer demand pattern, etc. They convey the most important variables for managing and controlling heat production in the district by means of the different heat production plants along the network.

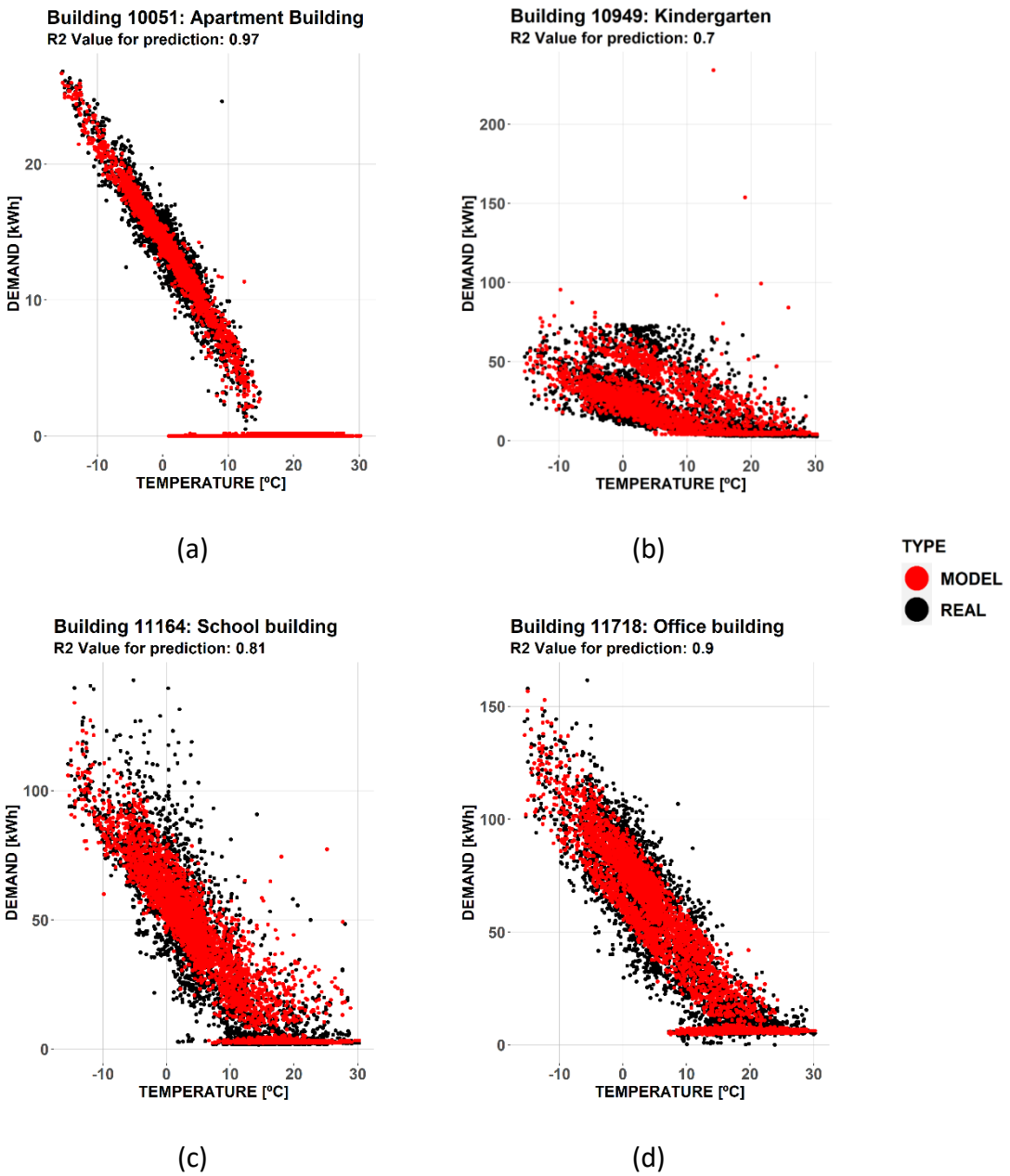


Fig. V-13. Hourly heat load vs outdoor temperature for (a) Building 10051, (b) Building 10949, (c) Building 11164 and (d) Building 11718.

From Fig. V-13 it is concluded that the model fits the general shape of the real data in the four buildings, with a minimum R^2 value of 0.85 in the school and a maximum R^2

value of 0.97 in Building 10051 (Fig. V-13a). A low scattering of the demand points in Building 10051 facilitates the gathering of very accurate results when applying the model to predict the heating demand. The high scattering of the demand in Building 11164 (Fig. V-13c) results in a lower R^2 value, probably caused by the greater variation of the set-point in the heating system due to the larger size of the building under study.

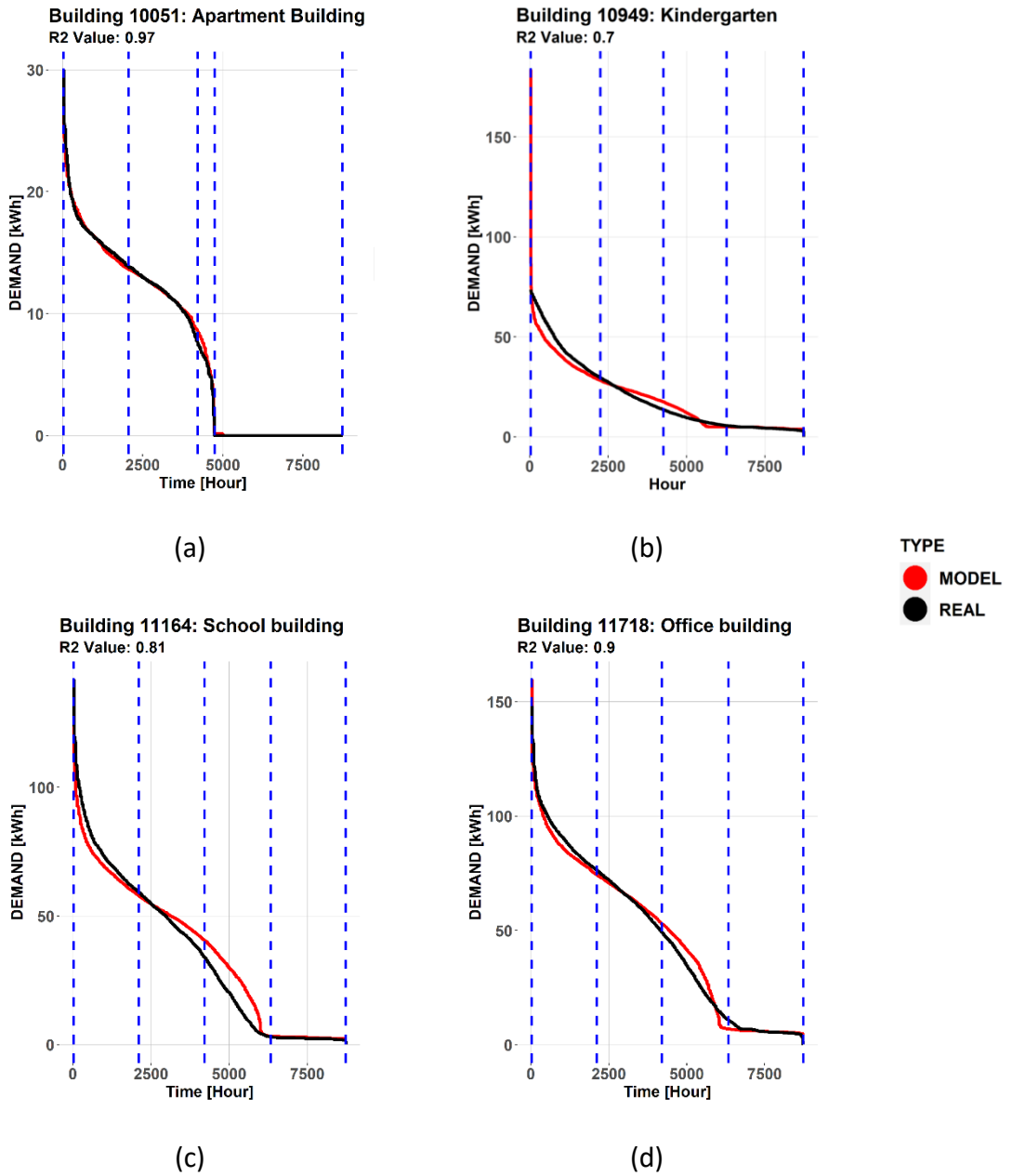


Fig. V-14. Monotonic function of Building 10051 (a), Building 11164, Building 10949 (c) and Building 11718 (d)

The monotonic function of the hourly heating demand shown in Fig. V-14 presents the general trend of the prediction profile from the model. Note that hour zero corresponds with the 00:00AM of 1st January in 2019. In Building 10051 (Fig. V-14a), the demand difference between both lines representing the real demand profile and the result from the model is always lower than 1% of the peak demand. However, in the other two buildings under study, similar results are obtained. In peak demand (first blue line starting from the left, 100% quartile) moments, the difference between the demand from the model and the real data is very low. At high demand moments up to the 3rd quartile, the model slightly underestimates the demand, as can be observed when the red line is below the black line in Fig. V-14b and Fig. V-14c. The inflection point in both cases is located in the hour 2500, after which the model slightly overestimates the real demand. Lastly, in the summer period, the model again fits the real demand.

An additional variable for measuring the accuracy of the model for the energy management of the DH network is the total yearly aggregated demand estimated for each building. The sum of the estimations of each building would anticipate the total energy required to be produced and distributed by the network. Due to the large thermal inertia within the network, the variation in hourly demand could be compensated with heat storage. However, the annual heat production requirement is a key variable for avoiding the overuse of resources to produce heat for the network. Table V-2 shows the total annual delivered heat monitored and estimated for each of the three buildings considered for the analysis.

Table V-2. Yearly demand in GWh for real data and results from the model.

	TRAINING DATA				TESTING DATA			
	REAL DATA		MODEL		REAL DATA		MODEL	
	GWh/Year	YEC	GWh/Year	YEC	GWh/Year	YEC	GWh/Year	YEC
Building 10051	31.70	0	31.93	0.71	31.13	0	31.23	0.32
Building 10949	80.38	0	79.36	1.27	79.50	0	77.87	2.05
Building 11165	150.55	0	156.96	4.25	148.52	0	154.48	4.01
Building 11718	204.02	0	204.01	0.01	201.00	0	201.61	0.30

Small variations between the real demand data and demand resulting from the model are observed. Table V-2 presents the yearly energy demand divided into training and testing data. The relative error of the real data is 0%. The total heat demand error remains below 5% of its real value and, in both Building 10051 & Building 11718, the error is near the top zero. Moreover, apart from one case (Building 11718 and training data), the rest always show a positive relative error; in other words, the model estimates a slightly higher demand than the real one, which ensures the comfort conditions in the buildings.

On the whole, the proposed model appears to be viable for both daily and hourly heat demand, considering the ease of application and the good accuracy of the estimations for most of the buildings. The application of this type of data-driven models in the operation and management of DH networks would be useful to reduce primary energy demand, as well as to achieve a more efficient operation within the flexibility allowed by the network.

5. Discussion & Conclusions

In this chapter, a data-driven model for the characterization and prediction of heating loads in buildings connected to a DH network has been presented. In a preliminary analysis of these heat loads time dependencies related to time-varying demand patterns

were found, as well as transient effects with a great effect on the instantaneous value of the heat demand. These time dependencies have been captured using decision trees with three levels, thus maintaining the simplicity and stationarity of the model. This supervised clustering method allows the implicit consideration of transient effects without the need for an explicit formulation of the thermal inertia in the model and allows to characterize the effects of users´ behavior.

The main objective of the chapter is the development of a simple model that can be deployed over a large set of buildings. This implies that the model needs to be generally applicable to any building, regardless of its usage pattern or construction characteristics. For this reason, no prior knowledge of the building has been incorporated into the model. Model inputs are limited to weather variables and calendar information, with hourly or daily heating demand being obtained as a prediction output.

The following conclusions can be drawn from the study:

- The part of the heat demand corresponding to SH is weather and time dependent, while demand for DHW is solely dependent on the heat demand patterns of the building. Supervised clustering enables the incorporation of this time-dependent demand patterns into the model.
- When the presented model is applied to hourly data for a full year, the results show good agreement with metered data in predicting yearly and daily heat load profiles. Therefore, the developed model is suitable for applications that require to analyze the long-term energy performance of buildings, such as measurement and verification processes.
- Weekly patterns are affected by occupancy schedules, mostly due to the weekday-weekend cycle. Generally, lower heat loads are found when the building remains unoccupied, with peak demands on the initial day of the week (presumably due to thermal inertia).
- Intra-daily patterns are also related to occupancy schedules, mostly business and leisure hours. However, additional variations have been found in heating

patterns due to night setbacks. The application of the model with hourly data is found to be significantly impacted by the dynamics of the building and manual heat switching systems.

- Statistically, the model obtains more accurate results in the prediction process for daily data resolution than for an hourly resolution. This can be attributed to the uncertainty of intra-daily demand patterns. Heat demand data of daily resolution presents less variability and deviation between demand points, which eases the modulation of the loads the high values obtained for daily R^2 in most of the buildings would make the deployment of the model viable for a larger building set.
- From a DH operator perspective, the hourly R^2 is not the most determining variable since the high thermal inertia of a DH network can assume the energy difference between production and demand in a short period of time. The model shows a good performance in predicting the total yearly aggregated heat demand in each of the buildings, with a maximum deviation of around 15% for the worst-fitted building.
- The data-driven model presented in this study is straightforward to implement and does not require a large computational capacity. The results of the study demonstrate that an accurate hourly heat load prediction is obtained for most of the buildings under study. The availability of such estimations for a range of different buildings in a DH network could enable the optimization of the resources for heat generation, deriving in both primary energy and economic savings.

6. Referred Appendix

The content of this chapter has been published as an article in the ENERGY journal by ELSEVIER. The literature details (title and DOI) and the first page of this article can be found in the Chapter XI: Appendix.

Chapter VI

Demand Pattern Recognition

Abstract

This chapter aims to identify energy demand patterns among the raw dataset of the group of buildings in the district-heating network in Tartu. While the previous chapter uses calendar variables to identify demand patterns, in this chapter unsupervised machine-learning algorithms will be applied only using energy demand profiles. This analysis will examine the optimal pre-processing actions to raw data and will determine the optimal unsupervised clustering algorithm using validation indexes. The study also presents individual patterns identified in four particular buildings.

Resumen

Este capítulo tiene como objetivo identificar los patrones de la demanda de energía en el conjunto de datos sin procesar del grupo de edificios conectados a la red de calefacción urbana en Tartu (Estonia). Mientras que el capítulo anterior ha utilizado variables de calendario para identificar patrones de demanda, en este capítulo se aplicarán algoritmos de aprendizaje automático no supervisado solo usando perfiles de demanda de energía. Este análisis examinará las acciones óptimas de preprocesamiento de los datos y determinará el algoritmo óptimo de agrupamiento no supervisado utilizando índices de validación. El estudio también presenta los patrones específicos identificados en cuatro edificios particulares.

Chapter VI Demand Pattern Recognition

1. Introduction

Energy demand patterns in buildings are daily loads or a fraction of the daily demand profiles that are repeated over time [57]. These energy demand patterns may be caused by a repetitive demand action by the users inside the building or by energy management strategies by the DH operator (in case the building is fed by a DH network) and they may be repeated over different days within a heating season. A correct understanding of the energy demand patterns and its causes will help in the characterization process of the heating demand in the buildings [58]. Moreover, the repetitive nature of these patterns could be used as an input variable for advanced models for the prediction of heat demand.

Unsupervised learning algorithms have been successfully applied to identifying usage patterns commonly used in electricity load analysis [59]–[62]; however, their use in heat-related applications has been limited so far. Amongst the existing references for electricity loads, Liu et al. (2021) studied the daily electricity usage pattern of three office buildings with a combination of unsupervised and supervised clustering techniques [96] and they also developed an application for anomaly detection. Carmo et al. (2016) clustered the electricity profile of the distributed heat pumps´ demand located in more than one hundred buildings in Denmark [47]. Two clusters representing weekend and weekdays were identified. A Demand-Response program is proposed by [97] based on electricity demand patterns identified in the electricity demand of a residential building, while Haben et al. (2016) presented a feature-based clustering method in which the computational costs of the algorithm were reduced by using representative variables of the raw dataset [98].

Even though some clustering works are applied over thermal energy, most of these studies are focused identifying electric energy demand patterns. This is mainly caused

by the fact that smart meters for electricity demand have been installed for a longer time than smart meters for thermal loads. Furthermore, the identified energy profiles and their approach to the real causes have hardly been discussed to date. However, the impact of external variables such as climatic variables or seasonal patterns is even more important in thermal loads than in electricity. Additionally, pre-processing activities are a key factor when using unsupervised algorithms. A wrong pre-processing of the original data could lead to inaccurate results, even though the methodology and algorithms used are the optimal ones.

Thus, there is a gap in current literature since there is no other studies in which different clustering algorithms are analyzed and applied to heating energy demand.

2. Objectives of this Chapter

The main objective of this chapter is to explore the use of unsupervised learning for the mining of heat-load patterns in the heating demand of buildings connected to a DH network.

The secondary objectives of this chapter are the following:

- Identification of heat demand patterns in buildings connected to the DH in Tartu and developing a general framework for this study.
    - Identification of optimal clustering method for pattern recognition.
    - Evaluation of several Clustering Validation Indexes or CVIs.
    - Data analysis, by means of different normalization processes and evaluation of the formed clusters with different pre-processing activities.
- Analysis of the identified clusters and heat demand patterns, identifying similarities and possible synergies between the buildings under study and their final use.

3. Methodology

This section outlines the general methodology followed in this chapter. In order to achieve the objectives listed in the previous paragraph, this chapter proposes a multistep method that is illustrated in Fig. VI-1.

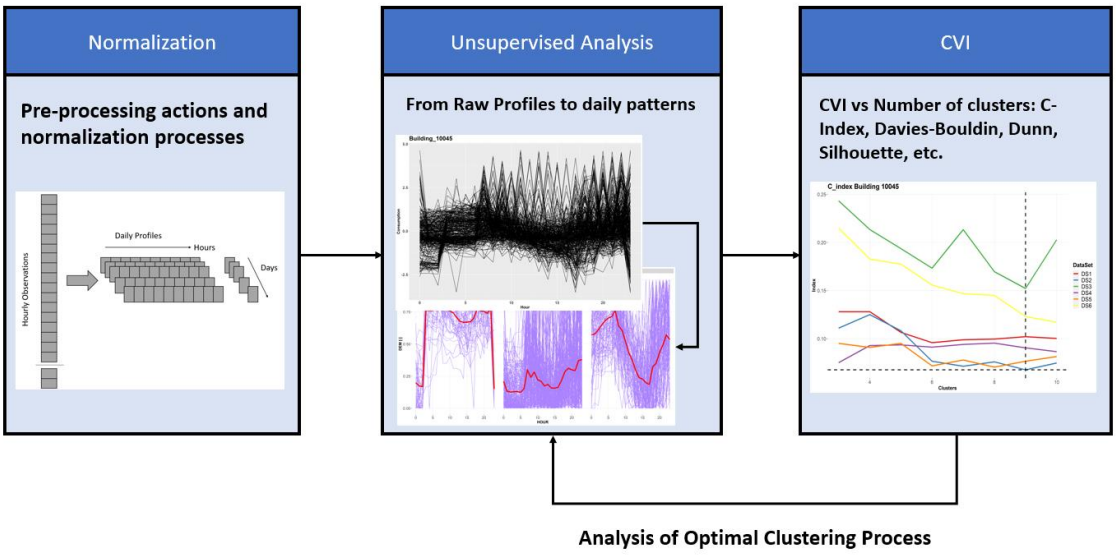


Fig. VI-1. General Methodology followed in Chapter VI

First, Section 3.1 analyzes the pre-processing activities carried out before the unsupervised study. Then, Section 3.2 describes the algorithms used for this purpose and finally in Section 3.3, the metrics used for the evaluation of the clustering process are presented. Additionally, Section 4 will show the results obtained and, as it was done in Chapter V, the results are divided into general results and special and deeper focus on some of the buildings in the DH in Tartu. Finally, Section 5 will summarize the most relevant conclusions and will present the next steps in the method.

3.1. Data-Preprocessing. Data Normalization

For this chapter also applies the outlier removal method presented in the previous chapter and the current dataset starts from the clean dataset after the application of DBSCAN algorithm.

Heat loads are known to vary in time. Lumbreras et al. (2022) and Chapter V showed there are two types of heat-load variations [2]:

- Intra-daily variations, where different load levels occur for each moment in time within the day. These variations might be caused by such variables and factors as climate, occupancy schedules, activation of thermostats and building management systems, as well as the transient response of buildings to the aforementioned issues.
- Inter-daily variations, where the variations are mainly associated to changes in how the building is used (i.e., public holidays).

Within this work, heat load profiles are considered as 1-day long datasets. Each profile contains the variation of the heat load along the day. Considering the 1-h resolution in the data, arrays of 24 values are generated. In each building, profiles for all individual days are generated. Days with data gaps and/or outliers are discarded.

The other main key data process for the application of efficient clustering process is the normalization of the energy daily profiles from the buildings. So, firstly the hourly observations for heat demand are re-order to daily vectors, corresponding each of the vector with the hourly observation of one day, as it is observed in Fig. VI-2. Thus, a vector $X_T = \{X_0, X_2 \ldots X_{23}\}$ containing hourly observations are obtained, corresponding each of the elements of the vectors with the parameters to be clustered.

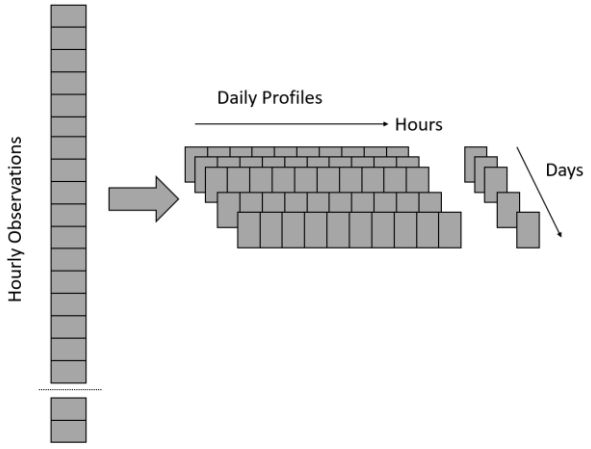


Fig. VI-2. Ordering Hourly observations to daily profiles

The normalization of the daily heat demand profiles is carried out for two reasons. First, the clustering process that is explained in the following section is optimized in terms of computational cost and second, and the most important, is that the main objective of this study is the recognition of the patterns of use of the energy demand. Thus, when dealing with pattern recognition, the absolute value of the load is not considered as relevant as its variation throughout the day. All the values of the energy demand are ranged between 0 and 1 (except the normalization process using Eq. 8)). For this normalization process, three different equations are proposed to identify the best pre-processing conditions for each type of data.

$$q_{norm1_t} = \frac{q_t - q_{min\,t}}{q_{max\,t} - q_{min\,t}} \quad \text{Eq. (6)}$$

$$q_{norm2_t} = \frac{q_t}{q_{max\,t}} \quad \text{Eq. (7)}$$

$$q_{norm3_t} = \frac{q_t - \bar{q}}{sd_{q_t}} \quad \text{Eq. (8)}$$

Where; q_t is the hourly value for heat demand, $q_{max\,t}$ and $q_{min\,t}$ are the maximum and minimum daily demand respectively. $\bar{q}$ is the daily mean value of the heat demand and finally, sd_{q_t} is the standard deviation of the demand profile.

The effectiveness of the clustering process is completely dependent on the normalization process applied to the original data. Fig. VI-3 shows the original data and the normalized profiles for one of the buildings under study (Building 10045) for the three normalized equations abovementioned. As it is observed, the normalized profiles generated with each of the equations are notably different and can affect the clustering process and the accuracy of the patterns that may be identified.

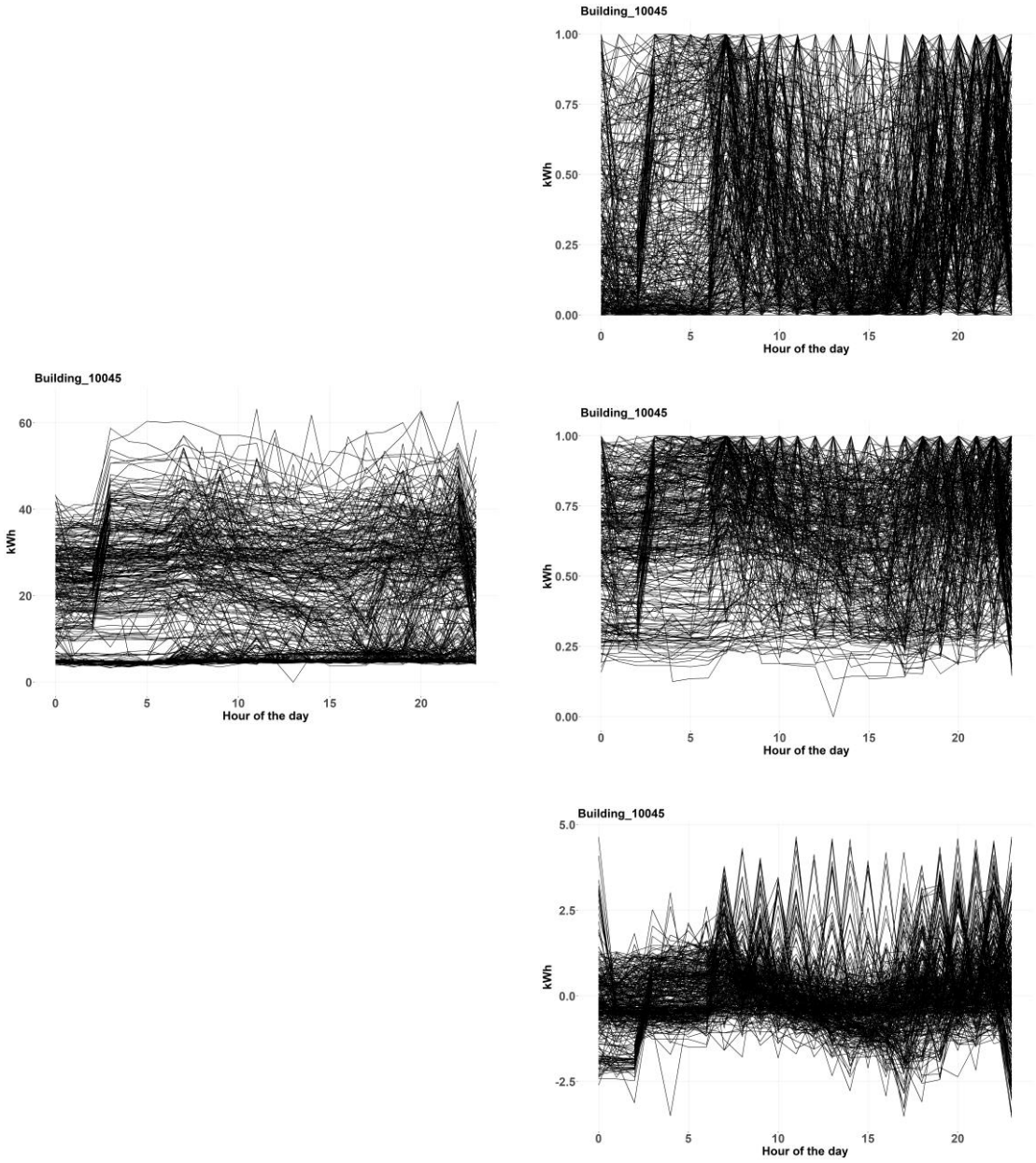


Fig. VI-3. Real Data (left) and the normalized data using Eq. (6), Eq. (7) and Eq. (8) on the right from the top to bottom, respectively.

The generation of the datasets (DS in Table VI-1) for the next steps is a combination of the different pre-processing activities proposed. Therefore, these datasets are compared in terms of efficiency levels, to determine what the optimal preprocessing method is for this process. The characteristics of each data set are shown in Table VI-1.

Table VI-1. Generation of the 6 datasets (DS) and their pre-processing actions

	Nomenclature	Outlier Removal	Norm. Eq. 6	Norm. Eq. 7	Norm. Eq. 8
DATA SET 1	DS1	YES	X		
DATA SET 2	DS2	YES		X	
DATA SET 3	DS3	YES			X
DATA SET 4	DS4	NO	X		
DATA SET 5	DS5	NO		X	
DATA SET 6	DS6	NO			X

The generation of these datasets enable to study the efficiency of different pre-processing actions before clustering.

3.2. Studied Clustering Methods

Clustering or the unsupervised classification of unlabeled patterns into groups is one of the most important tasks in data analysis and data mining. The main objective of clustering resides in gaining insights of the data, discovering patterns and information that are currently hidden. The clustering technique, unlike supervised classification and regression, is part of the unsupervised learning techniques and these unsupervised techniques enable to find all kind of unknown patterns from datasets, without the need of having known experience from previous data. This ML technique has been applied into a wide variety of scientific fields including biology, medicine, engineering and computer science. The first clustering algorithms are found in 1950s.

Among clustering algorithms, hard clustering and soft clustering are found. Whereas in hard clustering one observation can only belong to one cluster, in soft clustering each observation is given a probability likelihood to be part of each of the clusters pre-defined in the algorithm. In function of the mathematical approach, the most used clustering techniques are divided into partitioning clustering, hierarchical clustering, density-based

clustering and model-based clustering. Some of the most important and robust clustering algorithms are briefly explained in the following sections.

The following algorithms have been used in this study:

- K-means Algorithm
- Dynamic Time Warping or DWT
- Fuzzy c-means Algorithm

3.2.1. K-MEANS

This algorithm is part of the partitioning clustering algorithms and is one of the most used clustering algorithms, probably due its robustness and flexibility. Some studies ([99]-[100]) show that this algorithm is the most appropriate for the application in clustering electricity profiles, thus, it is very useful also in heat demand profiles.

The so-called K-means clustering [101] is used to partition the dataset into K predefined clusters and since it is a hard clustering algorithm each observation belongs only to one cluster. The algorithm starts with the random selection of K centroids. As the initialization of this algorithm is a random process, it is recommendable to run the algorithm more than once with different initial centroid selection. Then, each observation is assigned to the closest centroid based on the dissimilarity distance between the observation and the centroid. Different distances are used and studied:

- **Euclidean distance** is defined by Eq. (9) and defines the shortest path between two points which corresponds with the straight line between the observations.
- **Manhattan distance** is defined by Eq. (10) and is defined as the absolute difference between coordinates of the observations.
- **Pearson correlation** is a correlation-based clustering and is defined by Eq. (11).
- **Cosine distance** is defined by Eq. (12), and it is usually used for word clustering. However, as it is very useful when the real value is not important, this measure is also tested in this chapter.

For a pair of observations (X, Y) with n features the distances are calculated by the following equation:

$$d_{euclidean} = \sqrt{\sum_{i=1}^{N}(X_i - Y_i)^2} \quad \text{Eq. (9)}$$

$$d_{manhattan} = \sum_{i=1}^{N}|X_i - Y_i| \quad \text{Eq. (10)}$$

$$d_{pearson} = 1 - \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2}} \quad \text{Eq. (11)}$$

$$d_{cosine} = \frac{X \cdot Y}{\|X\| \cdot \|Y\|} \quad \text{Eq. (12)}$$

In K-means algorithm, as a partitioning clustering algorithm, each observation belongs only to one group. This partition starts with a random selection of K centroids. The objective function (J in Eq. 13) that hast to be minimized in this algorithm is the following:

$$J = \sum_{i=1}^{m}\sum_{k=1}^{K} w_{ik} \|x_i - \mu_k\|^2 \quad \text{Eq. (13)}$$

where w corresponds to a relative weight, x is the measurement value (in this case, heat load), and μ is the cluster center.

After the random mapping of the initial centroids, the Euclidean distance between each point and the centroid is calculated to assign the point to its closest cluster center. Then, the centroid is updated with new values and this process is repeated until the centers do not change. Thus, the initially chosen K centroids may vary the clustering results and, consequently, for choosing optimal clustering, the algorithm is applied 50 times with different initial conditions for every K.

There are no initial indications to determine which is the optimal number of clusters to identify the different energy demand patterns in the building. Therefore, the algorithm is applied for K = {3, 4… 10}. K = 2 is skipped in order to avoid weekday/weekend identification. Thus, for a specific building, eight different clustering processes are carried out.

3.2.2. DYNAMIC TIME WARPING

Dynamic Time Warping or DTW is a hierarchical clustering algorithm that is usually used in time series clustering [102]. This algorithm uses DTW distance as dissimilarity function and unlike the Euclidean distance, this metric enables to consider similarity when there is temporal translation between the patterns. This warping of two temporal sequences is represented in Fig. VI-4.

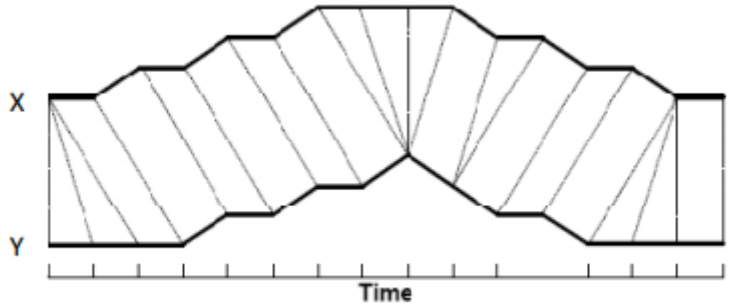


Fig. VI-4 A warping between two temporal signals. Source: [102]

This algorithm has the advantage over k-means if observations are shifted between each other and want to look rather at its shape. DTW calculated the smallest distance between all observations and for the implementation of this algorithm library *dtwclust* [103] has been used in R.

3.2.3. FUZZY C-MEANS CLUSTERING

Finally, Fuzzy C-means (or only C-means) is a partition soft clustering algorithm that was firstly proposed by [104] in 1970s and updated by [105] and later on by [106]. In this soft clustering algorithm, an observation is part of all the resulting clusters with varying degrees of fuzzy membership between 0 and 1. Therefore, the resulting cluster for an observation is the one with highest probability. This algorithm is initialized similarly as

K-means of previous section, starting with the specification of the number of centroids by the user. Similar to the previous case, the study analyses the clusters when using 3 to 10 centroids.

The implementation of this algorithm in R is carried out using library named *ppclust* [107]. A deeper analysis of the algebra behind this algorithm can be found in [106].

3.3. Cluster Validation Indexes

The efficiency of the pattern recognition process of previous section is evaluated by using normalized metrics for all the clustering processes, so that their efficiency can be numerically compared. Thus, the optimal data-set, number of clusters or data normalization process could be concluded. The metrics used for cluster validation are Cluster Validation indexes or CVIs and they can be categorized by three categories [108]:

- **Internal CVIs:** These indexes use the internal information of the clusters to evaluate this classification process and it can used for estimating the number of clusters when there are no initial conditions for the number of clusters.
- **External CVIs:** These indexes use external labelled data to calculate the effectiveness of the clustering process. This external data is considered as the true condition and these indexes are usually used for selecting the optimal clustering algorithm.
- **Relative CVIs:** These indexes evaluate the clustering structure by varying different parameter values for the same algorithm (e.g., varying the number of clusters). It is usually used for determining the optimal number of clusters.

Regarding the nature of the problem of pattern recognition, in which there is no initial conditions (or *true* conditions) to define the optimal number of clusters, **internal CVI**s are the most appropriate for cluster validation.

Each of the indexes analyzed present their own evaluation equation, but in general, these indexes evaluate the inter-cluster (distance between points in the same cluster) and intra-cluster (distance between points from different clusters) distances. Therefore,

a low intra-cluster distance and high inter-cluster distance mean that the identified clusters are separated and compact. In other words, these indexes evaluate the **compactness**, **separation** and **connectivity** of clusters. **Compactness** or **cluster-cohesion** measure how close the objects are within the same clusters (inter-luster) and **separation** refers to the separation between clusters centers and pairwise minimum distances between observations in the same cluster. Finally, **connectivity** corresponds with the measure when items are placed in the same cluster as their nearest neighbors.

Thus, many indexes are found in literature, but some of them are more robust than others and their use is widely applied. These indexes are Silhouette Index [109], [110], Davies-Bouldin [110], Dunn Index [111] or C-Index [112], among others. Even if a deeper focus is carried out for these indexes, a statistical study from more than 40 internal indexes is carried out using the library *ClusterCrit* [113] in R.

All the clustering processes mentioned in the previous section, including different buildings, data normalization processes, outlier removal and clustering algorithms are evaluated using CVIs, comparing the effect of each of the steps of the abovementioned methodology, defining the optimal framework for pattern recognition of heat demand.

Using too few clusters could not be useful to discover the patterns in the building, while using too many clusters could result in insignificant differences across some of the patterns. Therefore, the optimal number of clusters to be analyzed is selected to be from 3 to 10.

4. Results

This section summarizes the results obtained from different clustering processes, comparing normalization, clustering algorithms and number of clusters. First, the results comparing clustering algorithms are shown, followed by the general results for the different datasets proposed in

Table VI-1. Finally, the specific patterns of four of the buildings are presented.

4.1. General Results

4.1.1. Comparison between clustering Algorithms

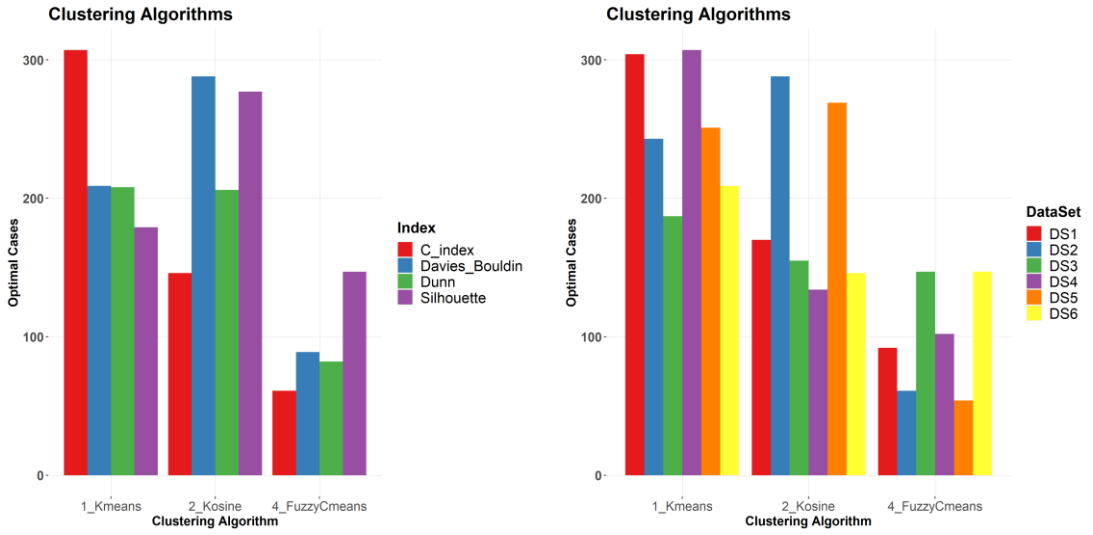
The nature of each clustering algorithm is different and the way they find clusters varies. Consequently, as the clustering algorithm proposed fully determines the efficiency of the methodology and regarding the large number of cases simulated, it is necessary to start with a global analysis for determining the algorithms that best perform according to the Cluster Validation Indexes (CVI) and for the case study that we are studying.

Firstly, and due to the very large convergence time required by the DTW algorithm and especially with a high number of clusters, Dynamic Time Warping algorithm was initially discarded. For a potential application of the method in a real operation of the DH network, the algorithm must converge in a reasonable time and even further for a high number of buildings. The instant and random nature of DHW demand profile makes inefficient the advantages of this algorithm. The most relevant advantage of this algorithm is not valuable for this type of energy profile.

On the other hand, K-means algorithm behaves very similarly with different dissimilarity distances (Euclidean, Manhattan and Pearson correlative distance) and no particular distance performs better than others do. Consequently, hereinafter the results will only be shown for the Euclidean dissimilarity distance. The K-mean variant using Cosine distance is separately shown in the following figures.

Thus, K-means and its variant Cosine distance K-means and along with Fuzzy c-means algorithm are tested for their effectiveness evaluation. For this purpose, all the simulated cases are accumulated in bar plots, where the number of optimal cases is shown for the different clustering algorithms. The abovementioned four CVIs (C_index, Davies-Bouldin, Dunn Index and Silhouette) are used for quantifying the clustering effectiveness and this process is run for the six datasets defined in

Table VI-1. This being so, Fig. VI-5 presents the sum of the number of optimal cases including the 43 buildings in the network for each of the clustering variants commented before.



(a) (b)

Fig. VI-5. Number of Optimal Cases for Different Clustering algorithms. (a) Divided by CVIs and (b) Divided by Datasets

From previous images can be concluded that Fuzzy c-means algorithm (and soft clustering algorithms in general) is not optimal for energy profile clustering. Even though it shows some cases that presents optimal results, it is the algorithm with lower number of cases, so it is discarded as the best clustering algorithm for this application. Besides, K-means using Euclidean distance performs the best followed by K-means with the Cosine distance. Thus, hereinafter, the identification of specific heat demand patterns in buildings is carried out with K-means algorithm and using the Euclidean distance as dissimilarity metric.

4.1.2. General Results. Comparison between Datasets

From previous paragraphs we concluded that among the different algorithms tested, K-means clustering using Euclidean distance is the most appropriate for this case. After analyzing the efficiency of different clustering algorithms, this section presents the results for different datasets (DS) generated and shown in Table VI-1.

The pre-processing of the original-raw data, by means of outlier removal and normalization process, defines the differences between datasets. Thus, K-means with Euclidean dissimilarity distance and Cosine distance are used and validated with the

same CVIs than in the previous section. Same methodology than for the comparison between algorithms is carried out, so K in K-means is varied from 3 to 10 for all the buildings under study and evaluated by 4 different CVIs for the two dissimilarity distances abovementioned.

Fig. VI-6 summarizes the number of cases in which each of the datasets generated results as the optimal dataset.

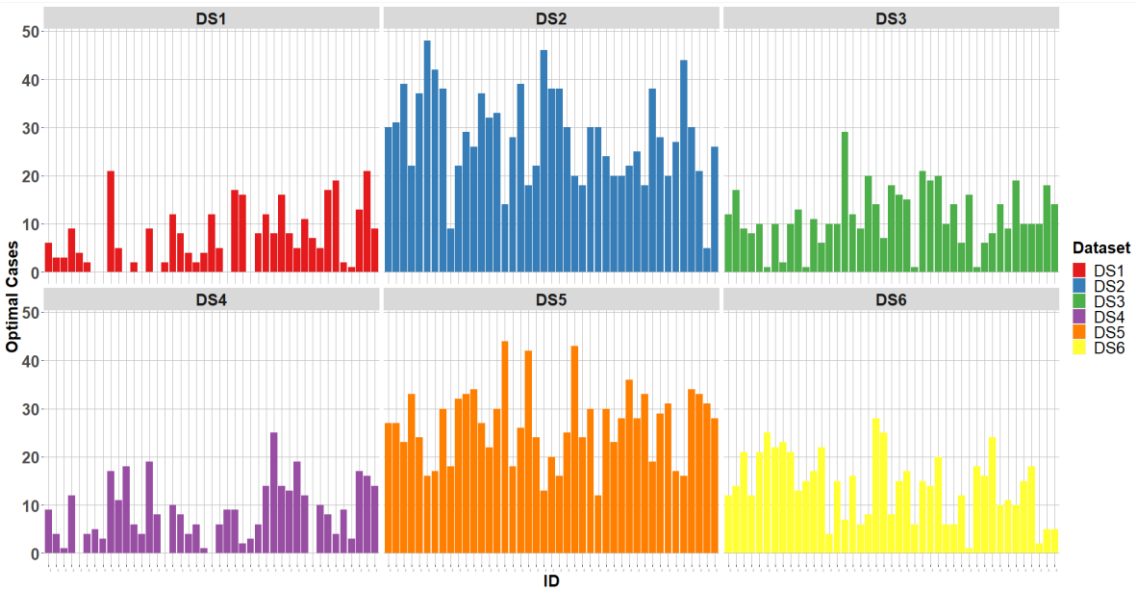


Fig. VI-6. Bar plot of the Number of optimal DS cases divided by buildings.

The ID of the building in Fig. VI-6 are removed for the sake of clarity.

Although there is not a unique dataset which groups all the most compact clustering processes, DS2 and DS5 presents the highest number of optimal cases. These datasets correspond with the normalization process in which the instant hourly heat demand is divided by the maximum daily heat demand value. Between DS2 and DS5, DS2 presents a slightly higher number of optimal cases, probably caused by the outlier removal. Nevertheless, the other datasets also present many positive results, especially DS3 and DS6 in which the normalization process includes more complex variables, such as the standard deviation of the heat demand daily profiles.

From Fig. VI-6 can be concluded that normalization of the heat demand profiles is a vital step for pattern recognition but identifying the optimal normalization equation is completely dependent on the shape of the data. The nature of the profiles (maximum value, minimum value, deviation, etc.) will determine the optimal way of normalizing data. Consequently, normalization process will not be an excluding condition and for each building the dataset with best results will be considered. Moreover, the optimal clustering results may not coincide with the most clarified pattern recognition, and therefore, the different data-sets are studied also in the following chapters. The relation between clusters, heat demand patterns and the external conditions affecting the heat demand will be analyzed in the following chapter (Chapter VIII).

4.2. Individual Buildings Analysis

The same way that it was done in Chapter V, this section will focus on the identification of four individual buildings´ patterns. The final use of the building, and consequently, the occupational behavior in the buildings will completely determine the potential patterns to be recognized. Thus, two residential buildings (Building 10045, with DHW and Building 10051, with no DHW demand), an educational building (Building 10949) and a commercial building (Building 11195) will be presented in the following paragraphs. At this point, we remember that the heating profiles of these buildings can be found in Chapter XI, Appendix.

The study for the heat demand patterns identification is divided into two main parts. In the first part, a statistical analysis of the resulting clusters using more than 30 CVIs (the four CVIs used in the previous paragraph are also included) is presented and followed by a detailed analysis of the four most used CVIs.

4.2.1. Building 10045 (Residential Apartment with DHW demand)

First, we will start with the analysis of the patterns of a residential apartment where the DHW demand is also fed from the DH network. Fig. 11 presents the number of optimal cases that are defined by all the CVIs under study in Building 10045. As it can be observed, and similarly to the algorithm decision study, there is not a unique answer

about which is the optimal clustering number. Moreover, the dataset and its associated normalization process also influence in the identification process of the patterns on that building.

On the one hand, Table VI-2 summarizes the number of CVIs that asserts which of the clustering processes developed is optimal for Building 10045.

Table VI-2. Number of CVIs for optimal clustering process in Building 10045

Nº of Clusters	BUILDING 10045					
	DS1	DS2	DS3	DS4	DS5	DS5
K = 3	15	24	15	28	20	25
K = 4	0	1	0	1	3	2
K = 5	7	1	0	0	0	0
K = 6	2	1	1	0	3	1
K = 7	0	0	1	0	0	0
K = 8	4	1	3	0	3	0
K = 9	1	3	11	0	0	0
K = 10	5	3	2	5	4	6

In general, the K=3 clustering process gather the highest share of optimal cases, especially for DS2, DS4 & DS6 datasets. Four different clusters (K=4) and five different clusters (K=5) also gather a large number of optimal cases for DS5 and DS1, respectively. Finally, for K=10, there are also a significant amount of optimal clustering process for all the datasets studied.

Focusing the study on the most common CVIs, the evolution of these indexes for the different datasets and cluster numbers are shown in Fig. VI-7. The dashed lines correspond with the optimal clustering process and the CVI value for each case. The optimization of C_Index and Davies-Bouldin index is obtained with the minimization of the index and optimization of Dunn Index and Silhouette is obtained with the maximization.

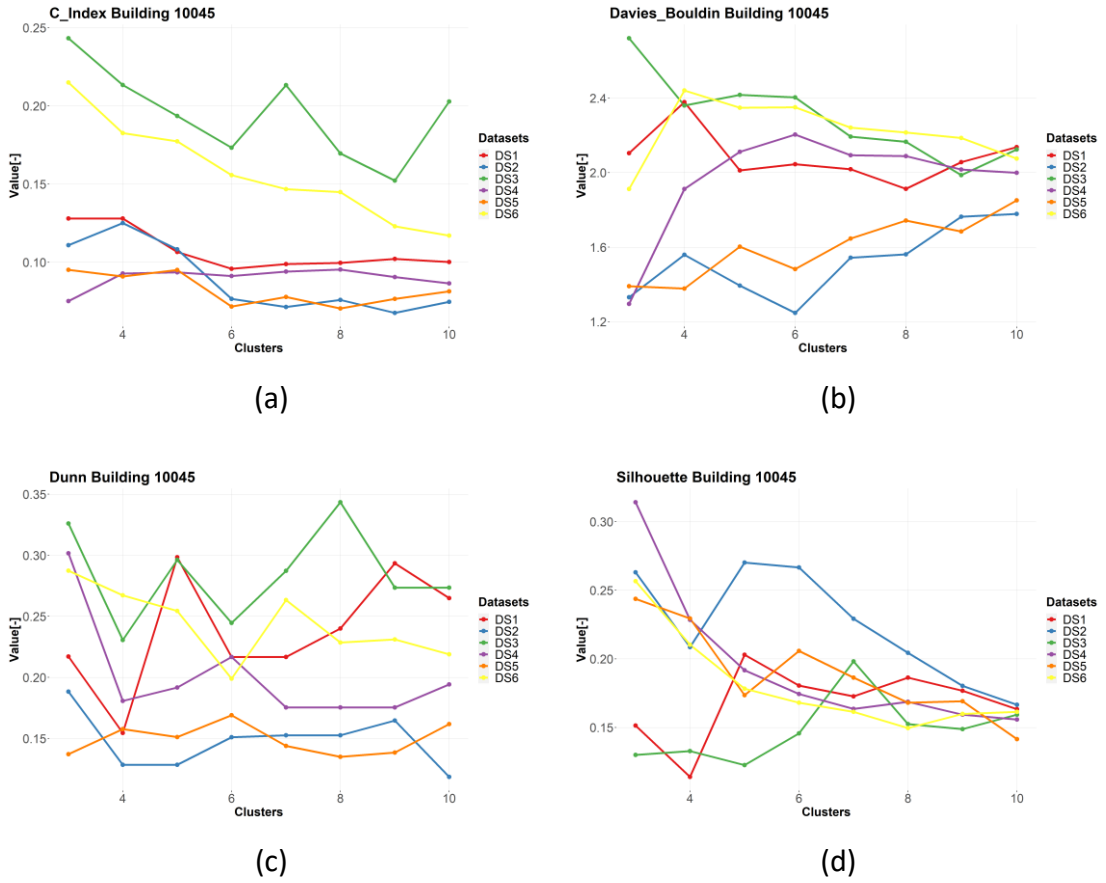
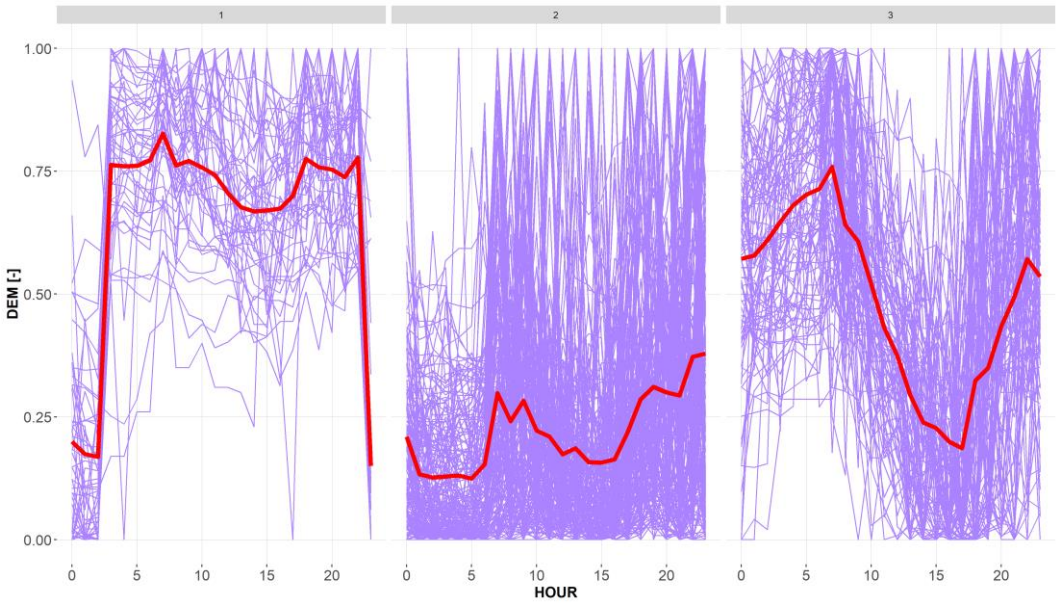
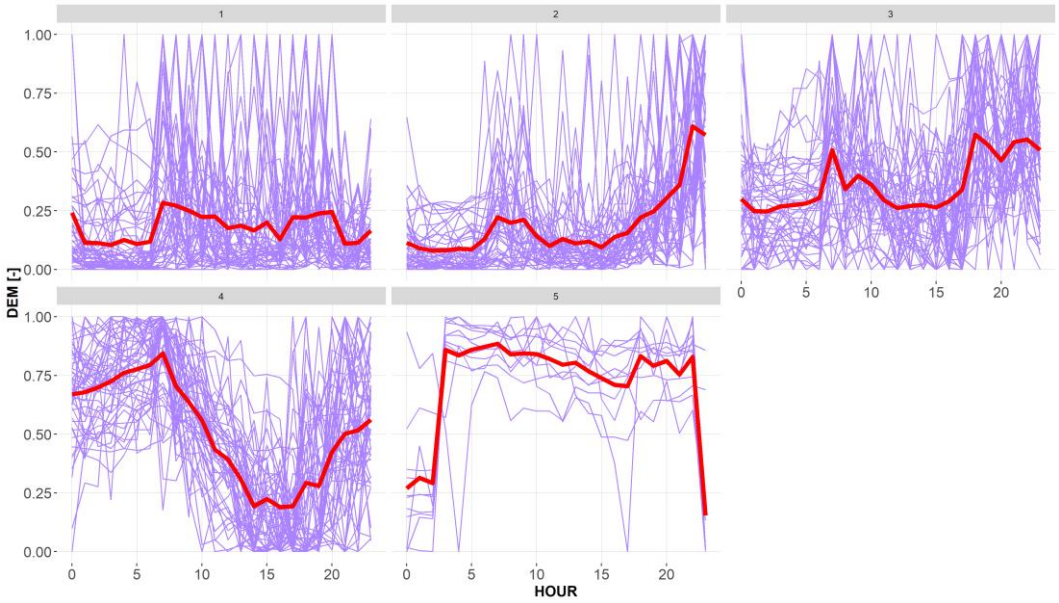


Fig. VI-7. Evolution of (a) C_Index, (b) Davies-Bouldin, (c) Dunn Index and (d) Silhouette Index in Building 10045

Observing the results for the CVI analysis, it seems to be a very complex and chaotic problem to be solved. However, when analyzing the individual clusters of the best cases, similar patterns are recognized. The optimal number of clustering seems to be K=3, for cases with normalized data, as per Eq. 2 (DS2 and DS4). There is also a relatively high number of CVIs that conclude that K=5 is the optimal clustering process, with 7 CVIs with DS1. Consequently, Fig. VI-8 presents the clustering results for these two cases.



(a)



(b)

Fig. VI-8. Daily energy demand clusters for normalized data in Building 10045: a) K=3 with DS4 and b) K=5 with DS1

If the energy demand patterns shown in Fig. VI-8 are visually analyzed, it can be inferred that both clustering approaches are not very different:

- Cluster 1 from DS4/k=3 (Fig. VI-8a) corresponds to Cluster 3 in DS1/k=5 (Fig. VI-8b).
- Cluster 2 from DS4/k=3 and Cluster 5 & Cluster 1 from DS1/k=5 correspond to the same pattern.
- Cluster 3 from DS4/k=3 corresponds to Cluster 2 and Cluster 4 from DS1/k=5.

Therefore, the following patterns were identified, based on the clusters from DS4, K=3:

- In cluster 1, the heat load is heavily increased between 3am to 5am. It remains relatively constant and at very high values from 5am to 11pm. At 11pm, another strong demand variation is identified and the levels of the demand before 3am are maintained. The high demands along the day are caused by the very cold temperatures that Tartu (Estonia) usually presents in winter and requires a constant demand for SH. The strong variations are caused by a night setback induced by the DH operator, in which the set-point temperature is reduced. It is expected that the users of this residential building will be sleeping and there will be no need to maintain the same comfort conditions as at other times. This night setback means that the energy demand differs from its dependency with climatic variables.
- The second cluster shows the most stable profile, grouping days with relatively constant energy demand along the day in the same cluster. A relative peak demand is identified at 7-8am, coinciding with the same peak demand of the other clusters.
- In Cluster 3, the energy demand gradually increases until approximately7-8am, when the peak demand is reached due to the DHW demand in these hours. From 8am onwards, energy demand decreases until 5pm, coinciding with the hours when the users of the building are supposed to be out of the building. After this hour, the demand starts to increase, up to the levels of the first hours of the day.

4.2.2. Building 10051 (Residential Apartment with NO DHW demand)

The same methodology followed in Building 10045 will be followed also in this building. For this building, Table VI-3 presents a summary of the number of optimal clusters for all the cases and evaluated with all the CVIs considered in this study.

Table VI-3. Number of CVIs for optimal clustering process in Building 10051

Nº of Clusters	BUILDING 10051					
	DS1	DS2	DS3	DS4	DS5	DS5
K = 3	15	24	15	28	20	25
K = 4	0	1	0	1	3	2
K = 5	7	1	0	0	0	0
K = 6	2	1	1	0	3	1
K = 7	0	0	1	0	0	0
K = 8	4	1	3	0	3	0
K = 9	1	3	11	0	0	0
K = 10	5	3	2	5	4	6

The number of optimal clustering processes are more concentrated than Building 10045, with a large share of cases in K=3 clusters. Besides, some optimal clustering cases can also be found for six, seven and ten clusters. Regarding datasets, DS2 & DS5 present a slightly higher number of cases, corresponding with the datasets normalized by the maximum value. Focusing on the 4 indexes, Fig. VI-9 illustrates the evolution of these indexes for different clustering processes.

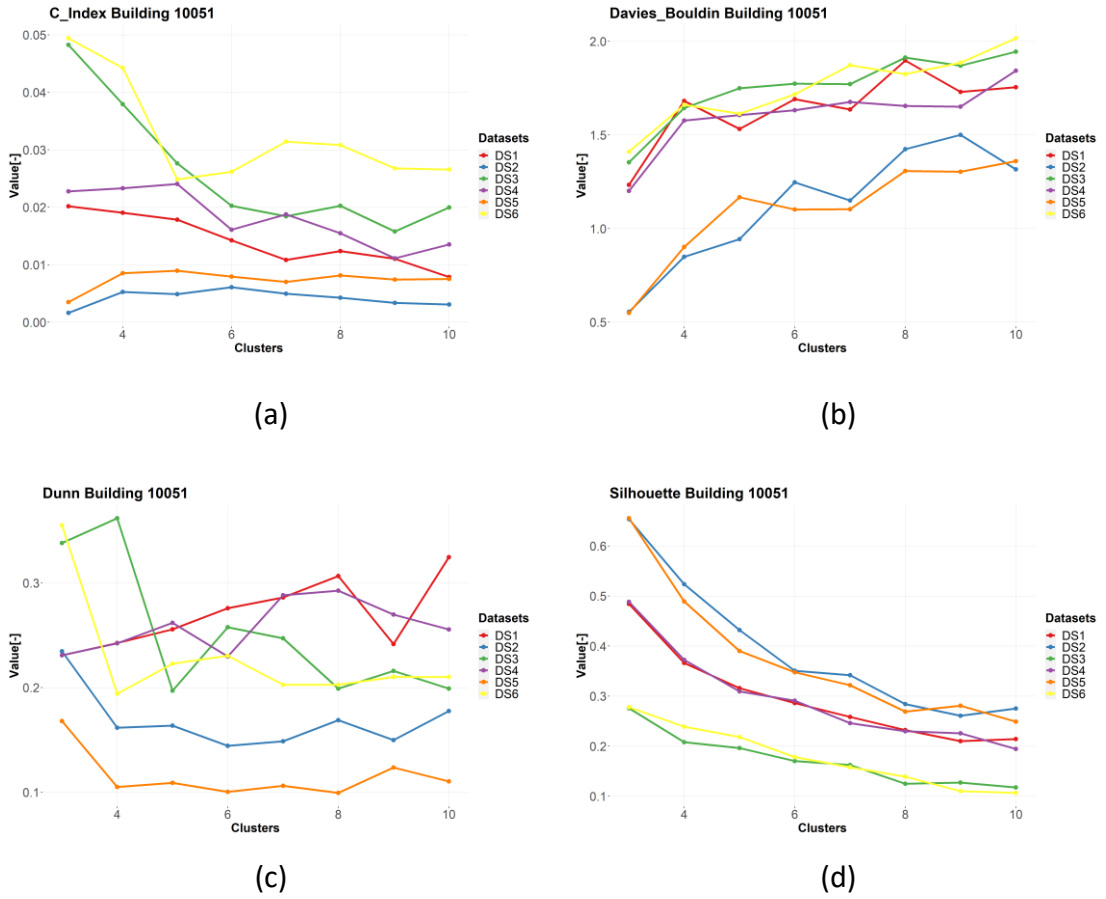
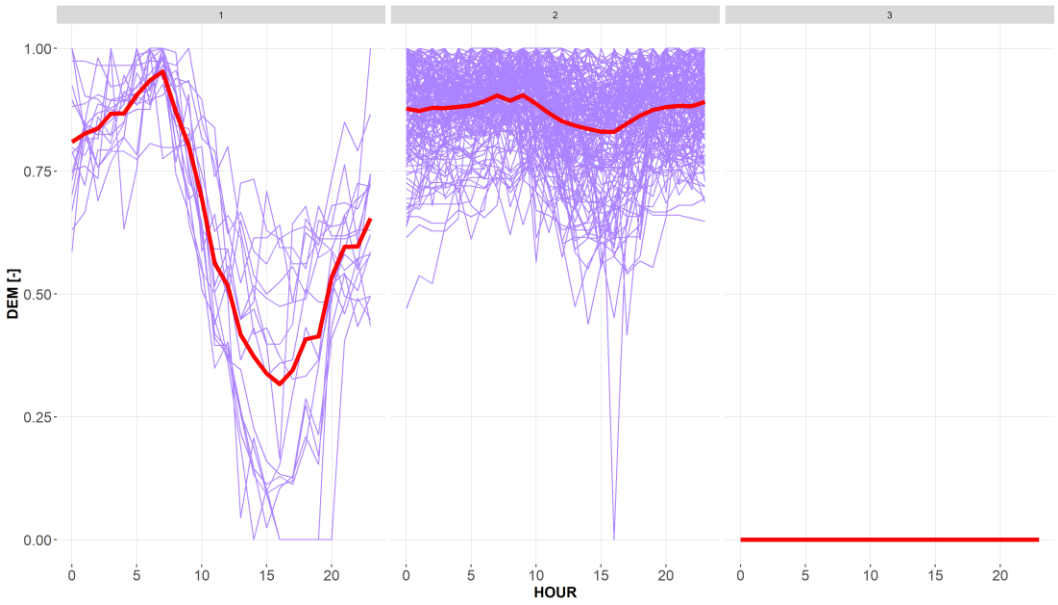
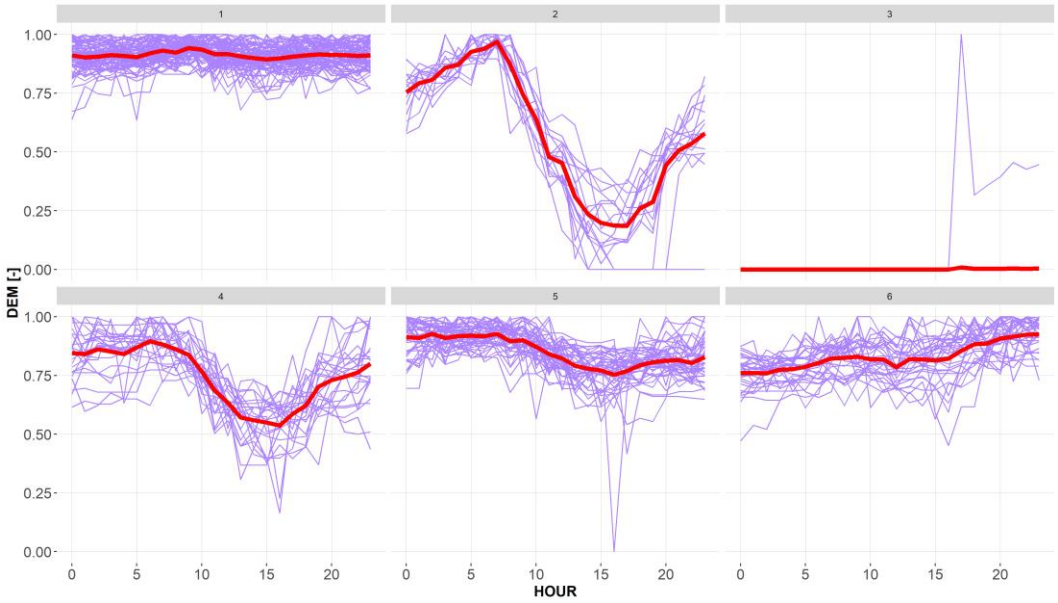


Fig. VI-9. Evolution of (a) C_Index, (b) Davies-Bouldin, (c) Dunn Index and (d) Silhouette Index in Building 10051.

In this building, C_Index, Davies-Bouldin, and Silhouette coincide that K=3 results in the optimal separation between clusters. Moreover, these optimal values are reached with DS2 & DS5. Regarding Dunn Index, the optimal value is obtained with K=4 and DS3. It could be considered an exceptional case because it is the only dataset that increases the clustering effectiveness when K increases from three to four. In order to study the differences between these clustering processes, Fig. VI-10 shows the result for two clustering processes.



(a)



(b)

Fig. VI-10. Daily energy demand clusters for normalized data in Building 0051 (apartments building): a) K=3 with DS4 and b) K=5 with DS1

Observing the separated energy demands profiles provided by the unsupervised clustering, the following heat demand patterns could be identified in Building 10051.

- Pattern 1: The demand in this cluster corresponds with the days in which there is no heat demand. Considering that this building is used as a residential apartment, the DHW may be supplied by an external heat source. Thus, this first pattern corresponds with summer period when there is no eat supply for space-heating purposes. This pattern can be identified in Cluster 3 with K=3 or Cluster 3 with K=6.
- Pattern 2: This energy pattern is similar to Pattern 2 in Building 10045. The energy demand is gradually increases until 7-8am approximately, when the peak demand is reached due to the DHW demand in these hours. From 7am onwards, energy demand decreases until 17pm, coinciding with the hours when the users of the building are supposed to be out of the building. After this hour, the demand starts to increase up to the levels of the first hours of the day. This pattern can be identified in Cluster 1 with K=3 or Cluster 2 & Cluster 4 with K=6.
- Pattern 3: This third pattern is constituted by days in which the heat demand remains relatively constant but with slight increase throughout the day. So, the minimum demand occurs at 0am and from then on, the heat demand increases throughout the hours, reaching the maximum demand value at 23pm. This pattern can be identified in Cluster 4 with K=4 or Cluster 6 with K=6. For K=3 case, this pattern is hidden in Cluster 2
- Pattern 4: This last pattern clusters the days with a constant heat demand all over the day, but contrary to the first pattern, heat is used in the building in these days. Since the normalized profile in some cases is similar to the days of pattern 1, Pattern 4 is hidden in Cluster 4 for K=4. However, it is necessary to distinguish between Pattern 1 and Pattern 4. This pattern can be identified in Cluster 2 with K=4 or Cluster 1 with K=6.

4.2.3. Building 10949 (Kindergarten)

The results from Table VI-4 show a greater concurrence that the optimal clustering process is obtained with K=3, especially in datasets DS3 and DS6. Table VI-4 presents the daily energy profiles obtained from this process with DS3. Similar to the analysis in Building 10045 (residential apartment), these results were compared to results from K=4 and DS4, which obtained 9 CVIs. As expected, different energy demand profile types from those in Building 10045 were found, since this building is used as a kindergarten.

Table VI-4. Number of CVIs for optimal clustering process in Building 10949

Nº of Clusters	BUILDING 10949					
	DS1	DS2	DS3	DS4	DS5	DS5
K = 3	15	17	28	19	22	26
K = 4	1	1	0	9	5	2
K = 5	0	3	0	0	1	0
K = 5	0	4	0	2	0	0
K = 6	0	3	0	0	0	0
K = 8	1	1	2	0	0	0
K = 9	0	0	0	0	1	1
K = 10	16	4	3	3	5	4

In the same way than for the rest of building analyzed, focuses on only 4 usual CVIs and presents the evolution of these indexes for the different clustering algorithms applied.

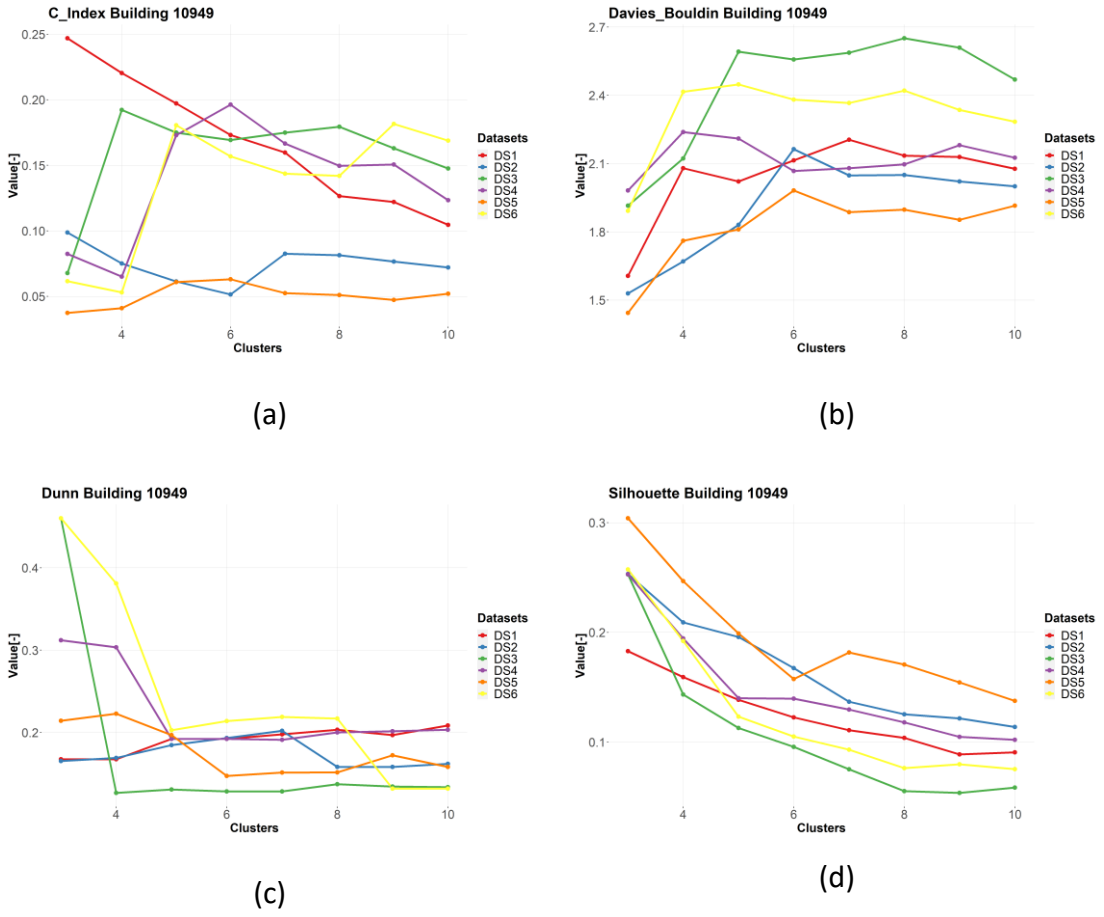
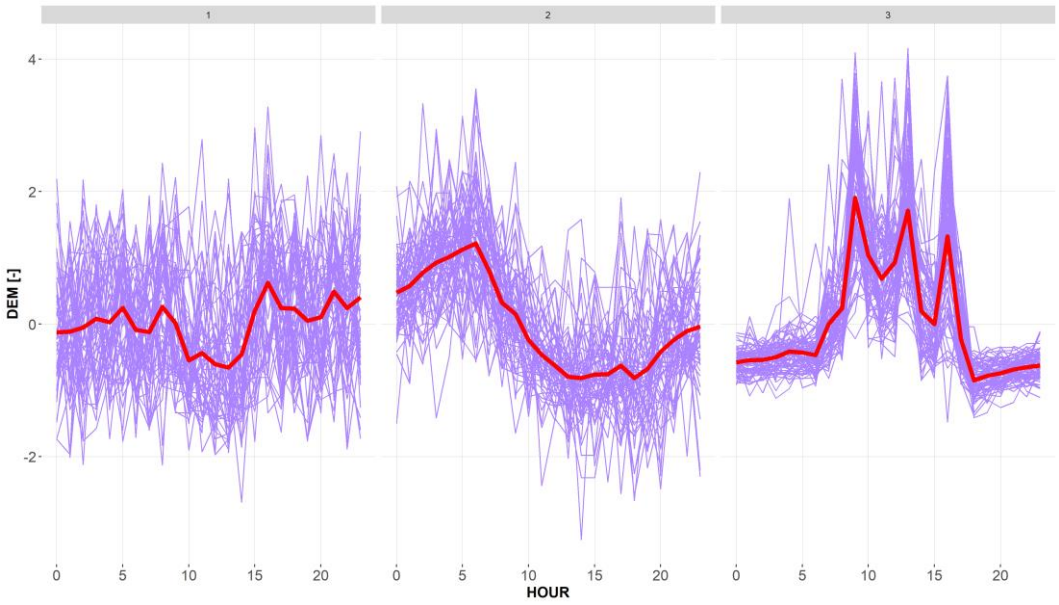
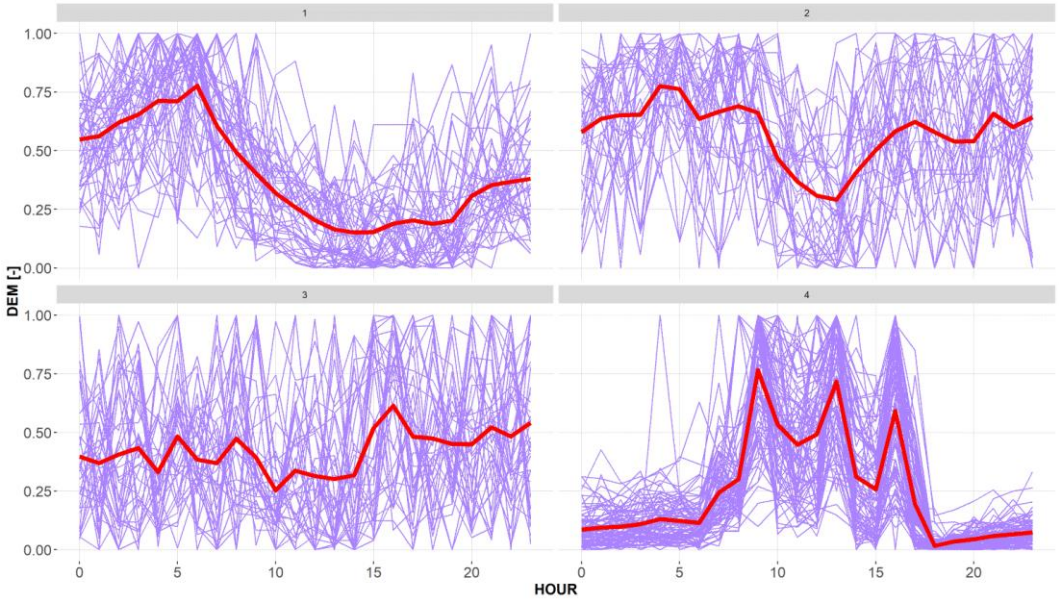


Fig. VI-11. Evolution of (a) C_Index, (b) Davies-Bouldin, (c) Dunn Index and (d) Silhouette Index in Building 10949

In this building, all specific indexes shown in Fig. VI-11 agree that K=3 clustering process is the optimal and regarding the datasets, all the indexes expect Dunn Index show their optimal value with DS5, followed by DS2. Dunn index presents its optimal value with DS3 and DS6. Fig. VI-12 shows the result for these clustering processes.



(a)



(b)

Fig. VI-12. Daily energy demand clusters for normalized data in Building 10949 (kindergarten): (a) K = 3 with DS3 and (b) K=4 with DS4

As occurred in Building 10045, slight differences between profiles could be found in Building 10949:

- Cluster 1 from DS3/k=3 (Fig. 4a) corresponds to Cluster 1 and Cluster 2 in DS4/k=4 (b).
- Cluster 2 from DS3/k=3 and Cluster 3 from DS4/k=4 correspond to the same pattern.
- Cluster 3 from DS3/k=3 corresponds to Cluster 4 from DS4/k=4.

Therefore, based on the energy demand profiles from K=3, the patterns recognized from Fig. VI-12 are the following:

- Cluster 1 presents a quite stable energy demand profile. This cluster groups the days in summer with no demand for SH and the days with very stable profiles with SH demand. A relative minimum demand is found at 12am. The energy profiles in this cluster show a very low correlation with climatic variables.
- Cluster 2 is similar to Cluster 1, but a greater load reduction is observed at approximately 12am, increasing dependency on the climatic variables. At noon and coinciding with the hours with the highest ambient temperature and highest solar irradiance levels, the demand is reduced.
- Cluster 3 shows the profile with greater variability. The energy demand profiles in this cluster remain relatively constant until approximately 7am. This time coincides with a common opening hour of kindergartens, or shortly before, so it is possible to condition the building before the arrival of the occupants. At this time, a steep increase in the demand is observed, reaching the first peak at around 10am. In the next hour, the demand slightly decreases, probably taking advantage of the thermal inertia of the building. Then from 12noon to 1pm (more or less), another increase in the demand is observed, and from 1pm to 3pm the demand increases again. A third maximum demand, in this case a relative maximum, is observed at around 4pm, before the demand starts to decrease to the levels of the first hours of the day. Finally, at around 6pm, the demand reaches a relatively constant value for the rest of the day.

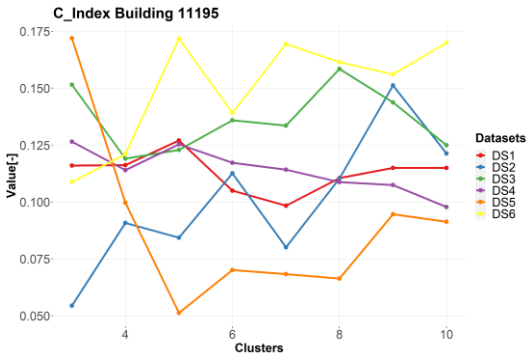
4.2.4. Building 11195 (Commercial/Shopping building)

The most predominant optimal clustering results are obtained with K=3 in all the datasets, but with K=4 (DS3) and K=5 (DS5) also showing good results, as can be observed in Table VI-5.

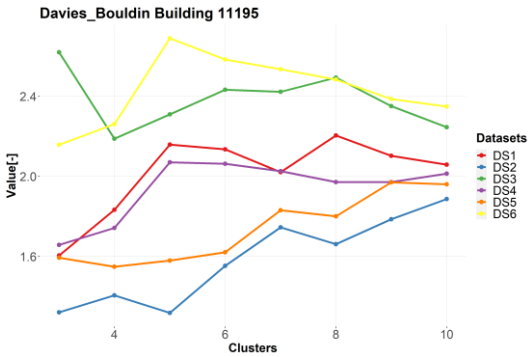
Table VI-5. Number of CVIs for optimal clustering process in Building 11195

Nº of Clusters	BUILDING 11195					
	DS1	DS2	DS3	DS4	DS5	DS5
K = 3	27	27	11	19	2	23
K = 4	0	0	15	5	3	5
K = 5	0	2	1	0	23	0
K = 5	0	0	0	0	1	0
K = 6	2	2	0	0	0	0
K = 8	0	0	1	0	1	0
K = 9	0	1	1	0	0	1
K = 10	5	2	4	10	4	4

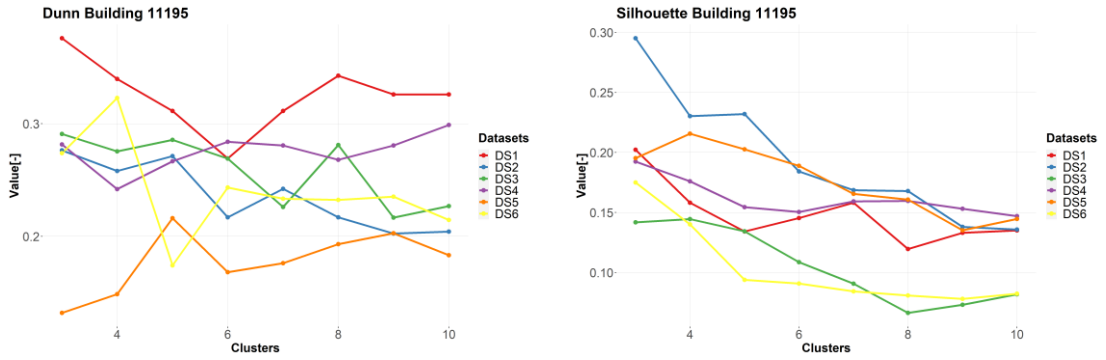
The same way than in the rest of the buildings, Fig. VI-13 presents the evolution of these four CVIs analyzed:



(a)



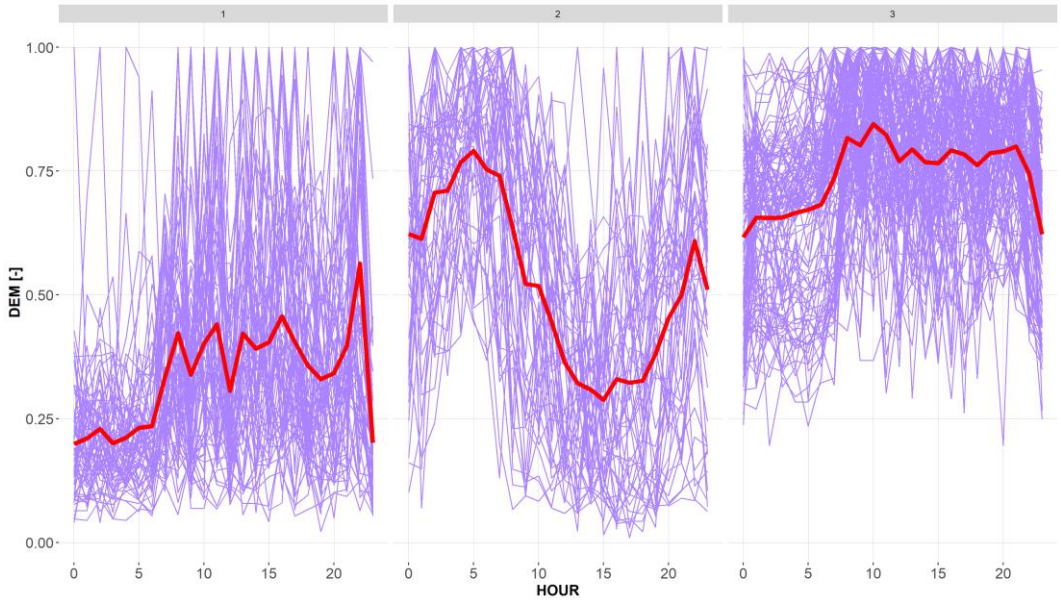
(b)



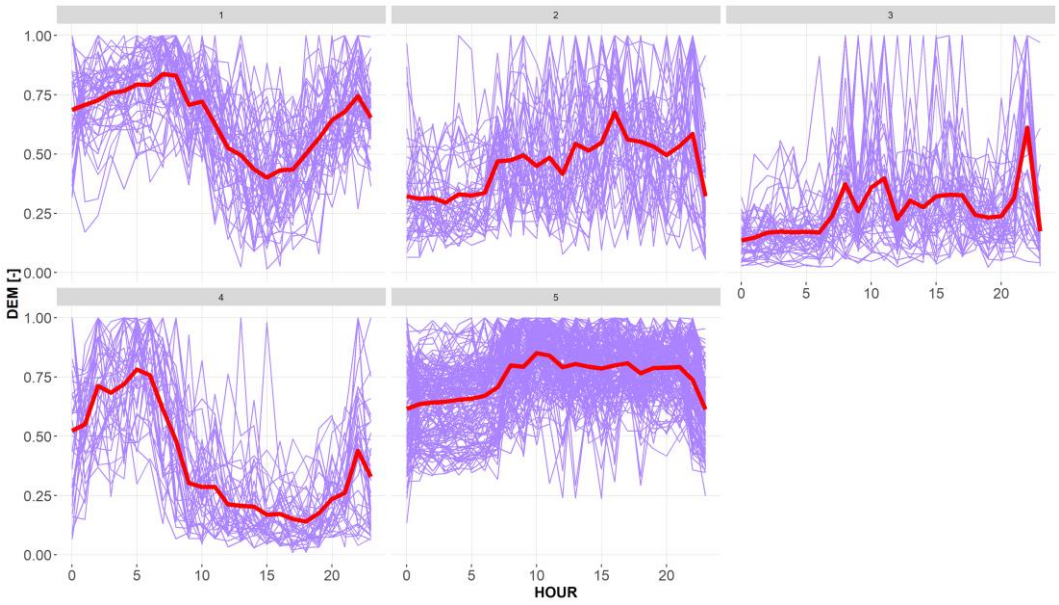
(c) (d)

Fig. VI-13. Evolution of (a) C_Index, (b) Davies-Bouldin, (c) Dunn Index and (d) Silhouette Index in Building 11195

Additionally, Fig. VI-14 presents the obtained clusters for K=3 with DS2 and K=5 with DS5, corresponding to the two processes with the largest amount of CVIs.



(a)



(b)

Fig. VI-14. Daily energy demand clusters for normalized in Building 11195 (Commercial building) (a) K = 3 with DS2 and (b) K = 5 with DS5

The following patterns are identified, based on the clusters from DS2, K=3:

- In Cluster 1, days with relatively constant profiles are grouped, including days with very low demand and other days with intermediate loads.
- The daily energy profiles in Cluster 2 correspond to intermediate load days. As in Cluster 1, the maximum demand is identified at around 7am. However, from 7am onwards, the demand decreases, probably due to the more favorable climatic conditions outside. The lowest demand period is found at around 1pm, and then the demand gradually increases until 9pm. At this time, the night setback decreases the set point of the SH, and the demand returns to values of the first hours of the day.
- In Cluster 3, a relatively constant demand is observed until 7am, when the demand drastically increases. At this time, the shopping building may open, coinciding with when the potential customers start to use this building. A high demand is maintained from approximately 7am to 9pm, probably coinciding with

the opening hours of this building. Finally, at 9pm, the heating demand returns to the same values as the first hours of the day.

5. Discussion & Conclusions

This chapter has proposed the use of unsupervised learning techniques for the identification of energy demand patterns in the buildings connected to the DH of Tartu. In order to study the accuracy of these mathematical ML techniques, different datasets have been tested in which different pre-processing activities are conducted. The great differences found between heat demand patterns in the buildings make impossible to present all of them. Therefore, a deeper analysis of four buildings has been presented, including buildings with significantly different energy demand profiles.

On the one hand, K-means algorithm using Euclidean distance as dissimilarity measure is the clustering algorithm that best performs among the tested algorithms. It is closely followed by the same algorithm but with Cosine distance. However, and with the objective of standardize the process, the identification of heat demand patterns is performed using K-means algorithm. Besides, between the six DS generated for each building, DS2 and DS5 are the datasets which obtain the largest amount of CVIs concluding the optimal clustering processes. Both data-sets coincide with the normalization process that only divided the hourly values by the maximum daily demand. However, the best normalization process completely depends on the profile shape of the demand in each building and cannot be a general conclusion for all the buildings under study. Thus, in some of the buildings one normalization process can be the best to separate demand patterns, whereas in other building other normalization process can be better.

The results for four buildings with different demand profiles are presented, showing the identified clusters and optimal clustering techniques for their respective profiles. Building 10045 corresponds with a residential apartment with SH and DHW demand fed by the district-heating network and Building 10051 is a residential apartment with no DHW demand. This building will probably have another heating system for this purpose.

However, as the smart meters are installed in the substations of the DH network, this demand is not measured. Building 10949 corresponds with a kindergarten in which the heat demand will be completely affected by the occupation of the buildings.

Regarding Building 10045, three main patterns have been identified. The first pattern reveals the night setback that rules the energy demand in that building from 11pm to 3am. However, the energy demand (including DHW + SH) increases and stays relatively constant throughout the day. This demand pattern matches the very cold months in winter, when the SH demand is much higher than the DHW and, therefore, there are no relevant energy demand peaks during these days. The second pattern reveals a typical energy demand for mid-season, when the energy demand for DHW is similar to that for SH and, consequently, the DHW demand peaks are not very relevant. Finally, the third demand pattern identified enables us to visualize the energy demand for summer days, when there is no demand for SH. Thus, the energy demand profile of this cluster roughly matches the DHW demand profile in this building.

Building 10051 identified four main patterns. This first pattern corresponds with summer period when there is no heat supply for space-heating purposes. The second pattern corresponds with cases in which the energy demand gradually increases until 7-8am approximately, when the peak demand is reached due to the DHW demand in these hours. From 7am onwards, energy demand decreases until 17pm, coinciding with the hours when the users of the building are supposed to be out of the building. This third pattern is constituted by days in which the heat demand remains relatively constant but with slight increase throughout the day and finally, the last pattern groups the days with a constant heat demand all over the day, but contrary to the first pattern, energy is consumed in the building in these days.

Regarding Building 10949, three main patterns were also identified. The first pattern of this building shows the most stable demand profile and matches the mid-season demand, when the demand for SH and DHW are similar. Thus, there are no relevant demand peaks throughout the day. The second pattern in Fig. VI-12 shows a similar

profile to the third pattern in Building 10045 and, similarly, this demand profile matches the typical demand in summer. In these days, the unique demand is the one independent from climatic conditions; even though the kindergarten is supposed to be empty of children, there might be activity inside the building. Finally, the third pattern matches the heat demand in the cold days in winter. The three demand peaks are caused by the SH demand required over these days and probably matches the occupational pattern of the building.

Finally, Building 11195 also presents three main patterns. The first pattern corresponds to the typical profile in summer days, when the heat demand in the building is very low. There is residual heat demand when the commercial building is supposed to be open. The second pattern in Building 11195 is very similar to the second pattern in Building 10949 and the third of Building 10045. Thus, the conclusions drawn are the same for this building. Finally, the third pattern identified in this building corresponds to the heat demand profile for winter days, when there is a high and relatively constant SH demand along the day due to the low outdoor temperature and the continuous comfort requirement in that building.

As a result of the abovementioned framework, the following conclusions are drawn:

- K-means algorithm using Euclidean distance as dissimilarity measure is the most robust clustering algorithm for this purpose.

The selection of the optimal normalization process is completely dependent on the heat demand values in each case. There are no general rules for the selection of the normalization equation. DS2 and DS5 from

- Table VI-1 are the most repeated optimal process along the 43 buildings under study.
- Accuracy evaluation of the different clustering processes is performed with more than 30 CVIs and mainly focusing on 4 mainly used indexes: *C_Index, Davies-Bouldin Index*, *Silhouette Index* and *Dunn Index*. Each of the analyzed indexes evaluate the clustering process using their own equation and therefore, different

results are obtained. From this study could be concluded that there is not a unique correct solution for determining the best clustering process.

- Theoretically, the optimal number of clusters is always equal to the number of daily profiles available. In other words, one day would correspond to one cluster. However, the objective of this study is to identify heat demand patterns that are repeated in different days, and this is why, the number of clusters that we are looking for is the minimal number of clusters that allow the visualization of all the patterns in the buildings. This is why, even though the optimal number is 10 in some of the clustering processes, a lower number of clusters is preferred.
- The final use of the buildings and consequently, the users' behaviors and energetic requirements determine the demand differences. Buildings for residential purposes present a night setback in the heating season (no setback in summer), whereas other type of buildings patterns depend on the particular use and occupation of the buildings.

6. Referred Appendix

The research presented along this chapter has been published by the author in JOURNAL OF BUILDING ENGINEERING journal by ELSEVIER. The reference (title and DOI) and the first page of this article can be found in the Chapter XI: Appendix.

Chapter VII

Classification Models for Pattern Prediction

Abstract

This chapter presents an exhaustive analysis for finding the solution to understand the clusters from previous chapter and to analyze the external variables that are determining the unsupervised classification. Thus, this chapter analyzes the use of different machine-learning classification models, influence of the dataset used or other conditions in order to optimize the classification accuracy. Additionally, classification and regression trees are applied to visually identify the heat demand patterns.

Resumen

Este capítulo presenta un análisis exhaustivo con el objetivo de encontrar una solución a la comprensión de los clústeres identificados en el capítulo anterior y analizar las variables externas que están determinando dicha clasificación no supervisada. Por lo tanto, este capítulo analiza el uso de diferentes modelos de clasificación de aprendizaje automático, la influencia del conjunto de datos utilizado u otras condiciones para optimizar la precisión de este proceso. Además, se aplican árboles de clasificación y regresión para identificar visualmente los patrones de consumo de calor en el distrito.

Chapter VII Classification Models for Pattern Prediction

1. Introduction

Previous chapter identified and analyzed different heat demand patterns among the different buildings under study and concluded the optimal processes for the identification of these patterns. Different unsupervised algorithms, by means of clustering, enabled to identify and classify the daily heat demand profiles by different datasets and number of clusters. It was concluded that there was not a unique correct clustering process but different processes in which each Cluster Validation Index (CVI) determine as optimal.

In this chapter, we are going one step further and we are developing classification models for predicting the patterns. Classification in ML is the process of predicting a categorical label using different variables and properties (predictors). For the case in this analysis, the categorical vector will correspond to the cluster/pattern variable resulting from the previous chapter. Thus, the main objective of this study is the prediction of the cluster classification, by means of external variables than are affecting this unsupervised clustering classification.

However, the pre-processing activities, by means of datasets and number of clusters will also determine the effectiveness of this process and it might differ from the conditions of the previous chapter. Therefore, a clustering process that was not considered optimal in the previous chapter might result to be the process in which the resulting classification has the closest correlation with the external factors affecting the heat demand in the buildings. While in the previous chapter the analysis was completely unsupervised and data-based, this chapter will find the relation between patterns and the variables affecting the demand, by means of different classification (supervised) models.

This chapter aims to gain insights on the possible conditions that determine the causes of the different heat demand patterns. In the way of developing a heat-load prediction model, it is necessary to identify the external conditions that could determine the classification. For this purpose, classification models, supervised ML models, are proposed where part of the data is used to train the model and the other part is used to as testing data. As a brief summary of the state of the art in this context, some classification models such as decision-trees (DT) [114] or association rules mining (ARM) enable to discover the potential variables affecting heat demand patterns [115]. ARM is widely used for discovering associations in large Boolean datasets; however, for numerical values such as outdoor temperature, this model is not valid. Thus, DT or specially classification and regression trees (CART) [116] are modelled to develop the linear relations between the external variables and the cluster classification. Besides, there are more complex classification models such as random forest (RF) classifier [117] or support vector machine (SVM) [118] for similar purposes. These complex models may allow obtaining greater efficiency levels at the expense of losing knowledge of the potential influencing variables and conditions.

In general, there are few works making the efforts towards the commented objectives and all the references found are referred to analyze electricity profiles. The variables affecting heat demand and electricity demand are different. Weather variables, and especially outdoor temperature, highly influence the heat demand (mainly in demand for space heating) in a building. Besides, domestic hot water demand is mostly affected by the users' behaviors inside the building. However, some of the methodologies found in literature about electricity pattern analysis could be partly used for the analysis of heating demand. The following lines summarize the main studies on this topic.

Regarding classification and regression trees, Capozzolli et al. developed a novel methodology combined with an adaptive symbolic aggregate approximation method and CART algorithm was proposed to identify infrequent and unexpected building energy patterns [119]. Two practical public buildings were used for case study analysis. Moreover, Liu et al. proposed a CART model using 6 different variables in order to

classify the electricity demand and improve the interpretability of clustering results [69]. This framework was applied to three practical offices in Chongqing (China).

On the other hand, McLoughlin et al. presented a clustering methodology for creating a series of representative electricity load profile classes and linked them with household characteristics using multi-nominal logistic regression [45]. The classification model developed by Viegas et al. enabled to classify new electricity customers using survey data and a limited amount of smart metering data [120]. This model obtained more than 50% accuracy in this classification task with only one week of smart metering data and this accuracy significantly improved with more data from the metering station. On its way to predicting electricity load profiles of buildings Vercamer et al. both Random Forest and Stochastic boosting in order to predict the load profile [121]. In this work, government and commercial data related to building characteristics were used as predictors for the classification models. Finally, Fabi et al. concluded that the demand patterns are influenced by the building characteristics [122], but they are also dependent on the behavior of the occupants and factors e.g., age, the number of children, lifestyle, etc.

However, in our study, the classification model is aimed to use more general information for two reasons:

(i) Lack of information. There is no additional information on building characteristics. The challenge is higher but the model resulting from this study will be widely applicable to any other case.

(ii) The model is proposed to be implemented at district-scale.

Summarizing, there is a gap of investigation of the variables affecting the previously defined daily heat demand profiles. Few references can be found on this topic even in the analysis of electricity profiles. Thus, this study attempts to study the most appropriate classification models for this purpose.

2. Objectives of this Chapter

The main and secondary objectives of this chapter are listed below:

- Development and analysis of the optimal CART characterize clustering and classifying each day to one of the clusters identified in Chapter VI.
  - Analysis of the effects of number of clusters and datasets into the effectiveness of the classification models developed.
  - Identification of the most affecting potential variables into clustering process.
- Analysis of the effectiveness of different classification models for the prediction of cluster classification in buildings connected to DH networks and comparison against CART.
- Evaluation of different error metrics in the prediction of the cluster classification.

3. Approach. General Methodology

For this study, we will start from the end of the previous chapter and we will use all the different datasets and pre-processing activities carried out in Chapter VI. The scope of this chapter is to develop the optimal classification model for predicting the pattern identified by the unsupervised learning, so that we can explain which are the variables that are affecting the demand in the different buildings.

We propose the evaluation of four different models, as it is illustrated in Fig. VII-1

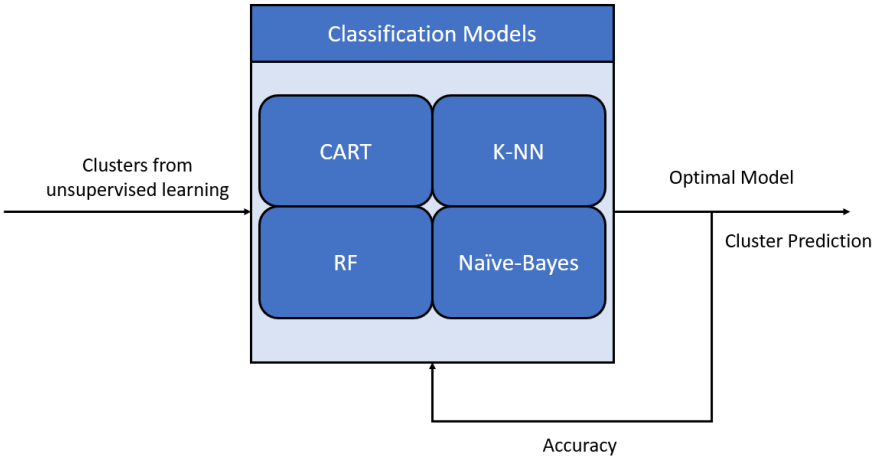


Fig. VII-1. General methodology followed in Chapter VII

The rest of the chapter is ordered as follows. First, in Section 3.1, we are introducing the basics for the different models used in the chapter and Section 3.2 explains the way we are evaluating the classification accuracy. Then, Section 4 shows the most relevant results, and as it was done in previous chapters, this section is divided into general results and particular results for individual buildings. Finally, Section 5 ends with the discussions and conclusions drawn from this study.

3.1. Studied Classification Models

This section will present the basis behind each of the algorithms used for the classification of the pattern. This section is as well divided into Classification & Regression Trees and the rest of the models, since the first one enables to visualize the results by showing the obtained tree.

3.1.1. Classification & Regression Trees (CART)

Classification & Regression Trees or CARTs algorithm is a supervised algorithm used for the construction of regression trees and was firstly proposed in the 80s by Breiman et al. [116]. These decision trees partition a whole dataset into smaller subgroups and then fit a simple constant for each observation in the subgroup. In this algorithm the partitioning of the dataset is carried out by successive binary partitioning, also called recursive partitioning.

The model begins with the whole data set, including predictors and the variable to be predicted (in this case, the pattern). In this first step the algorithm searches the predictor that divides the data into two datasets such that the Gini Index is minimized:

$$Gini\ Index = 1 - \sum_{i=1}^{N} (P_i)^2 \qquad \text{Eq. 14}$$

Where, P_i denotes the probability of an observation being classified to a particular class.

Having found the optimal best split that divides original dataset intro two subgroups, the process is repeated for each of the subgroups generated and this process continues until it is no longer possible to generate additional splits or some stopping criterion is reached. What results is, typically, a very deep, complex tree that may produce good predictions on the training set, but is likely to overfit the data, leading to poor performance on unseen (testing) data.

Thus, there is often a balance to be achieved in the depth and complexity of a tree to optimize predictive performance. To restrict a CART to an appropriate size, an early stop condition is set by providing the minimum number of observations in a node split. This variable is named as *MinSplit* and is used to avoid over partitioning. However, even though this stop condition is met, the tree can still be large and complex. Therefore, the normal method is to develop a large tree and then apply a post-pruning process using cost complexity parameter or *cp*. This cost complexity factor penalizes the cost function (*Gini Index*) for the number of terminal nodes of the tree. Usually, a hyper-parameter tuning is developed, evaluating multiple models across a spectrum of *cp* and use cross-validation to identify the optimal cost complexity factor.

For this study, the input variables or predictors used for tuning the tree are presented in Table VII-1. Note that even though these variables are all introduced to develop the tree, not necessarily all the predictors will be used. A simplified classification model (CART 2 in Table VII-1) is compared against the raw CART (CART1 in Table VII-1) with the above presented results, removing the hourly temperatures and holiday prevision from

the variables affecting the clusters. All this is summarized in the following table (Table VII-1).

Table VII-1. Summary of selected variables for developing CARTs

Variable	Type of Variable	Description	CART 1	CART 2
Weekday	Categorical	Day of the week: MON, TUE…	X	X
Month	Categorical	Month of the year: JAN, FEB…	X	X
Holiday	Categorical	Estonian holidays, including weekends: HOL/NO HOL	X	X
Holiday_Prev	Categorical	Holiday Prediction for next day: HOL/NO_HOL	X	
Mean_Temp	Numerical	Daily Mean temperature in ºC	X	X
Solar_Irradiation	Numerical	Daily total solar irradiation in kWh/m^2	X	X
Hourly_Temp.	Numerical	Hourly temperature readings in °C. (24 variables, one for each hour of the day)	X	
Summer*	Categorical	Divides summer and rest of the year[4]	X	X

The implementation of this algorithm in R has been made using rpart library [123], in which the tuning of the abovementioned parameters (*MinSplit* and C_p) is allowed.

3.1.2. Other Classification Models

As previously commented, CART models enable to rapidly visualize and understand the conditions that determines the cluster classification. However, it is important to compare the accuracy of CARTs with other frequently used classification models, even though the insights on how the predictors affect the classification is missed. The

[4] The Classification of this variable is presented in the Appendix (Chapter 3)

following paragraphs are used to introduce the classification models used in this study. These models use the same predictors than the CART 1 model shown in Table VII-1.

3.1.2.1. K-Nearest Neighbor (kNN)

K-Nearest Neighbor or k-NN classification model is one of the simplest supervised models in machine learning and can be used either for classification or for regression. In this model, the unique variable used for tuning the model is K, which refers to the number of nearest neighbors. The k-NN algorithm works as follows. For a specific observation to be classified, the Euclidean distance (or other distances: Manhattan, Minkowski, etc.) is calculated and the algorithm finds the K closest observations to that point. The label that is more times repeated within these observations will be the output for the classification.

This model performs better with low number of predictors. To avoid overfitting, the needed data needs to grow exponentially as the number of variables increase.

The implementation of this algorithm in R is made using *Class library* in R [124].

3.1.2.2. Naïve-Bayes (NB) Classifier

Naïve Bayes or NB classification model is a probabilistic classifier based on Bayes´ theorem, which assumes that each feature makes an independent and equal contribution to the target class. NB classifier assumes that each feature is independent and does not interact with each other, such that each feature independently and equally contributes to the probability of a sample to belong to a specific class. Bayes probabilistic theorem equation is the following:

$$P(A|B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \qquad \text{Eq. 15}$$

Thus, for each observation a table with probabilities is calculated and then used for prediction. For multi-label classification problems, the result for prediction is the class with maximum probability. The easiness of the model is compensated with usually poor prediction results.

The implementation of Naïve-Bayes algorithm is performed using *Naivebayes* library in R [125].

3.1.2.3. Support Vector Machine (SVM)

Following with complex classification models, Support Vector Machine or SVM enables the integration of linear and non-linear relations between the predictors. This supervised classification model is quite often used for multi-class classification problems by constructing hyperplanes in a multidimensional space that separates cases of different class labels. The hyperplane concept is only imaginable with variables with three or less dimensions. Thus, for three dimensions (three predictors) data, the hyperplane is a 2D plane. Thus, the hyperplane is one dimension less than the data.

For no linear relations between then variables, classification accuracy is drastically reduced. When the spatial separation between observations is not linearly possible, original dimension of the data is increased using Kernel functions. Kernel functions can be lineal kernel, polynomial Kernel or radial Kernel, among others.

For the implementation of this algorithm in R, *e1071 library* [126] is used.

3.1.2.4. Random Forest (RF) Classifier

The last classification model used in this study is Random Forest or RF. This model is like CART, since both of them are part of decision-trees. This model consists of many individual decision-trees that operates as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model's prediction.

The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well is the following: many relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

The tuning of this model is made using similar parameters than in CART and for the code implementation in R, library named *RandomForest* [127] is used.

3.2. Evaluation of the Models

3.2.1. K-fold Cross Validation

When evaluating the different classification models presented above, original datasets are divided into training and testing datasets. In order to obtain an accurate approach of the efficiency of the model, different training and testing data should be used for the calculation of the error metrics.

This study uses K-fold cross validation for the evaluation of the model performance. This methodology consists of dividing data into different subsets of the training data and calculate the average prediction error. This algorithm works as follows (see Fig. VII-2):

1. Randomly split the data into K subsets or K-folds.
2. K-1 subsets will be used as training data and the other subsets is used for testing.
3. Test the model and calculate the error metrics used for the evaluation of the model.
4. Repeat the process K-times, until each of the subsets is uses as testing data.
5. Calculation of the average metric error.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

Fig. VII-2. K-Fold Cross Validation example with K=5 (Training 80% and Testing 20%)

This study proposes a 5-fold cross validation, dividing data into 5 subsets of 20% of the data. Thus, 80% of data is used for training and 20% for testing and this process is

repeated 5 times. This process is carried out separately for each of the buildings in the network.

3.2.2. Error Metrics

For the evaluation of the cluster prediction by the CARTs, the classification accuracy defined in Eq. (16) is used. This accuracy metric evaluates the number of correct classifications against the total number of predictions. Thus, this metric will vary from 0 to 1, where 1 corresponds to the perfect classification.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \qquad \text{Eq. 16}$$

A more specific metrics coming from accuracy is the confusion matrix. More than an error metric, confusion matrix is a summary square matrix with same dimension as the number of classes to be predicted. The general shape of these matrixes is shown in Fig. VII-3. The diagonal of the matrix shown in Fig. VII-3 (colored in blue) corresponds with the correct classification predictions from the model, whereas the rest of the cells in the matrix are incorrect outputs.

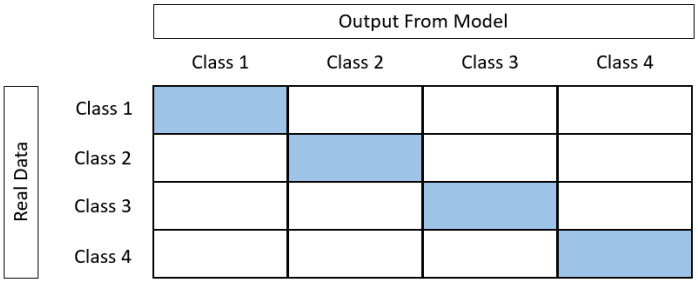


Fig. VII-3. Confusion Matric for a 4-Class Prediction

4. Results

This section presents the results obtained for the classification of the clusters and divided by the proposed models. Accuracy results for the different datasets, number of clusters and other variables are presented. Following the same structure than in the

previous chapter, this section is divided into general results analyzing the outcomes of all the buildings globally and a section focusing on the outcomes of four buildings.

4.1. General Results

4.1.1. CART

These models enable to obtain ease of visualization of the developed models, showing the lineal and binary relations between the predictors. In each of the buildings under study, two CART models are developed for the different datasets (normalization process) and the number of clusters resulting from the previous unsupervised clustering analysis. Thus, for each building 96 CART (six datasets and eighth clustering processes and the pruned variant[5]) models are developed in order to study the impact of each of the boundary conditions in the classification task.

Among all the analyzed cases, the maximum classification accuracy for each of the buildings is shown in Fig. VII-4.

Note that the accuracy that is shown in the following figures is the mean accuracy from the K-fold cross validation explained in Section 3.2.1. The model is tested five-times and the mean value is calculated.

[5] When we refer to the pruned variant we are using an special function in *rpart* library that enables to simplify the CART model

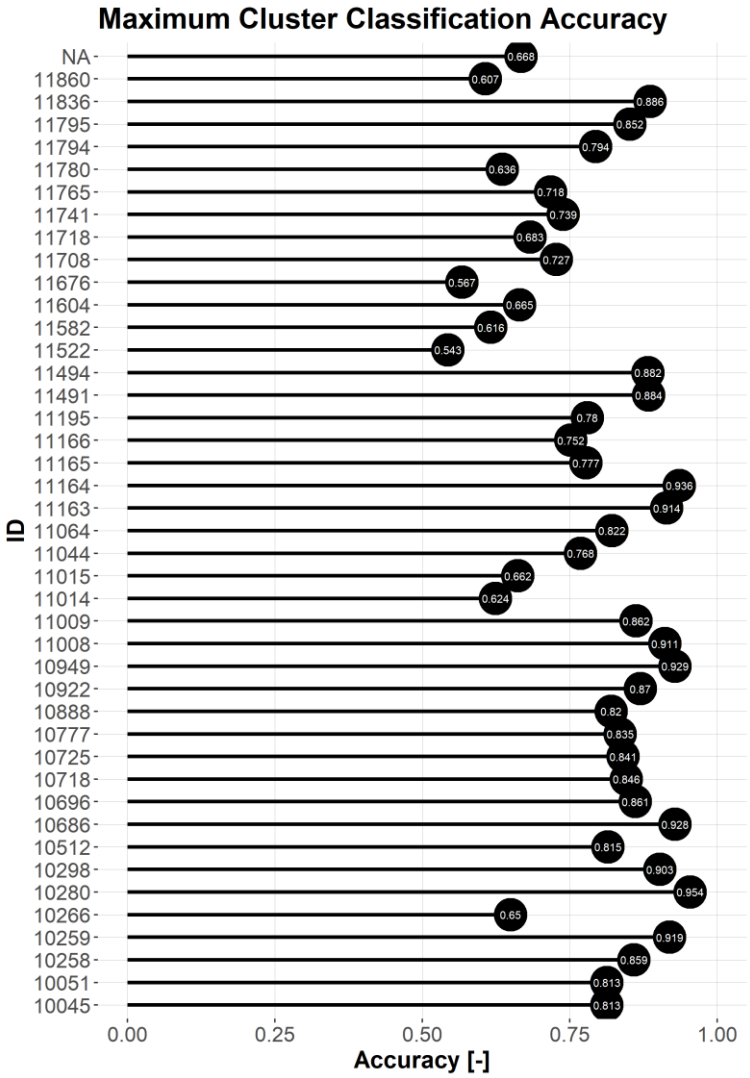
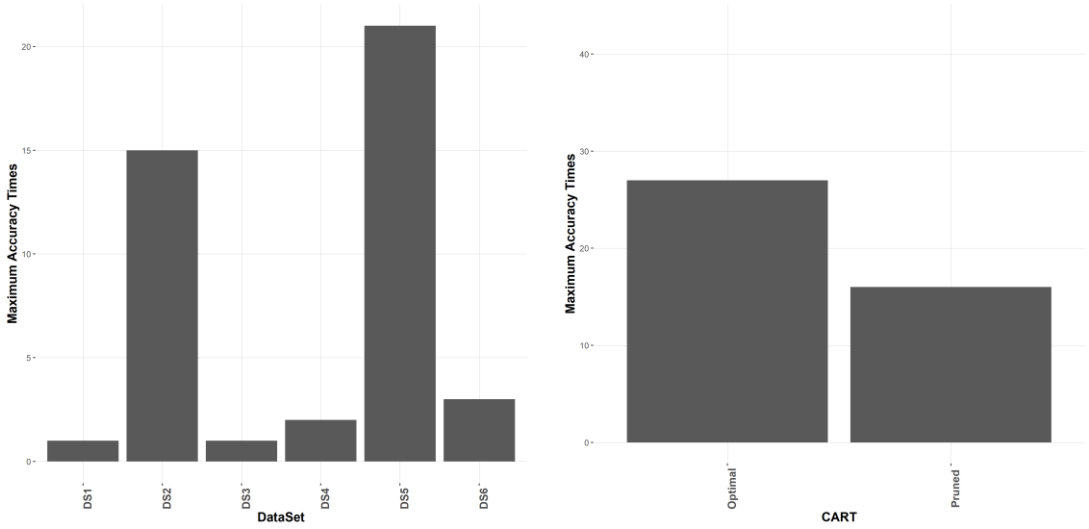


Fig. VII-4. Maximum classification accuracy obtained in each building using CART.

The accuracy metric (Eq. 16) measures the effectiveness of predicting the multi-class cluster variable, and in this case, accuracy ranges between 0.543 in Building 11522 and 0.954 in Building 10280. Using the same predictors than Table VII-1, effectiveness of this algorithm strongly varies from one building to another. It indicates that the relation between identified cluster and variables in Table VII-1 vary from one building to another. Thus, in all the cases the accuracy of the model overtakes the 50%.

Regarding the normalization process used for the generated datasets, the following figure (Fig. VII-5) classifies the maximum classification accuracies presented in Fig. VII-4 by the type of dataset used in each of the buildings under study. The figure on the left

side (Fig. VII-5a) presents the datasets with highest number of optimal cases, whereas figure on the right side (Fig. VII-5b) divides the accuracy for original and its pruned variant tree.



(a) (b)

Fig. VII-5. Number of Optimal cases divided by (a) Datasets and (b) Type of CART

DS2 and DS5, which corresponds with normalization process using only the daily maximum demand, obtain the greatest number of cases with highest classification accuracy among the buildings. These two datasets account for more than 80% of all the optimal classification process, regardless of if the outliers are removed from the dataset or not. Besides, the dataset in which the possible outliers are not removed get even more optimal cases than removing the outliers. This effect may be caused by the larger datasets trained when not removing the outliers. Thus, it is concluded that the effect of removing outliers is not a critical step in this classification process.

DS2 and DS5 are the most appropriate datasets for this purpose and this conclusion matches the one from the unsupervised clustering analysis. In that study, results using these datasets were also the optimal.

Besides, between using the optimal tree or the pruned tree, the results obtained do not clearly determine which model obtains better accuracy results. Around 60% of the

buildings obtain better results with the complex (not-pruned) CART model, whereas 40% of the building obtains the best results using the pruned tree. When using training data, it is obvious that the not-pruned model works better but for prediction purposes using testing data, this distribution is quite similar. The reduction of maximum depths and other variables of the tree simplifies the logic for the cluster classification, using general rules for the determination of the cluster. The use of more general rules leads to clearer conclusions and sometimes to better results than more complex CARTs.

On the other hand, among the different number of unsupervised clusters, a low number of clusters result in higher accuracy results. For the visualization of this conclusion, Fig. VII-6 shows the statistical variation of the accuracy obtained by the different models and divided for the number of clusters for all the buildings under study.

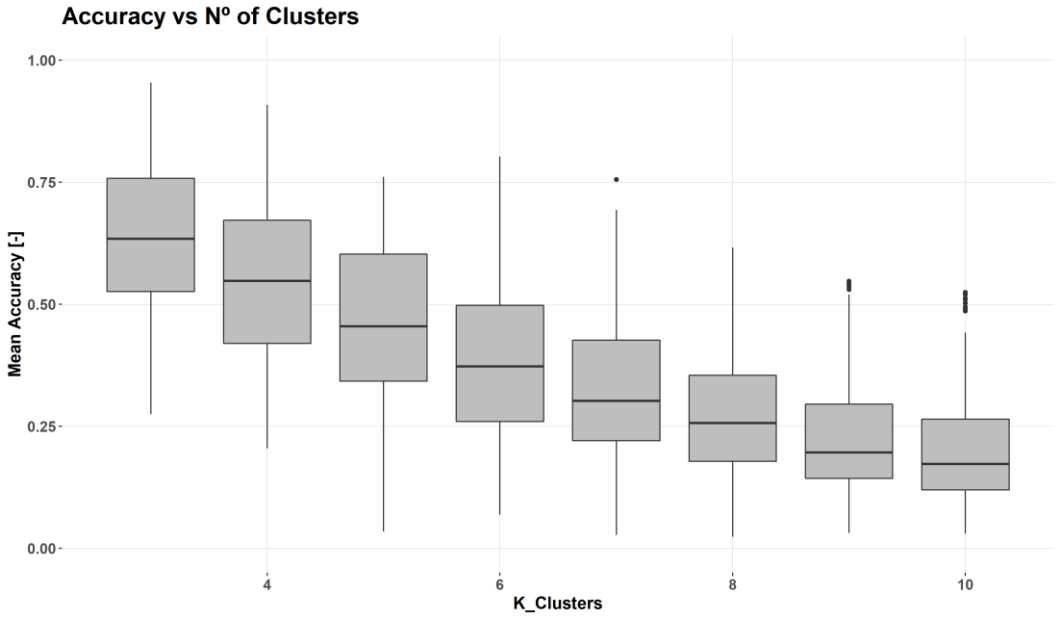


Fig. VII-6. Accuracy boxplot for the different number of clusters

The figure above (Fig. VII-6) shows how the accuracy of the model is reduced while increasing the number of clusters introduced. Maximum, minimum and mean values are reduced in almost all the cases. Besides, when considering the best classification results in function of the number of clusters, K=3 and K=4 (and specially K=3) are the most repeated number of clusters as optimal clustering among all the buildings under study.

Table VII-2 summarizes the obtained best classification results for each of the building as well as the dataset and number of clusters.

Table VII-2. Optimal Classification results for each building using CART.

Building ID	**10045**	**10051**	**10258**	**10259**	**10266**	**10280**	**10298**	**10512**	**10686**	**10696**	**10718**
Clusters	3	3	3	3	3	3	3	3	**3**	**3**	**3**
Dataset	DS6	DS6	DS1	DS5	DS5	DS5	DS5	DS2	DS2	DS3	DS5
Accuracy	0.813	0.813	0.859	0.919	0.650	0.954	0.903	0.815	0.928	0.861	0.846
Building ID	**10725**	**10777**	**10888**	**10922**	**10949**	**11008**	**11009**	**11014**	**11015**	**11044**	**11064**
Clusters	3	3	4	4	3	3	3	3	**3**	**3**	**3**
Dataset	DS2	DS5	DS5	DS5	DS2	DS6	DS5	DS4	DS2	DS5	DS2
Accuracy	0.841	0.835	0.82	0.870	0.929	0.911	0.862	0.624	0.662	0.768	0.822
Building ID	**11163**	**11164**	**11165**	**11166**	**11195**	**11491**	**11494**	**11522**	**11582**	**11604**	**11676**
Clusters	3	3	3	3	3	3	3	4	**3**	**3**	**3**
Dataset	DS2	DS2	DS5	DS5	DS5	DS2	DS5	DS5	DS3	DS2	DS5
Accuracy	0.914	0.936	0.777	0.752	0.780	0.884	0.882	0.543	0.616	0.665	0.567
Building ID	**11708**	**11718**	**11741**	**11765**	**11780**	**11794**	**11795**	**11836**	**11860**		
Clusters	3	3	4	3	3	3	3	3	**3**		
Dataset	DS5	DS2	DS2	DS5	DS2	DS5	DS2	DS2	DS5		
Accuracy	0.727	0.683	0.739	0.718	0.636	0.794	0.852	0.886	0.886		

To put an end of the presentation of the global results, a simplified classification model (CART 2 in Table VII-1) is compared with the above presented results, removing the hourly temperatures and holiday prevision from the variables affecting the clusters. Thus, this simplified model only considers the five variables shown in Table VII-1. Consequently, the number of affecting variables is reduced from 30 to only five variables.

This reduction of the classification accuracy is shown in Fig. VII-7.

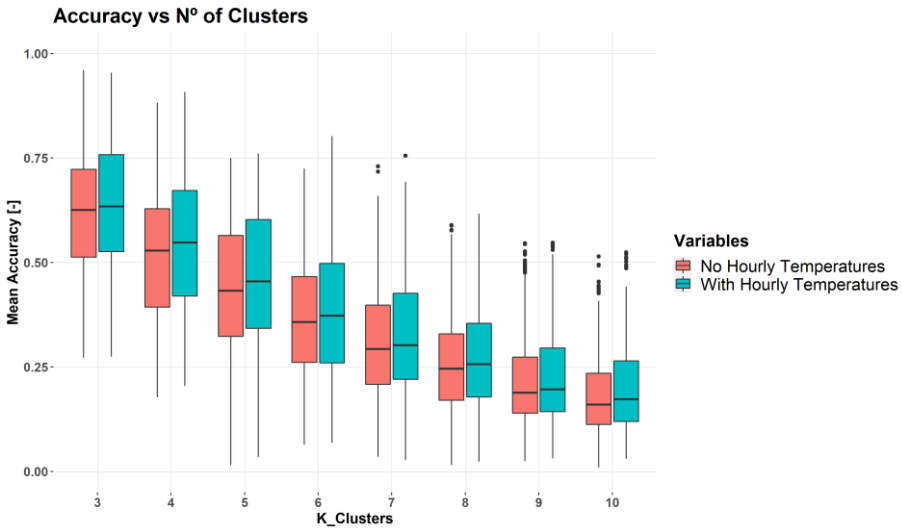


Fig. VII-7. Accuracy boxplot for the different number of clusters with and without hourly temperatures

As it was expected when using this number of variable reductions, the obtained classification accuracy is reduced in all the cases. However, this reduction of the accuracy of the classification is quite small compared to the reduction in the number of variables affecting the clusters.

Therefore, it is concluded that the importance of hourly temperature in the general shape of the heating profile is very significant. The hourly temperature completely affects the value of the demand but is not a crucial variable for energy profile classification. Moreover, the introduction of the hourly temperature variables in the model increases its effect in the classification of large number of clusters.

Table VII-3. Optimal Classification results for each building using CART without hourly temperatures.

Building ID	**10045**	**10051**	**10258**	**10259**	**10266**	**10280**	**10298**	**10512**	**10686**	**10696**	**10718**
Clusters	3	3	3	3	3	3	3	3	**3**	**3**	**3**
Dataset	DS6	DS6	DS5	DS2	DS2	DS5	DS5	DS2	DS5	DS2	DS5
Accuracy	0.707	0.707	0.859	0.934	0.688	0.947	0.887	0.79	0.9	0.860	0.773
Building ID	**10725**	**10777**	**10888**	**10922**	**10949**	**11008**	**11009**	**11014**	**11015**	**11044**	**11064**
Clusters	3	3	3	4	3	3	3	3	**3**	**3**	**3**
Dataset	DS2	DS5	DS1	DS5	DS2	DS6	DS5	DS1	DS2	DS5	DS2
Accuracy	0.834	0.843	0.79	0.882	0.912	0.892	0.881	0.65	0.68	0.831	0.838
Building ID	**11163**	**11164**	**11165**	**11166**	**11195**	**11491**	**11494**	**11522**	**11582**	**11604**	**11676**
Clusters	3	3	3	3	3	3	3	3	**3**	**3**	**3**
Dataset	DS2	DS5	DS5	DS5	DS2	DS2	DS5	DS2	DS5	DS2	DS5
Accuracy	0.961	0.943	0.828	0.747	0.784	0.816	0.839	0.557	0.677	0.699	0.59
Building ID	**11708**	**11718**	**11741**	**11765**	**11780**	**11794**	**11795**	**11836**	**11860**		
Clusters	3	3	4	4	3	3	3	3	**3**		
Dataset	DS5	DS5	DS2	DS5	DS5	DS5	DS2	DS2	DS5		
Accuracy	0.722	0.625	0.702	0.701	0.701	0.763	0.876	0.841	0.653		

4.1.2. Other Models

Apart from CARTs, other classification models have been tested in order to compare with the accuracy results obtained by CARTs. The main advantage of CART is the ease of visualization of the classification and consequently, the direct interpretation of the generated clusters. However, other models may obtain better accuracy results at the cost of a lower interpretation of this classification. In this section, an overview of the general results regarding all the buildings is shown.

Following the same order as for the CARTs, Fig. 7 presents the maximum mean accuracy in the different buildings and showing the three models applied: k-NN, Support Vector Machine or SVM and Naive Bayes Classifiers. Note that the presented accuracy is the mean accuracy obtained by the K-fold cross validation in the K testing options. Therefore, if the mean accuracy is one, it means that in the five cross-validations the obtained accuracy is one.

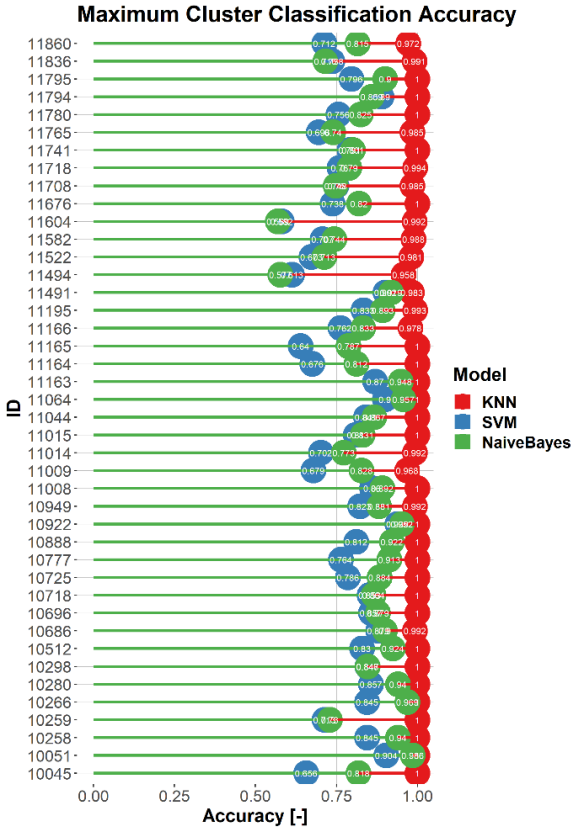
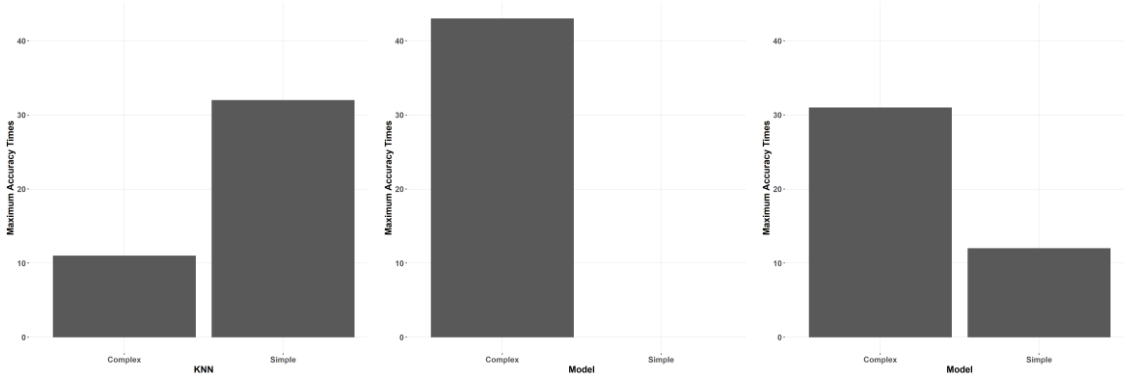


Fig. VII-8. Maximum classification accuracy in each building using kNN, Naïve-Bayes & SVM.

Previous figure presents the optimal accuracy among all the simulations including different datasets and number of clusters in which the demand profiles are divided. Thus, it is important to identify the conditions in which the maximum classification accuracy is obtained for each of the buildings. At a first look at the figure above, the best classification results are obtained with the simplest model: k-NN. This model obtains a classification accuracy of 1 in many buildings and in other cases the accuracy is very near to 1. Besides, Naïve-Bayes classifier obtains better results than SVM in almost all the buildings.

First, the complexity of the model is studied. For this purpose, Fig. VII-9 presents the number of cases in which the maximum accuracy results are obtained divided by the complexity of the variables including in the model. Thus, the "complex" model considers the hourly temperatures as input predictors (CART 1 in Table VII-1), whereas "simple" model does not include hourly temperature (CART 2 in Table VII-1).



(a) (b) (c)

Fig. VII-9. Comparison between simple and complex model for (a) k-NN, (b) SVM and (c) Naïve-Bayes

kNN model (Fig. VII-9a) is the unique model in which increasing the number of predictors does not increase the accuracy of the results, at least in most of the cases. On the other hand, SVM (Fig. VII-9b) always obtain better results when increasing the number of variables if these variables correlate with the prediction variable. Naïve-Bayes model (Fig. VII-9c) is situated in the middle of the two previous models and the model performs better when including hourly temperatures in most of the buildings.

Secondly, the type of DS used is also a key factor in this analysis. For this purpose, Fig. VII-10 presents the sum of cases in which the optimal classification is reached by the type of dataset described in previous chapter. Note that three normalization processes and outlier removal are considered within these six datasets. As it occurred with CARTs, the most repeated datasets are DS2 & DS5.

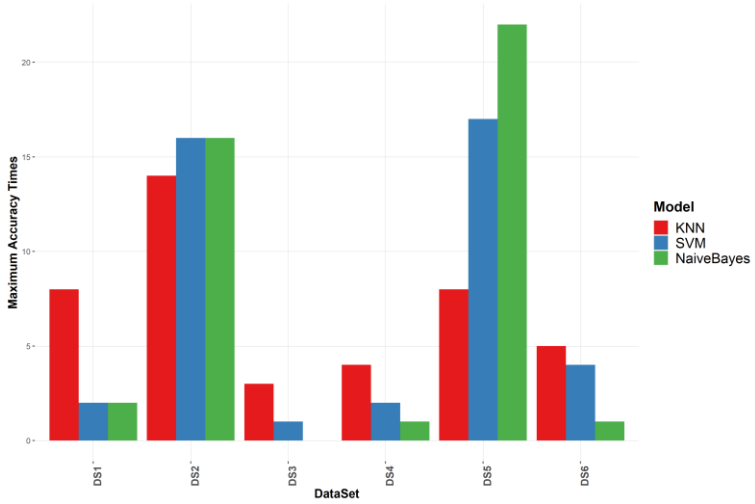


Fig. VII-10. Number of optimal cases divided by datasets for kNN, SVM & Naïve-Bayes models.

Observing the results presented in Fig. VII-10, kNN model present the best accuracy using DS2, followed by DS1 & DS5. However, DS5 is the most repeated optimal dataset in both SVM and Naïve-Bayes classifiers, closely followed by DS2.

Ending the analysis of the general results for these three models, a comparison between the accuracy results by the number of clusters is presented in the following figure. Therefore, Fig. VII-11 shows the statistical distribution of the accuracy by the number of clusters and for the three models presented.

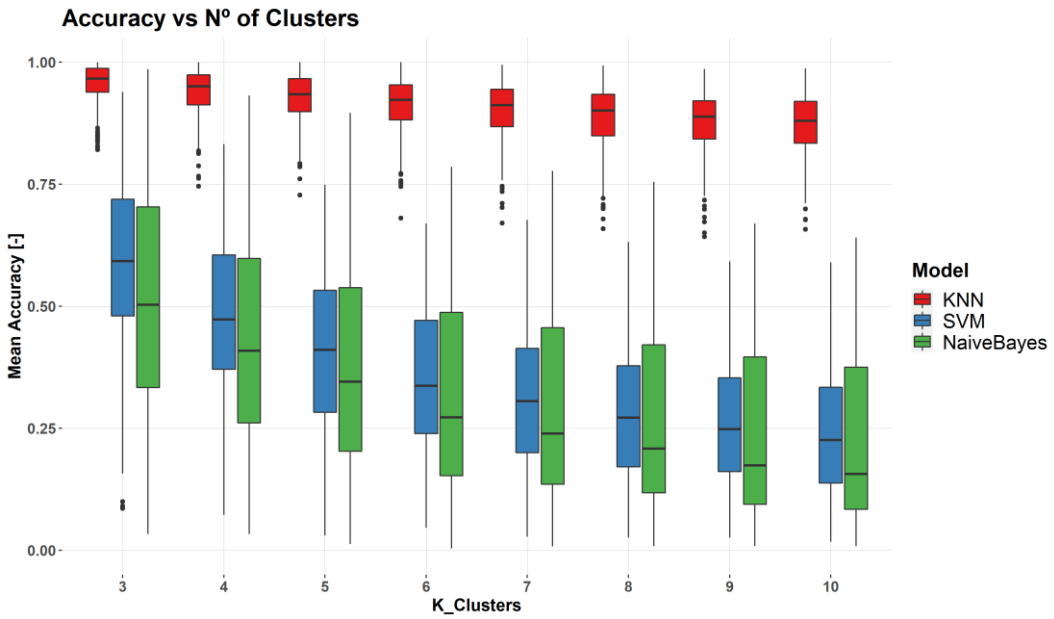


Fig. VII-11. Accuracy Boxplot for different number of clusters for the three models

As it was previously concluded, kNN is the model with highest accuracy results by far. Support Vector Machines and Naïve-Bayes classifier have similar accuracy results but quite far away from kNN. In all the cases, the best class prediction results are obtained with the smallest number of clusters (K=3). The efficiency of kNN model slowly decreases as long as the number of clusters increases, whereas the efficiency reduction of the other two models rapidly increases. Specially, the accuracy is drastically reduced in the smallest number of clusters, following a negative exponential reduction trend.

As it can be observed in Fig. VII-11, for K=3 & K=4, the accuracy of SVM model is higher than Naïve-Bayes, whereas this trend is reversed from K=5.

Thus, the simplicity of kNN model results in the most accurate prediction of the cluster in function of the predictor variables shown in Fig. VII-11. The simplicity of this algorithm and the relatively low amount of predictor variables used in this study makes this model to be the most appropriate for cluster prediction. This is why, kNN model is analyzed in the particular building assessment (Section 4.2.2) along with the results for classification and regression trees.

4.2. Specific Buildings' Analysis

Following the same structure than previous study about clustering algorithms, special focus on four buildings is presented. The four buildings chosen for this section are the same than presented in Chapter V and Chapter VI. Thus, two residential buildings (Building 10045, with DHW and Building 10051, with no DHW demand), an educational building (Building 10949) and a commercial building (Building 11195) will be presented in the following paragraphs. At this point, we remember that the heating profiles of these buildings can be found in Chapter XI, Appendix.

4.2.1. CART

4.2.1.1. Building 10045 (Residential Apartment with DHW demand)

To start with the results, Fig. VII-12 presents the evaluation of the accuracy by the number of clusters in which the heating demand is divided for the two models above-presented. Fig. VII-12a presents the accuracy results by the application of CART model using hourly temperatures, while Fig. VII-12b presents the results obtained with the CART without hourly temperatures. In both cases, DS6 presents the highest accuracy (yellow line), closely followed by DS4 and in both cases, the maximum is obtained in K=3 clusters.

(a) (b)

Fig. VII-12. Evolution of Accuracy by number of clusters (a) with hourly temperature and (b) without hourly temperature in Building 10045

Additionally, the trees formed using hourly temperatures are very complex to visualize due to the high number of variables affecting the clusters and consequently, the high number of branches of the tree. Nevertheless, the pruned tree of the model without hourly temperatures is shown in Fig. VII-13 along with the classified clusters.

Therefore, in general terms, Cluster 3 incorporates days from summer and mid-season periods with low and stable loads. Cluster 2 groups days in January and February, coinciding with cold days (high heat load), when there is a night setback in the demand. Finally, Cluster 1 groups the rest of the days, when the demand is highly reduced at mid-day. Thus, the main variables affecting this model are the Month of the year (with particular focus on months 1 and 2), the summer period, the day of the week, and the daily mean temperature.

All this explanation is visualized in Fig. VII-13. In each of the boxes that conform the models, the first number defines the cluster number of the predominant cluster in that step of the model. In the second row, the distribution of the existing clusters is presented and finally, the number in the third row indicates the fraction of the data

remaining after the previous classification step. Thus, in the first box of all the CARTs, this number in the third row is 100%.

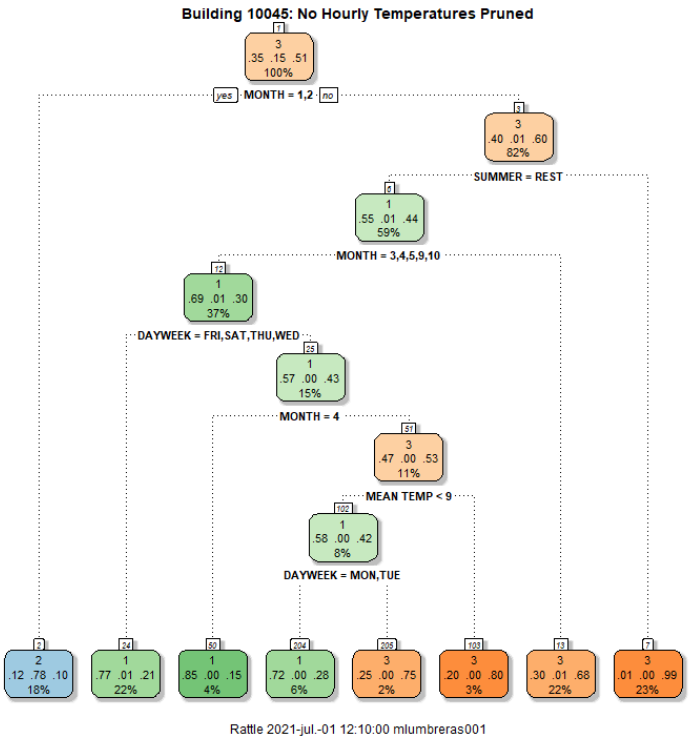
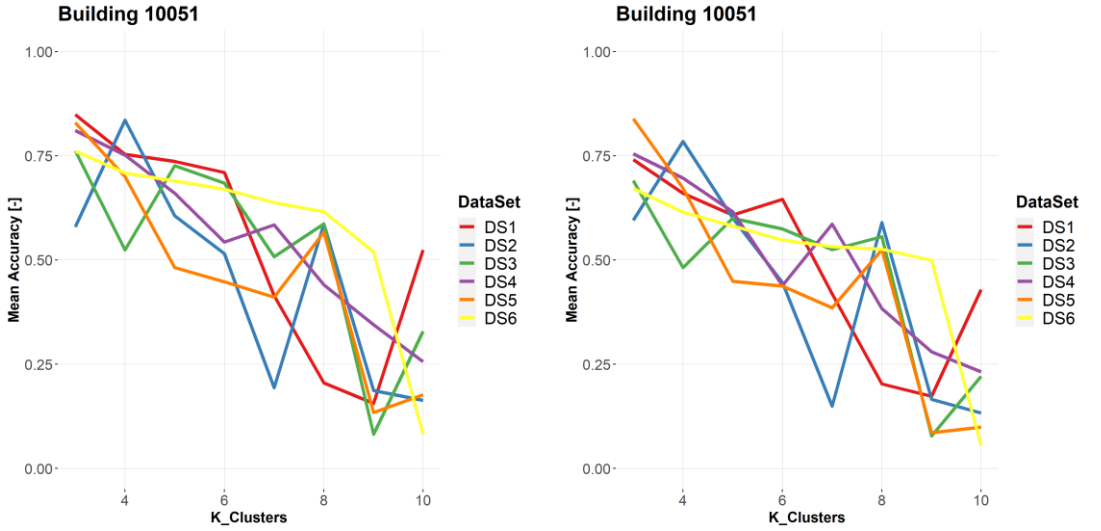


Fig. VII-13. CART Pruned Model without hourly temperatures in Building 10045

4.2.1.2. Building 10051 (Residential Apartment with NO DHW demand)

The accuracy results from all the CART models agree that K=3 clustering distribution is the optimal for this building. Firstly, Fig. VII-14 presents the evolution of the accuracy in the two CARTs proposed models by the different number of clusters. It is observed that DS1 obtains the highest accuracy in the CART model with hourly temperatures Fig. VII-14a), followed closely by the accuracy of DS4 & DS5. On the other hand, when removing hourly temperatures, the DS2 is the one that obtains the highest accuracy. Note that the maximum accuracy of the simplified model is higher than the more complex CART model. Comparing with Building 10045, the classification accuracy in Building 10051 is higher.



(a) (b)

Fig. VII-14. Evolution of Accuracy by number of clusters (a) with hourly temperature and (b) without hourly temperature in Building 10051

The Fig. VII-15 shows the CART scheme/logic of the simplified pruned model.

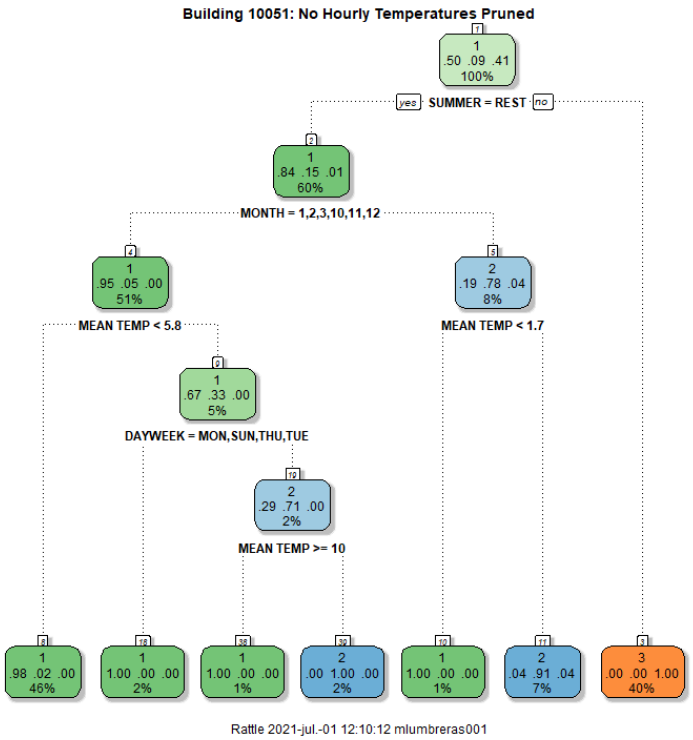


Fig. VII-15. CART Pruned Model without hourly temperatures in Building 10051

The first classification logic is determined by SUMMER variable. Thus, all the heating profiles in the third Cluster correspond with the days that are considered summer in this variable. The demand of the days grouped in this cluster (except one) is cero, including SH and DHW demand. Considering that the building is a residential apartment, it is concluded that DHW is supplied from an external heat source (not the DH).

Mid-season in this building is defined as all the months except from January, February, March, October, November & December, and the summer months. Therefore, winter or heating season is composed by the 6 months mentioned before.

On the other hand, Cluster 2 groups the following days:

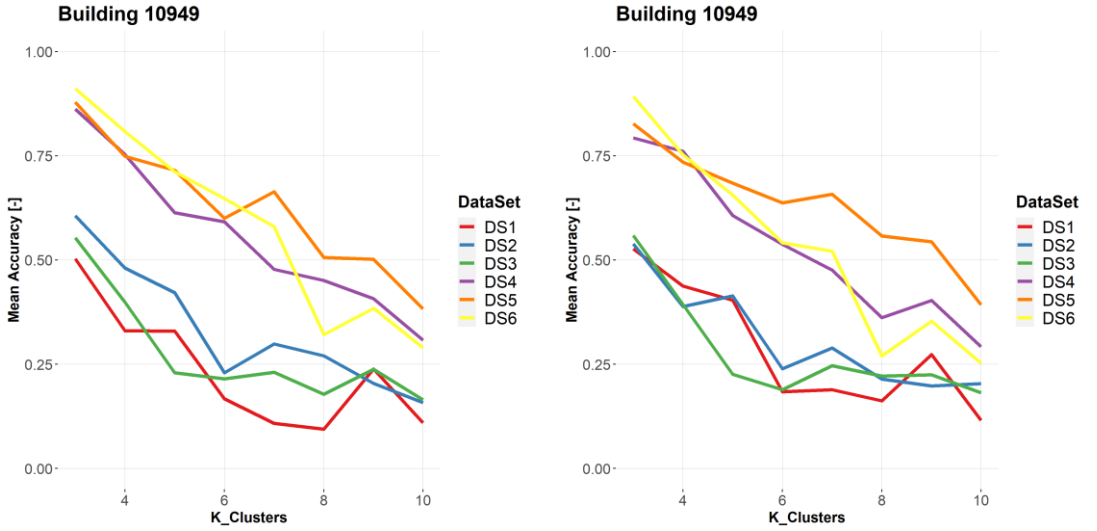
- Mid-season in which the mean temperature is above 1.7°C.
- Heating season in which the daily mean temperature is between 5.8 and 10°C and the day of the week is Wednesday, Friday and Saturday.

Finally, Cluster 1 is composed by the following logic:

- Days in the heating season in which the daily mean temperature is below 5.8°C.
- Mondays, Tuesdays, Thursdays and Sundays in the heating season with daily mean temperature above 5.8°C.
- Wednesdays, Fridays and Saturdays in the heating season in which the daily mean temperature is above 10°C.
- Mid-season in which the mean temperature is below 1.7°C.

4.2.1.3. Building 10949 (Kindergarten)

The heat demand pattern identified in the previous chapter were quite different from the ones in Building 10045 & 10051 due to the different use of the building. Fig. VII-16 presents the evolution of the accuracy results for the CART model (Fig. VII-16a) and its simplified version (Fig. VII-16b) by the different number of clusters. In both models, the highest accuracy results are obtained with three clusters. However, the best results in this building are obtained with DS6 but closely followed DS5 & DS4.



(a) (b)

Fig. VII-16. Evolution of Accuracy by number of clusters (a) with hourly temperature and (b) without hourly temperature in Building 10949

The pruned tree of the simplified model is shown in Fig. VII-17.

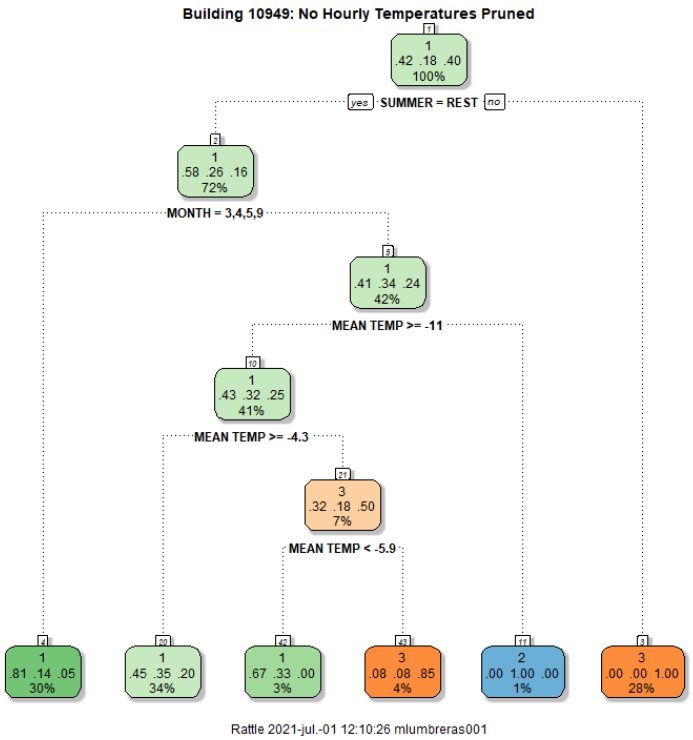
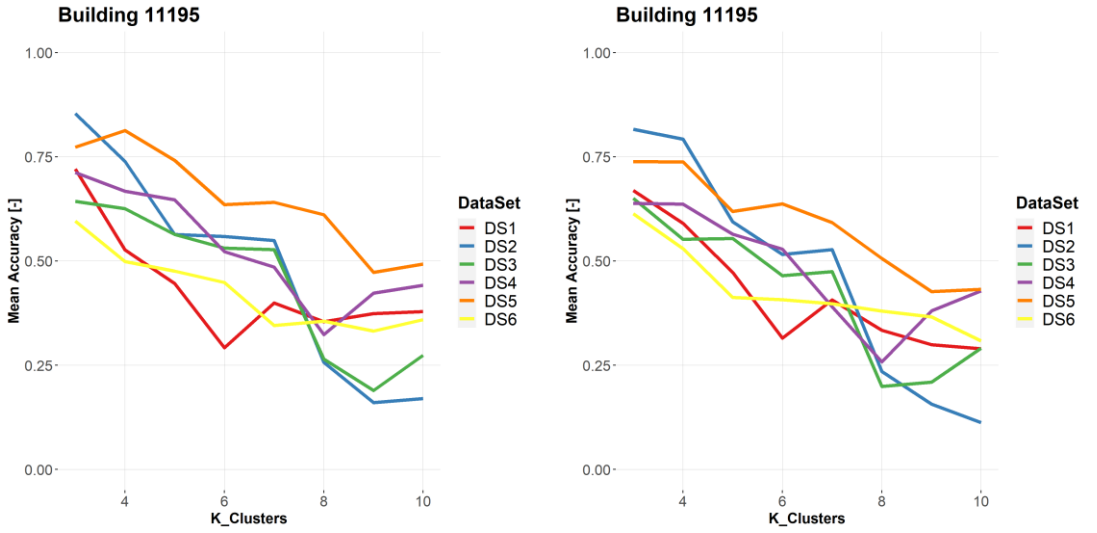


Fig. VII-17. CART Pruned Model without hourly temperatures in Building 10949

This classification model shows that the most determining variable for classification is the seasonal period (95% of the days were properly classified, considering only "summer" and "month" variables). Thus, Cluster 3 is composed by heating days in the summer period and some days in the heating season, when the daily mean temperature is above -5.9°C. Cluster 2 gathers daily heating profiles of the days in the heating season, when the daily mean temperature is below -11°C in the first and last months of the year. The days grouped in Cluster 2 are the rest of the days in the mid-season.

4.2.1.4. Building 11195 (Commercial/Shopping building)

Finally, Fig. VII-18 presents the evolution of the accuracy results for the CART model (Fig. VII-18a) and its simplified version (Fig. VII-18b) by the different number of clusters in the commercial building. In both models, the highest accuracy results are obtained with three clusters. However, the best results in this building are obtained with DS2 and followed DS5.



(a) (b)

Fig. VII-18. Evolution of Accuracy by number of clusters (a) with hourly temperature and (b) without hourly temperature in Building 11195

Then, Fig. VII-19 presents the form of the model. This classification model also shows that the most determining classification variable is the season, followed by the daily mean temperature (91% of the days were properly classified, considering only the

"summer" and "Mean outdoor temperature" variables). Thus, Cluster 2 is formed by days in the summer period and the mid-season with daily mean temperatures above (or equal to) 16°C. Moreover, Cluster 1 groups demand profiles in the first and last months of the year (heating season) and the mid-season with low outdoor temperatures. Finally, Cluster 3 gathers the days that are not classified in Cluster 1 or 2.

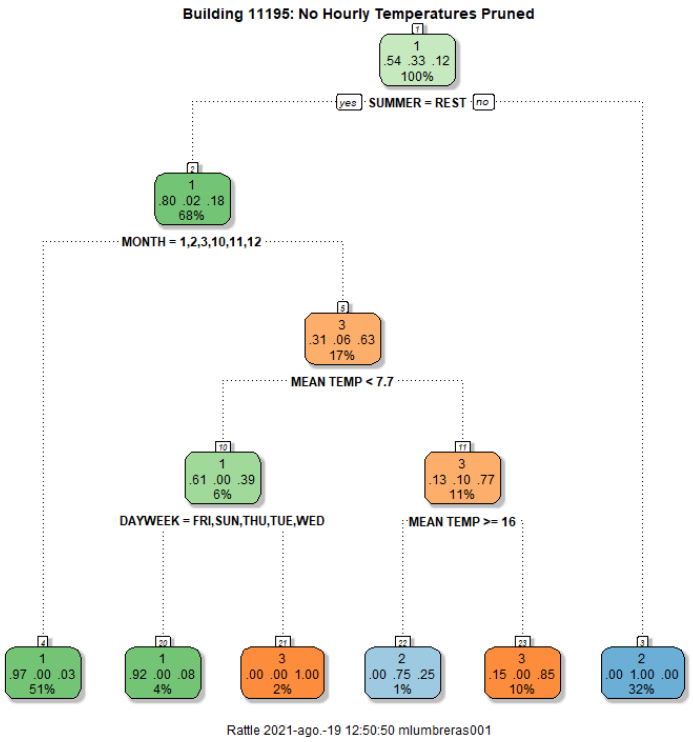


Fig. VII-19. CART Pruned Model without hourly temperatures in Building 10949

4.2.2. Other Models: kNN

As concluded in section 4.1.2, kNN algorithm shows the greatest accuracy for the prediction of the clusters. The following table presents the accuracy evolution for the optimal dataset in the four buildings under analysis.

Table VII-4. Evolution of the accuracy by clusters in the four buildings

Building	DS	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
10045	DS2	0.964	0.967	0.911	0.949	0.874	0.946	0.914	0.911
10051	DS5	0.987	0.977	0.985	0.957	0.906	0.954	0.939	0.960
10949	DS5	0.981	0.956	0.989	0.926	0.978	0.958	0.936	0.970
11195	DS2	0.965	0.961	0.979	0.989	0.951	0.941	0.951	0.951

5. Discussion & Conclusions

This chapter have presented a wide comparison between different classification models to predict the multiclass cluster variable coming from the previous study about unsupervised clustering algorithms. Unsupervised clustering groups data only by the "shape" of the heating profiles in the building, whereas this supervised study enables to obtain the correlation between heat demand patterns and the external variables (climatic + calendar variables) that could be the main causes for the existence of different heating profiles and heat demand patterns. For this purpose, four different models have been developed, including the use of classification and regression trees (CARTs). The great differences found between heat demand patterns in the buildings make impossible to present all of them. Therefore, a deeper analysis of four buildings has been presented.

Regarding the effectiveness of the analyzed models, the best accuracy results are obtained with kNN model. This model enables to obtain very accurate prediction of the cluster, with accuracy values above 0.9 in almost all the buildings and cases analyzed. The simplicity of the algorithm behind this model contrasts the very accurate classification results obtained. However, the main disadvantage of this algorithm is the lack of knowledge of the factors determining the cluster classification. The rest of the models present similar accuracy results for cluster prediction in which the best result is obtained by different models in different buildings. Particularities of each model have been previously presented.

Moreover, all the models agree that a low number of clusters are easier to be characterized and predicted. Thus, the clusters formed with K=3 from previous chapter are the one with highest prediction accuracy and consequently, the clusters with greater relation with the external variables or predictors. CART, SVM & Naïve-Bayes models strongly reduce the accuracy as well as the number of clusters increase. However, accuracy from kNN model is also reduced with the number of clusters but with a lower slope comparing with the rest of the models.

As for the datasets used for the various simulations, all the models also agree that DS2 and DS5 are the optimal for classification purposes. These datasets correspond with the normalization process in which the instant heat demand value is divided only by the maximum daily heat demand. This result also agrees with the results obtained in the unsupervised clustering algorithm, in which these datasets obtained the highest share of CVIs. Thus, these two datasets (removing or not the possible outlier from the original data) enable to better differentiate heat demand patterns and typical heat profiles as well as the formed clusters show the greater relation with the external variables including weather and calendar variables (Table VII-1).

Finally, CART algorithm enabled to obtain a greater knowledge of the causes affecting the different heat demand profiles. This study showed that in general, this algorithm is effective for cluster prediction based on the obtained results. The results proved that the models developed using variables in Table VII-1 (CART 1) perform slightly better than the models developed using variables in Table VII-1 (CART 2). However, the “complex” model uses 24 more variables than the simplified model and consequently, the accuracy difference between these two models is not big enough to justify the use of the “complex” models. Thus, we concluded that the CART model without hourly temperatures as predictors is the optimal CART for this purpose. Besides, among the variables in Table VII-1, SUMMER/REST, DAY of the week and mean temperature are the variables that most affect the clustering process. In general, the simplest models are more effective for prediction (testing data), whereas complex models are valid for characterization purposes (training data).

To sum up, the most interesting model is the CART, since it allows visualizing the rules that determine the cluster classification. This model results to be quite effective with low number of clusters and its effectiveness strongly reduces when the number of clusters increase. On the other hand, kNN model results to be the most effective classifier with very accurate results for almost all the cases. Note that the advanced tuning process of a SVM model requires the grid study of all the variables in the model and it requires a large computational cost and time, which makes it very difficult to implement in large scale (district or city scale).

6. Referred Appendix

The research presented along this chapter has been published by the author in JOURNAL OF BUILDING ENGINEERING journal by ELSEVIER. The reference (title and DOI) and the first page of this article can be found in the Chapter XI: Appendix.

Chapter VIII

Advanced Models for Demand Prediction

Abstract

This chapter presents the final step of the methodology developed for the characterization and hourly forecasting of heating demand in buildings connected to DH networks. In this chapter, different predictive algorithms and model configurations are trained and tested against real data. This chapter analyzes the influence of number of clusters, classification model or predictor variables, among others, in the accuracy of the predictions. As a result, it will determine which are the optimal conditions for this purpose.

Resumen

En este capítulo se presenta el paso final de la metodología desarrollada para la caracterización y predicción horaria de la demanda de calor en edificios conectados a redes de DH. En este capítulo, se entrenan diferentes algoritmos predictivos y configuraciones de dichos modelos, para posteriormente validarlos con datos reales. Este capítulo analiza la influencia del número de clústeres, el modelo de clasificación utilizado o las variables predictoras, entre otras, en la precisión de las predicciones. Como resultado, se determinarán las condiciones óptimas para este fin.

Chapter VIII Advanced Models for Demand Prediction

1. Introduction

In the previous chapters we studied how heat demand patterns can be identified using unsupervised learning and we developed classification models for predicting the mentioned pattern. This chapter will finish this multistep methodology by the prediction of the heating demand using the knowledge obtained from Chapter IV onwards.

Among data-driven models, ML (supervised) algorithms turn out to be effective for the forecasting of multiple operational variables in buildings´ demand, such as energy demand or supply temperature in the SH loop. ML models have been extensively applied in electricity demand modelling ([128] or [129]) and to a lesser extent for heating energy demand in buildings (e.g., [130]). Data from electricity demand in buildings has been more accessible than the heating demand data from heat-meters and consequently, a wider literature is found in ML algorithms applied to electricity management.

Heating energy demand in buildings have been traditionally simulated using the commonly known as "white-box" models which are based on computer programs that include the equations of the physics of a building. An example of this software can be: *TRNSYS, Design Builder*[6] *& Casanova,* among others. However, this methodology presents two main problems: (i) high-computational cost and time and (ii) a lot of information about buildings´ characteristics is needed: wall and windows transmittance, window to wall ratio or other constructive characteristics. Moreover, this type of methodologies does not present high accuracy if there is not much information available, including the occupational behavior of people inside the building. On the other

[6] We will use this computer program for simulating two small districts in Chapter IX

hand, the commonly known as “black-box” models or machine learning models try to develop less cost computing models based on demand data from previous years. There is a wide variety of models with different efficiency levels that will be analyzed in the following paragraphs, but in general terms, machine learning models enable to use the knowledge obtained from the data to forecast the demand, specially based on the weather and occupational data of the building. In the middle of white box and black box models coexist the grey box models. These models integrate prior physical knowledge and are typically formulated as state-space models through a set of stochastic linear differential equations, either in discrete or continuous time. Grey box models require a deep understanding of all relevant phenomena in a building that impact instantaneous or cumulated values of the load. Nevertheless, this study is only focused on black-box data-driven models since there is a need for low computational cost and high flexibility in order to adapt the model to all building typologies. And there is no extra information about the buildings under study.

Therefore, this chapter will be focused on buildings connected to a DH network, a factor that determines the objective of predicting the demand. The objective is not just the forecast of the heating energy in a single building but in a wide variety of buildings with different final uses and different heat demand profiles. This means that all the knowledge obtained for a unique building cannot be applied to the rest of the buildings (different demand patterns, as studied in previous chapter) and a general method valid for all buildings is needed.

The introduction of digital devices in the system allows the instant measure and gathering of all the operational variables in the network. This process opens the door to new opportunities in data-driven energy management. With operational data available, it is possible to characterize energy demand in buildings and consequently in the DH network. A more extensive review of the literature regarding heat load prediction has been presented in chapters before.

This chapter aims to present a novel method that combines both supervised and unsupervised learning algorithms for the hourly prediction of the heating demand in the buildings connected to the DH network of Tartu (Estonia). This study uses the data from the substations of the mentioned buildings and aims to compare the obtained results, in terms of prediction accuracy and computational time, against the model previously developed by the authors in [2] (explained in Chapter V), the so-called Q-T algorithm.

The previous chapters summarized the studies made in: (i) unsupervised clustering for the identification of heat load patterns and (ii) classification models for the prediction/characterization of the identified patterns. These studies were carried out using multiple datasets and machine-learning algorithms in order to identify the most suitable for this case study. Chapter VI concluded that K-means algorithm was the most suitable clustering algorithm to be applied but the CVI study was quite relative. From the classification model chapter (Chapter VII), the kNN model results the highest accuracy and CARTs were used to identify the most determining variables affecting heat-load energy patterns.

2. Objectives of this Chapter

This chapter aims to follow and finish with the methodology started in the previous chapters. Thus, this chapter aims to predict the hourly heat demand in respective buildings by using different machine-learning algorithms and using the knowledge obtained in the pattern identification and prediction studies. The prediction of a variable in machine learning is the process of obtaining a numerical value of a particular variable. Therefore, the main objective of this chapter is to obtain the value of the hourly heat demand using the supervised machine-learning models.

The technical objectives of the present chapter are the followings:

1. Development and analysis of different machine learning prediction models and evaluation of the error metrics.
    - Analysis of the effect of different cluster numbers

2. Evaluation of the effect of introducing the heat load patterns as input variable for the prediction models against not introducing them.
3. Analysis of the error metrics and the time spent for the computation of the models and evaluate the effect of:
    - Number of clusters
    - Machine-learning models

3. Approach. General Methodology

In this chapter, the final step of the multistep method for the heat load prediction in buildings is carried out. The chapter starts from the end of the previous chapter, therefore by the end of the cluster classification and uses this class prediction for evaluating the efficiency of different forecasting algorithms.

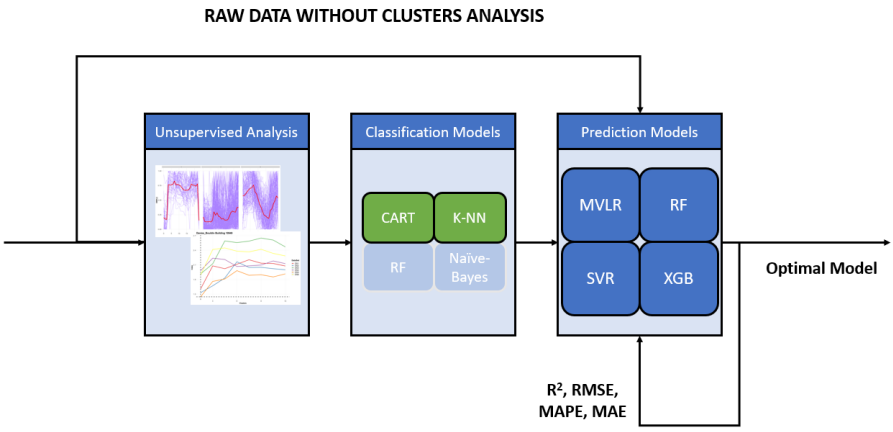


Fig. VIII-1. General methodology followed in Chapter VIII.

The rest of the chapter is ordered as follows. First, Section 3.1 describes the algorithms used for the final step of heat-load prediction and presents the variables used as predictors in each the variants of the models. Then, Section 3.2 presents the metrics used for the evaluation of the forecasting process. Additionally, Section 4 will show the results obtained and, as it was done in the previous chapters, the results are divided into general results and special and deeper focus on some of the buildings in the DH in Tartu. Finally, Section 5 will summarize the most relevant conclusions.

3.1. Studied Heat-Load Prediction Models

This section will provide an overview of the different algorithms used in this study. We present a brief overview of the mathematical approaches of 3 different models apart from the regression explained in Q-T algorithm.

Regarding the physical meaning of these models, Q-T algorithm, and its evolutions shown in this chapter, is the unique model in which the heat transfer effects are introduced. This is why, the study of the comparison between regression models and Q-T algorithm is presented separately. The rest of the section is divided as follows: Section 3.1 analyses the Q-T algorithm and is compared with other new regression models while Section 3.2 provides a study for the rest of the models developed: Support Vector regressor (SVR), random forest (RF) and finally the advanced extreme gradient boosting algorithm (XGB).

3.1.1. Q-T Algorithms versus other Regression Models

In Chapter V the Q-T algorithm was developed and analyzed. This model is based on a multi-variable linear regression (MVLR) and using three decision-trees four hourly data (LVL1, LVL2 and LVL3). Thus, the model was formed by 336 equations and more than 1000 parameters for each building. Even though R^2 values above 0.7 were obtained in the predictions carried out for a vast majority of buildings, this model was time and resource consuming. So, in this chapter, the original Q-T algorithm will be compared, in terms of error and computational cost, against other regression models using the unsupervised patterns identified for each building.

In general terms, linear regression attempts to model the relationship between two variables fitting a linear equation. In mathematical notation if y is the value to be predicted and x the independent variables:

$$y\,(w,x) = w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + x_p \cdot x_p \qquad \text{Eq. 17}$$

For determining the curve of the vector w, a cost function has to be minimized. For this case the function that to has to be minimized is the least square's function. This way, all

the parameters in the w vector are calculated. Using the same nomenclature, the function is the following, determined as L (y, x):

$$L\,(y,x) = \sum_{i=1}^{N} (y_i - x_i)^2 \qquad \text{Eq. 18}$$

Thus, this case study compares the following multivariable regressions:

1. Q-T algorithm (Chapter V)
2. Linear regression without clusters (MVLR_1)
3. Unsupervised Clusters (Chapter VI) + kNN classifier (Chapter VII) + Regression using clusters (MVLR_2)
4. Unsupervised Clusters (Chapter VI) + CART classifier (Chapter VII) + Regression using clusters (MVLR_3).

The input variables (independent) for each of the multi variable regressions are shown in Table VIII-1.

Table VIII-1. MVLR models and the input variables used in each case.

Model	T_{OUT}	G_T	Cluster	Weekday	Month	Hour Day	Holiday	Classification Model
Q-T algorithm (Chapter V)	X	X		X	X	X	X	--
MVLR_1	X	X				X	X	--
MVLR_2	X	X	X			X	X	KNN
MVLR_3	X	X	X			X	X	CART

3.1.2. Other Advanced Prediction Models

Besides, the multivariable linear regressions are compared against other less intuitive and more complex machine-learning models. The following three algorithms are used for the construction of the models.

1. Support Vector Regressor (SVR)
2. Random Forest (RF)
3. Extreme-Gradient Boosting (XGB)

3.1.2.1. Support Vector Regressor (SVR)

This model had also been previously used for classification purpose in Chapter VII. Specifically, this model was used as the classification model for clusters. SVR are one of the most popular and widely used machine-learning models and the functionality is the same as the SV for classification problems. The main advantages of this algorithms are the following:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

In SVR models, a hyperplane that divides data with the largest margin has to be found. This way, the hyperplane will have the dimensionality of the data minus one. Thus, if data has two dimensions, the hyperplane will be a line and if it has three dimensions, then hyperplane will have a shape similar to the one in Fig. VIII-2.

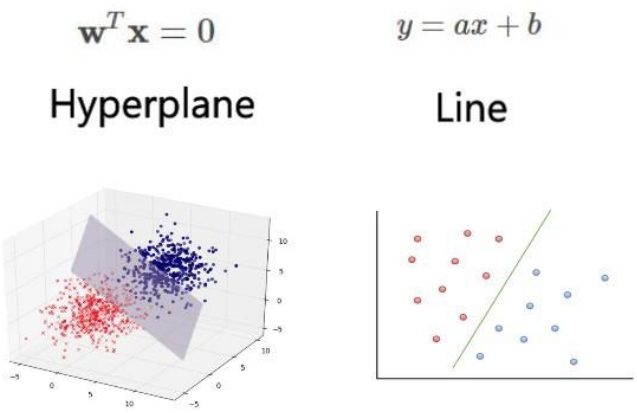


Fig. VIII-2. Support vector hyperplane example

The implementation of this algorithm in R is made using e1071 library in R.

For the evaluation of this algorithm, three models have been developed:

1. Support Vector Regressor without clusters (SVR1)
2. SVR with clusters using kNN classifier (SVR2)
3. SVR with clusters and CART classifier (SVR3)

Table VIII-3 summarizes the variables used as independent variables for training the models in the three variants. Note that some of the variables used in the models are naturally categorical (such as months or day of the week) cannot be included directly in the model and have to be divided into different features. This process of converting categorical into numerical variables is named **encoding**. Thus, the variable day of the week is divided into seven different columns and the values for observations are 0 or 1. Table VIII-3 presents the variables included in each of the models for this algorithm.

Table VIII-2. SVR models and the input variables used in each case.

Model	T_{OUT}	G_T	Cluster	Weekday	Month	Hour Day	Holiday	Classification Model
SVR_1	X	X		X	X	X	X	--
SVR _2	X	X	X	X	X	X	X	KNN
SVR _3	X	X	X	X	X	X	X	CART

3.1.2.2. Random Forest Regressor (RF)

The random forest (RF) is a supervised machine-learning algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. This bagging (also known as bootstrap aggregation) consists of random sampling of the data with replacement, enabling a better understand the bias and the variance of the dataset. Bagging makes each model run independently and then aggregates the outputs at the end without preference to any model, as it can be observed in Fig. VIII-3.

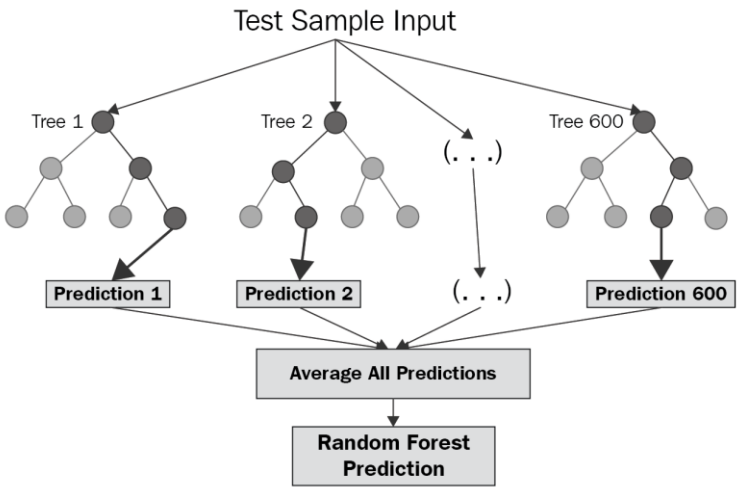


Fig. VIII-3. General functioning scheme of the random forest regressor for predictions

The main advantages of this algorithms are the followings:

- **Reduced risk of overfitting**: Decision trees run the risk of overfitting as they tend to tightly fit all the samples within training data. However, when there's a robust number of decision trees in a random forest, the classifier won't overfit the model since the averaging of uncorrelated trees lowers the overall variance and prediction error.
- **High efficiency,** and specially in large databases where there are lots of input variables without variable deletion.
- **Provides flexibility**: Since random forest can handle both regression and classification tasks with a high degree of accuracy, it is a popular method among data scientists. Feature bagging also makes the random forest classifier an effective tool for estimating missing values as it maintains accuracy when a portion of the data is missing.
- **Easy to determine feature importance**: Random-forest makes it easy to evaluate variable importance, or contribution, to the model. There are a few ways to evaluate feature importance. Gini importance and mean decrease in impurity (MDI) are usually used to measure how much the model's accuracy decreases when a given variable is excluded. However, permutation importance, also known as mean decrease accuracy (MDA), is another importance measure. MDA identifies the average decrease in accuracy by randomly permutating the feature values in samples.

On the other hand, the key challenges for this algorithm are the following:

- **Time-consuming process**: Since random forest algorithms can handle large data sets, they can provide more accurate predictions, but can be slow to process data as they are computing data for each individual decision tree.
- **Requires more resources**: Since random forests process larger data sets, they'll require more resources to store that data.
- **More complex**: The prediction of a single decision tree is easier to interpret when compared to a forest of them.

The implementation of this algorithm in R is made using *randomForest* [127] library in R.

For the evaluation of this algorithm, three models have been developed:

1. Random Forest Regressor without clusters (RF1)
2. Random Forest with clusters using kNN classifier (RF2)
3. Random Forest with clusters and CART classifier (RF3)

The following table summarizes the variables used as independent variables for training the model. The same way as it was done with SVR, the categorical variables are converted into binary predictors using the one-hot-encoding method. The variables included in each of the models are shown in Table VIII-3.

Table VIII-3. Random Forest regression models and the input variables used in each case.

Model	T_{OUT}	G_T	Cluster	Weekday	Month	Hour Day	Holiday	Classification Model
RF_1	X	X		X	X	X	X	--
RF _2	X	X	X	X	X	X	X	KNN
RF _3	X	X	X	X	X	X	X	CART

3.1.2.3. Extreme Gradient Boosting (XGB)

The final machine learning algorithm chosen for this study is the advanced extreme gradient boosting or XGB [131]. This algorithm is supervised learning model that evolves from gradient boosting algorithm. The same way than random forest work, this algorithm uses multiple decision trees to obtain the results. The result of the algorithm is obtained using all the trees developed, what is called ensemble method. Fig. VIII-4 shows a graphical illustration of how the algorithms performs.

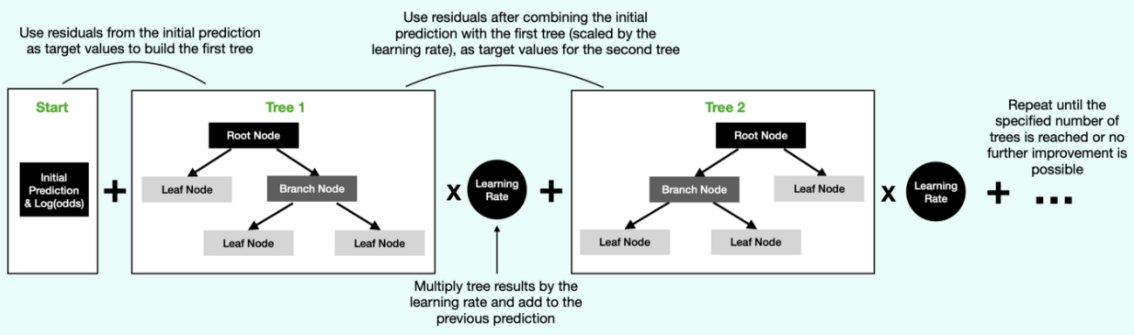


Fig. VIII-4. Extreme Gradient Boosting functioning scheme

For the total comprehension of the workflow of this algorithm, it is necessary to introduce the following parameters:

- **Learning Rate**: the value of each tree is scaled by the learning rate. This enables the algorithm to have a more gradual and steady improvement at each step.
- **Tree depth**: the algorithm allows you to control the maximum size of the trees to minimize the risk of overfitting the data.
- **Residuals**: actual (observed) value — predicted value.
- **Similarity Score and Gain determine** are the variables that will determine the shape of the models and the node splits in each of the steps.

The equation that governs Similarity Score is the following (Eq. 19):

$$Similarity\ Score = \frac{(\sum_{i=1}^{n} Residuals_i)^2}{\sum_{i=1}^{n}[Previous\ Probability_i * (1 - Previous\ Probability_i)] + \lambda} \quad \text{Eq. 19}$$

Where λ is a regularization parameter that is used to influence the weight of small leaves in the tree. As λ increases the importance of small leaves is reduced. Thus, the gain in each step is determined as follows:

$$Gain = Left\ leaf_{similarity} + Right\ leaf_{similarity} - Root_{similarity} \quad \text{Eq. 20}$$

So, the tuning of the algorithm is carried out using the number of trees, the depth of each tree and modifying the learning rate. The process consists in developing successive weak trees that use the results from the previous tree until the incorporation of more trees does not increase the precision of the results.

The implementation of this algorithm in R is made using *xgboost* library in R [132]

For the evaluation of this algorithm, two models have been developed:

1. Extreme Gradient Boosting without clusters (XGB1)
2. Extreme Gradient Boosting with clusters using kNN classifier (XGB2)

Since this model is expected to obtain higher accuracy results than the rest of the algorithms, when introducing clusters analysis, only kNN classifier is used. There are not relevant differences between using kNN or CART models for the prediction of the pattern class. Thus, the variables used are the same than in the rest of models. Note that this algorithm also needs the one-hot-encoding process before introducing variables to the model.

Table VIII-4. XGB models and the input variables used in each case.

Model	T_{OUT}	G_T	Cluster	Weekday	Month	Hour Day	Holiday	Classification Model
XGB_1	X	X		X	X	X	X	--
XGB_2	X	X	X	X	X	X	X	KNN

3.2. Model Validation and Error Metrics

This chapter tries to evaluate the efficiency of the models detailed in the previous section and to use common metrics that allow comparing results. The problem faced in this study is to predict a numerical variable, so it corresponds with a regression problem. Thus, the error metrics that needs to be used cannot be the same used in Chapter VII.

The same way than it was done for classification models, the evaluation of the prediction models is carried out using k-fold cross validation (Section 3.2.1 in Chapter VII). As in any machine learning model, the dataset is divided into training and testing datasets and the efficiency of the models is completely dependent on the way we divide the data. In this study, different training and testing datasets are used so that the obtained results can be correctly interpreted. Cross validation performance scheme was shown in Fig. VII-2.

This study proposes a 5-fold cross validation, dividing data into 5 subsets of 20% of the data. Thus, 80% of data is used for training and 20% for testing and this process is repeated 5 times.

Regarding the error metrics used for the evaluation of the prediction, the following metrics are used:

1. **R squared value (R^2) or Coefficient of Determination**: It is one of the most used error metrics in regression. Its value ranges between 0 and 1 and represents the proportion of the variation in the dependent (regressed) variable. Zero is the worst regression and R^2=1 means that the prediction is perfect. An R^2 value equal to 0.9 can be interpreted as: "Ninety percent of the variance in the baseline values can be explained by the modeled values".

   This metrics is calculated using the following equation:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \bar{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \mu)^2} \quad \text{Eq. 21}$$

   Where, Yi is the predicted vector, $\bar{Y}$ is the known vector and μ is the mean value of Y.

2. **Root Mean Square Error (RMSE):** This metric represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences. This metric ranges between 0 and ∞ and a low value is desired. The following equation governs this metric:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \bar{Y}_i)^2}{n}} \quad \text{Eq. 22}$$

   Where n is length of the vectors.

3. **Mean Absolute Percentage Error (MAPE):** This metric is usually used as a loss function in regression problems (for example in extreme gradient boosting)

and present in percentage units the relative error of the predicted vector. This metrics ranges between 0% and 100% and the equation is the following:

$$MAPE = \frac{100}{n} * \sum_{i=1}^{N} \left| \frac{Y_i - \bar{Y}_i}{Y_i} \right| \quad \text{Eq. 23}$$

4. **Mean Absolute Error (MAE):** This final metric is similar to the MAPE, but the error is given in predicted vectors´ unit. In this case for example, this metric is given in energy units, kWh. Thus, this metric will range between 0 and ∞.

$$MAE = \frac{\sum_{i=1}^{n} |Y_i - \bar{Y}_i|}{n} \quad \text{Eq. 24}$$

Additionally, the computation time spent for each of the algorithms is also measured and analyzed as an indicator of the quality of the model. Execution time for all the models has been monitored using Sys.time() function in R. The computational time is compared against the time used in the Q-T algorithm, so that the real reduction obtained by the reduction of regressions by the unsupervised learning may be observed. This variable is defined as the time required by the computer to train and test the corresponding multi-step model. The computational time required in this type of application is a critical variable when forecasting hour-ahead predictions and the model is extended to a high number of buildings connected to a DHN. The developed models are run in a personal laptop with no special requirements. The processor of the machine is an Intel(R) Core (TM) i5-10210U CPU@1.60GHz 2.11 GHz with 8 GB RAM. Thus, computational time should be understood as a relative variable since these times would be highly reduced if the models were run in a dedicated server.

4. Results

Following the same results' scheme than in previous chapters, results are divided into a section for general results comparing algorithms´ efficiency and other initial conditions and another section focusing on the results of the four buildings that we have analyzed throughout all the dissertation.

4.1. General Results

The most important and relevant metric used for the evaluation of forecasting heat load error has been the R^2 value. Fig. VIII-5 shows the value of this error metric for all the buildings analyzed using the four prediction models explained and additionally includes the results obtained by the Q-T algorithm (presented in Chapter V). Therefore, this figure shows the optimal R^2 values for all the buildings when including clusters/patterns as predictor variables: MVLR_3, SVR_3, RF_3 and XGB_3 from Table VIII-1.

We have presented some tables (from Table VIII-2 to Table VIII-4) that summarize the variables used as independent variables for training the models. Note that some of the variables used in the models are categorical (such as months or day of the week) and cannot be included directly in the model so that it is necessary to be divided into different features. This process of converting categorical into numerical variables is named encoding. For example, the variable day of the week is divided into seven different columns (from Monday to Sunday) and the values for each observation are 0 or 1.

Therefore, Table VIII-2 presents the variables included in each of the models for this algorithm. The same way as it was done with SVR, the categorical variables are converted into binary predictors using the one-hot-encoding method. The variables included in each of the models are shown in Table VIII-3 and Table VIII-4, respectively. Since this model is expected to obtain higher accuracy results than the rest of the algorithms, when introducing clusters analysis, only kNN classifier is used. There are not relevant differences between using kNN or CART models for the prediction of the pattern class. Thus, the variables used are the same than in the rest of models. Note that this algorithm also needs the one-hot-encoding process before introducing variables to the model.

Fig. VIII-5 illustrates the maximum R^2 values obtained for all the buildings among the different models, including MVLR, SVR, RF, XGB and Q-T algorithm.

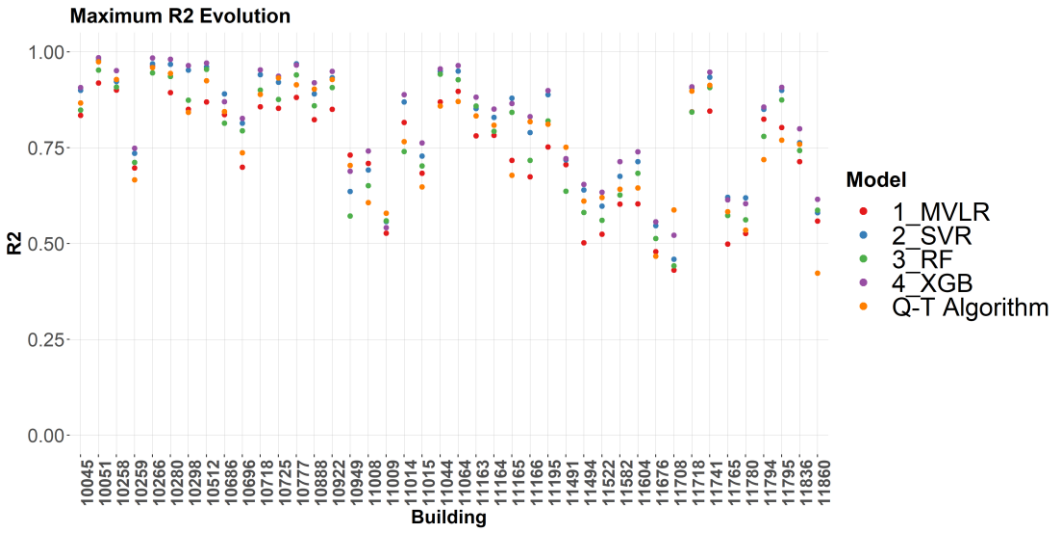


Fig. VIII-5. Maximum R^2 values for all the buildings in the district

It is observed that in all the cases, this error metric shows the best results when using the extreme gradient boosting algorithm. Only in a few buildings, the Q-T algorithm is the best method: Building 11009 and Building 11708. Moreover, the ranges obtained with the Q-T Algorithms are maintained when using more complex ML algorithms. Thus, the buildings with the worst results with Q-T Algorithm also resulted to be the worst with XGB or another algorithm. In the case of the optimal method, XGB, the R2 values range between 0.52 in the worst case and 0.99 in the best case. The highest the deviation between the points the worst results are obtained since the climatic variables result to be less dependent on the final heat demand.

In general terms, XGB is capable of increasing the R^2 metric in around 10% compared with the Q-T algorithm. Analyzing the other methods tested, apart from XGB, the best algorithm results to be the SVR followed by RF and finally, the simple MVLR. There is only one case, Building 10949, where the MVLR obtains the highest accuracy and it is closely followed by the Q-T algorithm. This is one case of very high point dispersion in which the complex models are not capable of obtaining good results.

In addition to R^2 metric, the Fig. VIII-6 presents the optimal MAE values for the buildings under evaluation.

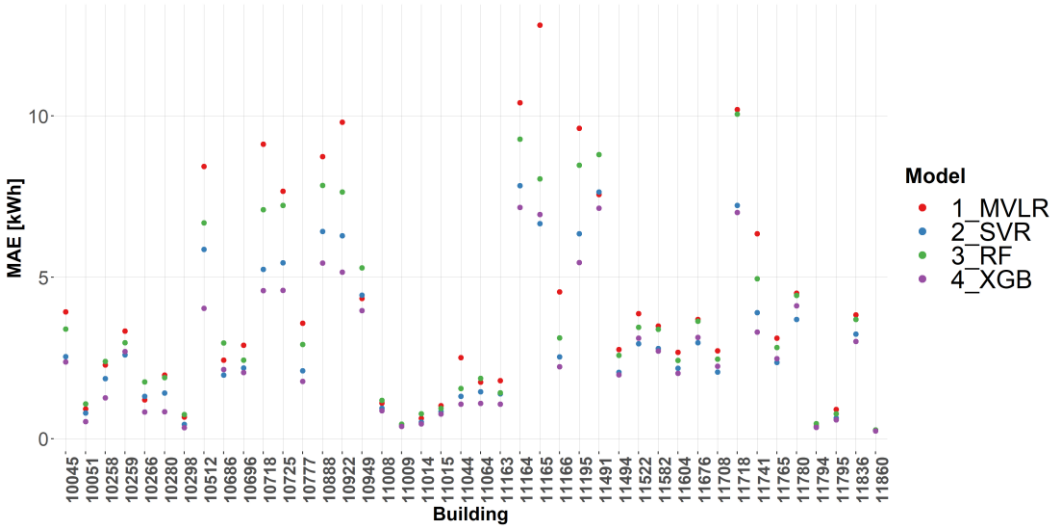


Fig. VIII-6. MAE values for all the buildings in the district

This metric presents an absolute value of the error, and this is why it is very case-dependent. So, in buildings with higher demand, this metrics is supposed to be higher than in buildings with lower demands. As a general conclusion, discussion from the previous figure is the same than the one shown with Fig. VIII-5. This metric ranges between almost 0 to a maximum error around 12 kWh in Building 11165.

4.2. Individual Buildings´ Analysis

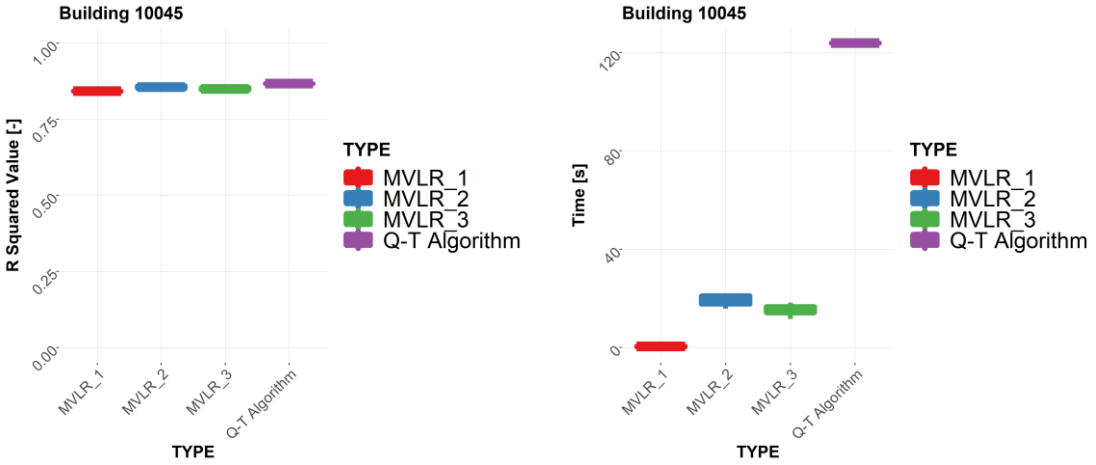
As it was previously stated, special focus on four buildings is presented. The four buildings chosen for this section are the same than presented previous chapters. Thus, two residential buildings (Building 10045, with DHW and Building 10051, with no DHW demand), an educational building (Building 10949) and a commercial building (Building 11195) will be presented.

For all these buildings, the following figures are shown and discussed:

- R^2 and Computation time comparison between MVLR and Q-T Algorithm.
- Comparison between the computational time required by the optimal methods.
- Summary of the optimal error metrics for all the models.

4.2.1. Building 10045 (Residential Apartment with DHW demand)

The following figure (Fig. VIII-7) shows the comparison between the results from Q-T algorithm and the three regression models proposed.



(a) (b)

Fig. VIII-7. (a) R^2 and (b) computation time of MVLR against Q-T Algorithm in Building 10045

These images show that the prediction errors for MVLR are a higher than Q-T algorithm but the difference always remains below 5%. However, the most significant differences are found in time analysis, since the introduction of unsupervised patterns reduce the number of clusters and consequently, the number of regression coefficients and mathematical operations. The R^2 value varies from 0.842 in Q-T algorithm up to 0.863 in MVLR_2 and 7 clusters. Moreover, the computational time required in these two cases goes from 123,81 seconds in Q-T algorithm to only 19,89 seconds. Thus, apart from obtaining better accuracy results, the unsupervised study enables to reduce the computational time to 20%.

The following figure presents the computational time required by the four studied algorithms when the number of clusters is three.

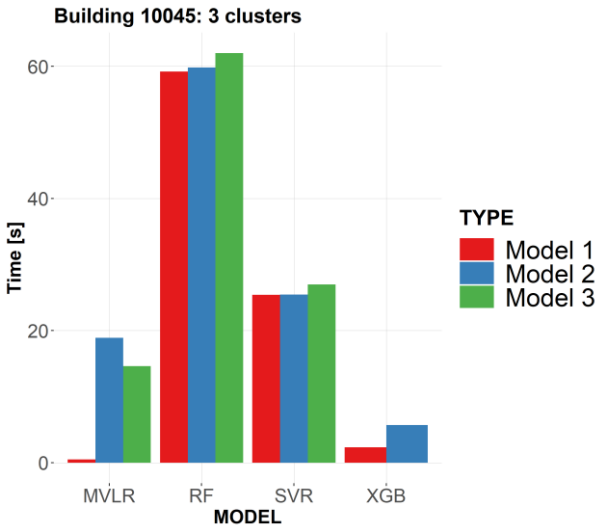


Fig. VIII-8. Computational time of the four models for three clusters in Building 10045

Model 1, Model 2 and Model 3 refers to the models developed using the same nomenclature than in Table VIII-1, Table VIII-2, Table VIII-3 and Table VIII-4. That is to say, Model 1 for MVLR refers to MVLR_1 in Table VIII-1.

Among the four algorithms and 11 models, the lowest computational time is obtained with MVLR_1, the simplest regression in which the cluster variable is not used as a predictor. Among the other models, XGB is capable, in any of its models, to forecast the hourly demand in 1.38 seconds in the first model (XGB_1) and 2.3 seconds by the second model (XGB_2). Thus, apart from obtaining the highest prediction accuracy, this model is the fastest except for the simple linear regression. The highest computation time is achieved when applying the RF model, with more than one minute required for achieving the results.

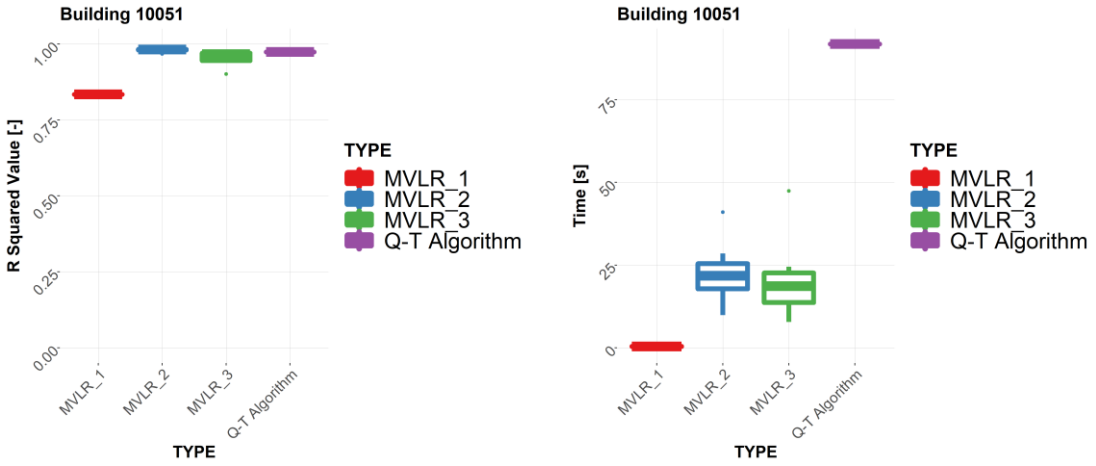
To finish with this section, Table VIII-5 presents the minimum error for each of the models:

Table VIII-5. Minimum error metrics for the three models in Building 10045

Model	RMSE	Nº Clusters	Model	MAPE	Nº Clusters	Model
SVR	3.5457	9	SVR_2	0.2118	3	SVR_2
RF	4.5218	6	RF_2	0.3374	3	RF_2
XGB	3.3533	8	XGB_2	0.1535	6	XGB_2

4.2.2. Building 10051 (Residential Apartment with NO DHW demand)

Similar to the methodology followed for Building 10045, this section analyzes the prediction accuracy and computational time required for each of the multi variable regression models. Fig. VIII-9 presents the R^2 value and the computational time [seconds] for the for variants of the regression.



(a) (b)

Fig. VIII-9. (a) R^2 and (b) computation time of MVLR against Q-T Algorithm in Building 10051

The Q-T algorithm explained in a previous chapter achieved better accuracy results for this building due to the lower distribution of the heating demand over the time. As it happened in Building 10045, the lowest accuracy results are obtained with MVLR_1, with R^2 values slightly above 0.8. Due to the quite linear trend of the heat-load demand and the outdoor temperature, the Q-T algorithm achieved a R^2 value of 0.974. However,

when applying MVLR_2, the R^2 value is even higher in most cases reaching the maximum prediction accuracy of 0.982. In the case of MVLR_3, the accuracy is a bit lower than the Q-T algorithm but very near any case, with R^2 values near 0.95.

Moreover, these two multivariable regression models achieve similar prediction accuracy than Q-T algorithm with very much less computational cost. While in Q-T algorithm, the computational cost was around 92 seconds, the MVLR_2 results from 7 seconds (two clusters) to 42 seconds (in the case of using 10 clusters). The maximum prediction accuracy in MVLR_2 is obtained when using three clusters (K=3) and the computational cost in this case was 12 seconds. Thus, the use of unsupervised clusters, apart from improving the prediction accuracy, enable to reduce the computational time to 13% of the time used in Q-T algorithm.

This time reduction could be critical for live predictions when managing a real DH network and a large number of simulations have to be simultaneously run. Fig. VIII-10 shows the computational time required by all the models under study.

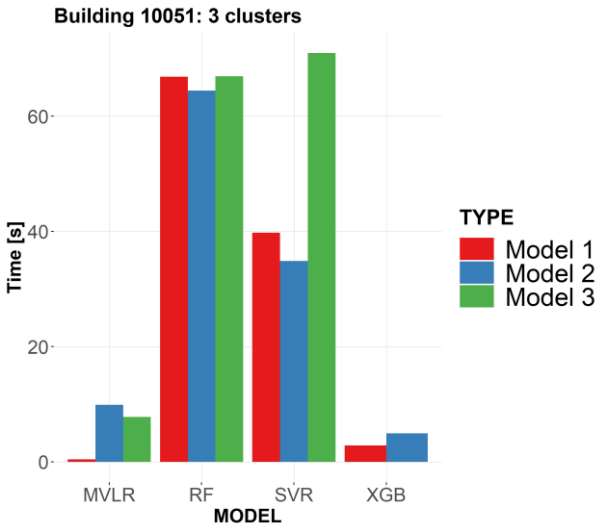


Fig. VIII-10. Computational time of the four models for three clusters in Building 10051

Lowest computational time is achieved with the simplest linear regression, MVLR_1, and as it happened in the previous buildings, it is followed by extreme gradient boosting. In this case, the computational time required is 2.53 seconds and 4.53 seconds for XGB_1

and XGB_2, respectively. In contrast with Building 10045, the SVR_3 required more time than RF to achieve the results.

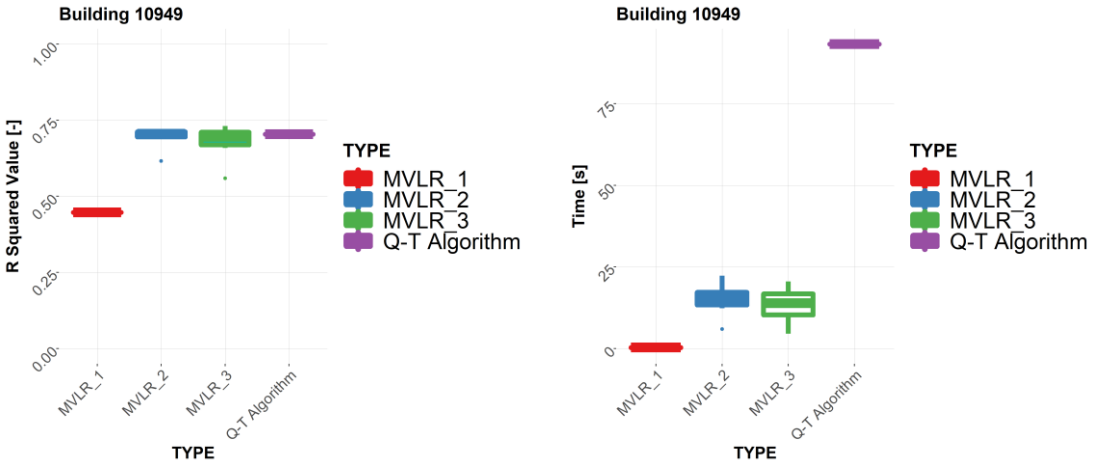
To finish the study for Building 10051, Table VIII-6 presents the minimum error for each of the models. The MAPE metric is *Inf* caused by the no heating demand hours that can be observed in some of the summer days in this building. As this building does not feed the DHW demand from the district-heating network, the total heat demand in summer days is zero. Thus, regarding the equation of the metric, this value will always be *Inf*.

Table VIII-6. Minimum error metrics for the three models in Building 10051

Model	RMSE	Nº Clusters	Model	MAPE	Nº Clusters	Model
SVR	1.0765	3	SVR_2	Inf	***	***
RF	1.5397	3	RF_2	Inf	***	***
XGB	0.8472	7	XGB_2	Inf	***	***

4.2.3. Building 10949 (Kindergarten)

The final use of this building differs from the previous two buildings and the distribution nature of the demand in this building is more chaotic and dispersed than in the other cases (Fig. IV-11). Worse prediction accuracy results were obtained in Q-T algorithm and similar patterns were expected also for the rest of linear regression models. Following the same method than in the rest of buildings, Fig. VIII-11 shows the prediction accuracy range, by means of the R^2 value, for the three models simulated against the Q-T algorithms and the computational cost necessary to run the models.



(a) (b)

Fig. VIII-11. (a) R^2 and (b) computation time of MVLR against Q-T Algorithm in Building 10949

The results observed in the figures above are similar to those obtained in the other two buildings, concluding that there is a common pattern for most of the buildings. The predictions obtained using MVLR_1, the model without unsupervised clusters, have much lower accuracy that Q-T algorithm. Whereas Q-T algorithm enabled to reach R^2 value of 0.704 in this building, results from MVLR_1 only reach values below 0.5. In contrast to MVLR_1, the other two models enable to reach predictions accuracy very near, even higher in some cases, to the results from Q-T algorithm. Additionally, the computational cost is still highly reduced from 93 seconds to around 20 seconds.

Continuing with computational time, Fig. VIII-12 presents all the computational times used by the models.

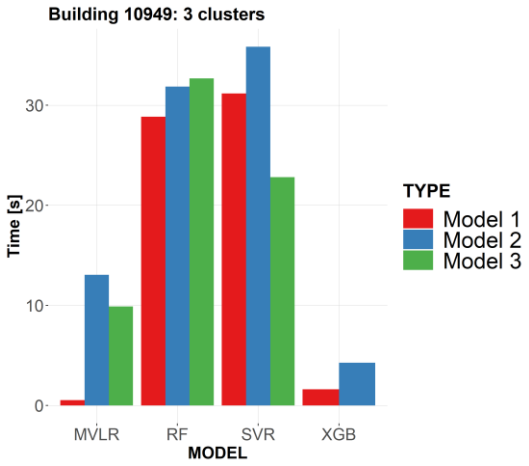


Fig. VIII-12. Computational time of the four models for three clusters in Building 10949

Again, the MVLR_1 is used the minimum time, only needing 0.26 seconds to predict the demand. This model is closely followed by the extreme gradient boosting variants. In this building, the time required by these models is 1.00 seconds and 2.06 seconds for XGB_1 and XGB_2, respectively. Moreover, the time required by the rest of the models is lower than in the previously analyzed buildings. The maximum time required in this case is slight above 30 seconds with SVR_2, around half of the time required in Building 10045 and Building 10051.

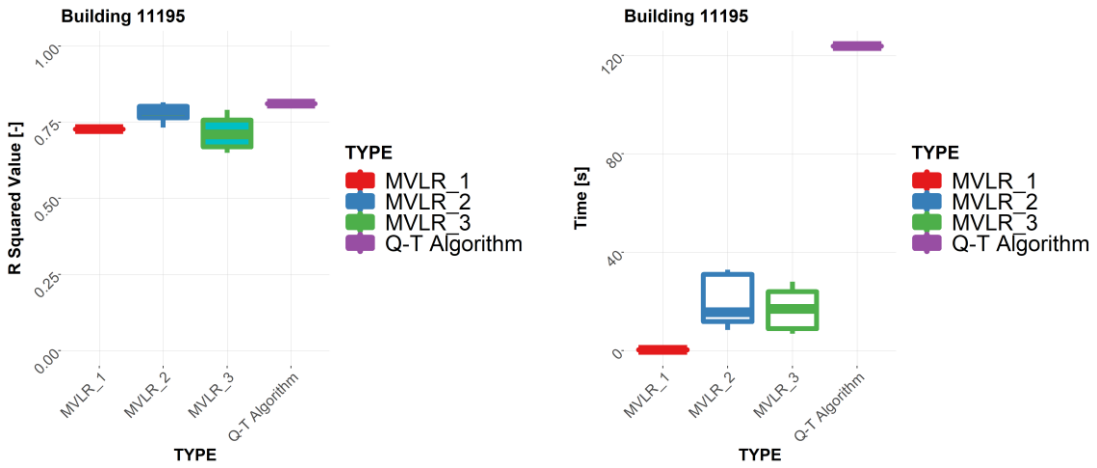
Finally, Table VIII-7 presents the optimal error metrics, RMSE & MAPE, for the three models applied to Building 10949.

Table VIII-7. Minimum error metrics for the three models in Building 10949

Model	RMSE	Nº Clusters	Model	MAPE	Nº Clusters	Model
SVR	6.9382	5	SVR_2	0.2955	3	SVR_2
RF	7.5155	3	RF_2	0.4345	3	RF_2
XGB	6.4517	3	XGB_2	0.2786	3	XGB_2

4.2.4. Building 11195 (Commercial/Shopping building)

The last particular building analyzed in this chapter is the commercial building, Building 11195. On the one hand Fig. VIII-13a shows the R^2 value of the Q-T algorithm against the regression variants, whereas Fig. VIII-13b presents a similar figure but for the computational time required.



(a) (b)

Fig. VIII-13. (a) R^2 and (b) computation time of MVLR against Q-T Algorithm in Building 11195

In this building, as it was previously commented, the regression variants are not capable to reach the accuracy of the Q-T algorithm, even introducing the cluster predictor, MVLR_2 and MVLR_3. Thus, while the R^2 value obtained in Q-T algorithm reached 0.811, the maximum R^2 obtained with the regression has been 0.791 in the best case. However, the computational time is reduced from more than 120 seconds to mean value near 20 seconds (it is reduced to a sixth part) and 37 seconds in the worst case.

Following with the study of the computational time in Building 11195, Fig. VIII-14 presents the time used by all the models in the case of using three clusters as predictor variables.

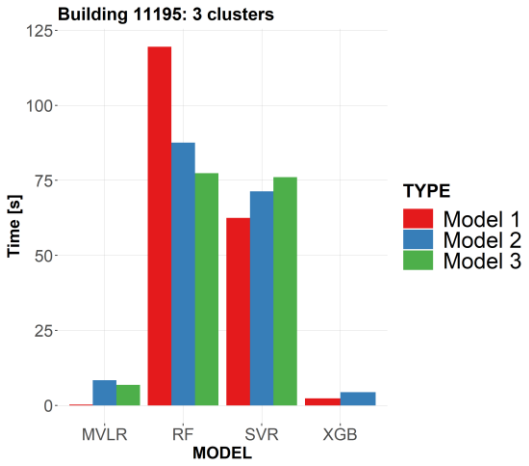


Fig. VIII-14. Computational time of the four models for three clusters in Building 11195

The lowest time is obtained with the simple linear regression named as MVLR_1. The time required in this case was 0.341 seconds. However, the accuracy of this models was clearly worse than the rest of the models. The variants of the extreme gradient boosting achieved low accuracy results, showing 2.07 and 4.40 seconds in XGB_1 and XGB_2, respectively. RF needed the longest time for obtaining the prediction with around two minutes in the worst case.

Finally, Table VIII-7 presents the optimal error metrics, RMSE & MAPE, for the three models applied to Building 10949.

Table VIII-8. Minimum error metrics for the three models in Building 11195

Model	RMSE	Nº Clusters	Model	MAPE	Nº Clusters	Model
SVR	9.049	3	SVR_2	0.722	3	SVR_2
RF	11.552	3	RF_2	1.064	3	RF_2
XGB	8.274	3	XGB_2	0.434	3	XGB_2

5. Discussion & Conclusions

This chapter presents the final step of this new methodology for the application in the prediction task for heating demands in buildings connected to DH networks. This chapter has analyzed different ML algorithms to study the prediction accuracy and

evaluate which are the optimal conditions for this forecasting task. The whole methodology combines unsupervised and supervised learning steps, finalizing with the supervised task of forecasting the demand. While unsupervised learning was used for the clustering of daily heat load profiles (Chapter VI), supervised algorithms are used to classify the unsupervised clusters (Chapter VII) and to forecast the hourly heat demand of the buildings (Chapter VIII).

For this chapter, four different predictive algorithms have been trained and tested with data from the DH of Tartu:

- Multi-variable Linear regressions or MVLR.
- Support Vector Machines or SVR.
- Random Forest or RF.
- Extreme Gradient Boosting or XGB.

Different variants have been modelled using different predictor variables and classification algorithms.

Their prediction efficiency is analyzed using three error metrics, and the computational time required to train and test the models is also evaluated. The chapter has presented some general results for all the models and special focus on four buildings is carried out. Thus, the methodology is applied to four buildings connected to a DHN in Tartu (Estonia) as a demonstration case. The optimal model is aimed for large-scale application among a large number of buildings of varied use. This application will be shown in the following chapter of this dissertation, proving that the method can be transferred to other buildings with different circumstances and the impact on economic savings will be quantified.

From the results in this chapter, the following conclusions could be drawn:

- The previous Q-T algorithm used 366 equations to characterize the hourly demand, based only on calendar variables. The use of unsupervised learning, by means of K-means clustering, drastically reduces the number of equations

needed to characterize the heat load profiles, enabling to group daily profiles by the shape or patterns of the demand. Thus, the number of equations is reduced to the number of clusters generated. In general, all the algorithms evaluated in this paper reduce the computational time required. The computation cost of MVLR models was around 90% lower than for the Q-T algorithm in all the cases.

- The methodology using cluster analysis improves the efficiency for prediction in building scale compared against Q-T algorithm. In the four buildings in which this study was focused, the R^2 value was improved and compared with the obtained in Q-T algorithm.
- Computation cost is a key variable for the operation of large DHN, where the hourly demand of several buildings is characterized. For high frequency predictions, such as hourly or sub-hourly forecasting, the response time of the model needs to be as fast as possible, so that the prediction of hundreds or thousands of buildings is feasible within the prediction interval. The Q-T algorithm enabled to discover the correlation between calendar variables and the instant demand, and it achieved remarkable prediction results. However, the large computation time of this model made necessary to analyze alternative ML models.
- The use of unsupervised learning before the application of the predictive algorithm enables to increase the predictive performance of the ML models in most of the cases. This efficiency gain, comparing use or not using clusters, ranges between 2% to around 50% (MVLR in Building 10949), always maintaining a reasonable computation time.
- When extending the study to the rest of machine-learning models, the XGB method is the one with highest prediction results, regardless the number of clusters and classifier model. Moreover, these accurate predictions are obtained with the lowest computation time among all the models simulated. XGB is followed, in terms of prediction accuracy, by SVR and RF, respectively.

- Among the ML models trained and analysed in this paper, the models developed using extreme gradient boosting algorithms reach the highest predictive performance in the three buildings. Moreover, apart from MVLR_1, XGB models (both XGB_1 and XGB_2) required the lowest computation time. Therefore, the multistep method presented in this paper using extreme gradient boosting as the predictive algorithm becomes a promising alternative to the most common operation algorithms used in the current DHN.

6. Referred Appendix

All these studies are summarized in the article that will be published by the author in ENERGY journal by ELSEVIER. The article was sent on the 11th of November of 2022 and it is still under review. The title of this article is: “Advanced Heat-Load Prediction Models in Buildings Combining Supervised & Unsupervised Learning”.

Chapter IX

Applicability of the Models

Abstract

In this chapter, the machine-learning models developed for the district-heating in Tartu (Estonia) are transferred and applied to two simulated districts in Bilbao (Spain). The objectives of this chapter are to analyze the efficiency of the models in softest climates and quantify the economic savings derived from a potential better forecasting accuracy. For the simulation of these two districts, DESING BUILDER simulations and a simplified model based on heating degree days will be used.

Resumen

En este capítulo, los modelos de aprendizaje automático desarrollados para la red de distrito en Tartu (Estonia) se transfieren y aplican a dos redes de distrito simuladas y localizadas en Bilbao (España). Los objetivos de este capítulo son analizar la eficiencia de los modelos en los climas más suaves y cuantificar los ahorros económicos derivados de una posible mejor precisión en la predicción. Para la simulación de estos dos distritos se utilizarán simulaciones de DESING BUILDER y un modelo simplificado basado en grados día de calefacción.

Chapter IX Applicability of the Models

1. Introduction

So far, the developed machine-learning algorithms have been applied to the same data retrieved from the real DH network located in Tartu (Estonia). This city is located in a humid continental climate with sever winters, classified as D_{fb} by Köppen-Geiger classification [84]. However, the main objective of this chapter is to study the applicability of the developed model in other DH networks with other boundary conditions. For this purpose, this chapter is going to apply these algorithms to two new DH networks located in Bilbao (Spain) with a completely different climate in order to study the efficiency of the heat demand prediction algorithms. The advanced algorithm including unsupervised and supervised learning developed and explained along Chapter VIII will be applied for the management of the energy production in these new networks. The efficiency of this advanced algorithm will be compared against the commonly used temperature-only management system. This commonly used algorithm will be used as the baseline for the quantification of the benefits of applying the advanced algorithm using unsupervised and supervised learning and theoretically obtaining better accuracy results in the prediction process.

DH network management is not an easy issue. Usually heat production in real DH networks is only based on the temperature prediction for the following hours. For example, when the temperature is expected to be reduced in the following hours, the demand is expected to reduce as well. It also depends on the heat production system and the inertia and flexibility of this plant to increase and decrease the instant heat production. For example, the ease to increase the demand in a medium-size gas boiler is not the same than the one that needs a large CHP (Combined Heat & Power plant) system in which the turbine requires some time to reach a steady and secure status.

Another variable to be considered in DH energy management is the size of the network. Thus, the energy generated for the network requires some time to reach all the buildings connected to the heating grid. A large distance from the production point(s) to the buildings increases the heat losses in the distribution pipelines and increases the time required for the hot water to reach the buildings (substations). Moreover, the distribution pipeline that forms the DH network itself also provides an additional thermal inertia that is characterized with the following equation.

$$C_{DH} = \sum_{i=1}^{N} \rho \cdot C_P \cdot L_i \cdot \pi \cdot D_i^2/4 \qquad \text{Eq. (25)}$$

Where, ρ is the density of the heat carrier fluid (usually, water), C_P is the specific heat of the fluid inside the pipeline, L is the length of the pipeline and D is the diameter of the pipeline. Therefore, the thermal inertia of the network could be used to satisfy small variations of the demand prediction.

To sum up, the management of a real network includes the analysis of many variables in the system and depends on the specific network to be managed.

2. Objectives of this Chapter

The objectives of this chapter are the followings, divided into main objectives and secondary objectives:

1. Prove that the developed methodology also works efficiently in other DH networks where the climate is not so extreme. Thus, DH networks in Bilbao (Spain) are proposed. The description of the climate in this location is described throughout the chapter.
    a. Evaluate the efficiency difference between the DH in Tartu and the DH in Bilbao. Note that the models were developed originally with data of Tartu´s buildings. It can be understood as another testing case study for the models.

b. Modelling two different districts, based on two different methods and evaluate the differences. Therefore, in this chapter, the models are applied against simulated (white-box) data.

2. Quantify the economic savings of implementing the new model for managing the demand in the network against the baseline method that is usually used nowadays.

3. Approach. General Methodology

The general method followed in this chapter is illustrated in Fig. IX-1

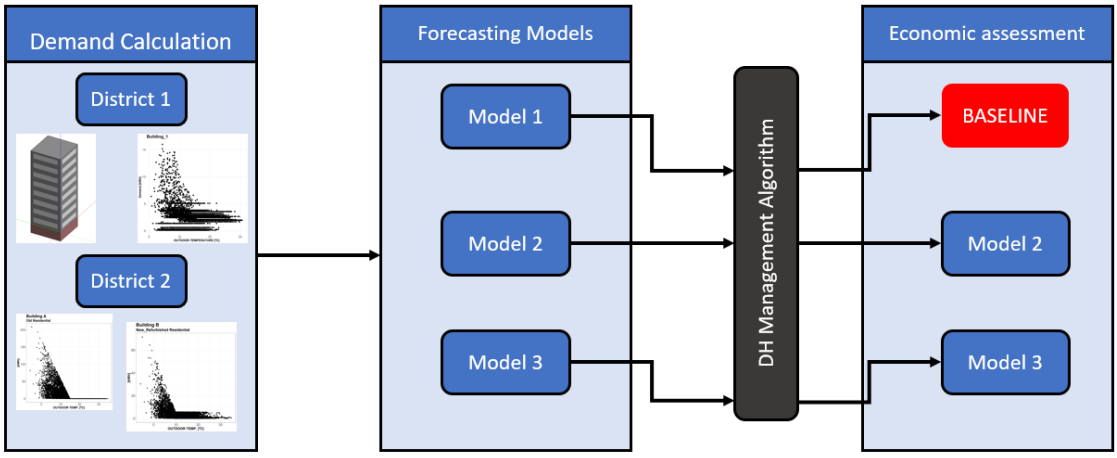


Fig. IX-1. General methodology followed in Chapter IX.

As it can be observed in that figure, the first step of this chapter will be the simulation of two "theoretical" districts, by means of simulating the demand in the buildings that conform the demand of the thermal grid. Then, a baseline scenario will be simulated, including the energy management of the district using a unique heating plant. The baseline scenario will include a common prediction algorithm using only the outdoor temperature and this scenario will be compared against different prediction algorithms using the same energy management strategy. This way, a comparison between the actual and commonly used energy management algorithm and the one using the self-developed prediction algorithm will be carried out.

The rest of the chapter is ordered as follows. First, Section 3.1 describes the buildings that will conform the DH networks in Bilbao and will analyze how these demands have been calculated. Then, Section 3.2 presents the three forecasting models analyzed, including the Baseline scenario and finally, Section 3.3 will outline the algorithm used for the management of the heat production in the network. As it will be explained, the heat production algorithm strategy will be the same regardless of the forecasting model used for the demand. Section 4 will present the results of this activity, following the same structure than Section 3. Finally, Section 5 will summarize the most relevant conclusions.

3.1. DH Networks Description

To achieve the objectives of this chapter, we will simulate two different networks:

1. District 1: Formed by 4 Buildings simulated using Design Builder [133].
2. District 2: Formed by 100 Buildings using a simplified method [134] for calculating the demand for SH + DHW.

As it has been commented, these two DH networks (District 1 & District 2) are supposed to be located in Bilbao, Spain. The location of the city is [43.26, -2.93].

According to Köppen-Geiger classification [84], this city is classified as Cfb or oceanic climate, also known as a marine climate, named as the humid temperate climate sub-type in Köppen classification, typical of west coasts in higher middle latitudes of continents, generally featuring cool summers and mild winters (for their latitude), with a relatively narrow annual temperature range and few extremes of temperature.

A typical climatic year is obtained from [135] with hourly frequency, including hourly temperature, solar irradiance and other climatic variables.

There is a significant difference between the climatic conditions in one city and the other. The heating demand in Tartu will be specially based on SH demand, whereas the demand for SH in Bilbao is not so relevant against the total demand. This is caused by

the cool minimum temperatures in this location. Nevertheless, in most of the winter, approximately from October to May, it is usual to have SH demand in Bilbao.

3.1.1. District 1

For this first district, we will simulate a small network where only four buildings are connected to it. The energy demand of these buildings is simulated using an often-used building simulator: Design Builder software [133]. The program works as a modular system integrated with a heat balance-based zone simulation with time-steps of less than an hour.

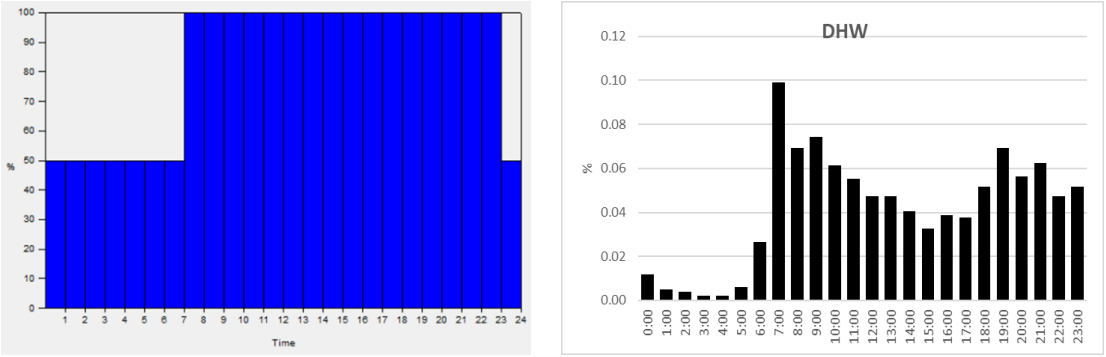
Regarding the use of these buildings, District 1 is formed by two residential buildings (Building_1 and Building_2) and two commercial buildings (Building_3 and Building_4) and the constructive characteristics used for these simulations are summarized in Table IX-1. These four buildings simulate two building typologies that are easy to find in this location. Whereas Building 1 and Building 3 are modelled as new buildings (from 2010 onwards), Building 2 and Building 4 represent construction in 1980s.

Table IX-1. Constructive Characteristics of the buildings in District 1

	Building_1	Building_2	Building_3	Building_4
Heated Surface [m^2]	800	432	1440	864
Windows to Wall Ratio	40	40	60	60
U_{window} [$W/(m^2K)$]	1.96	2.64	1.96	2.64
$U_{building}$ [$W/(m^2K)$]	1.03	2.09	1.30	2.26
DHW	Yes	Yes	No	No

Moreover, in order to define the SH and DHW demand profiles, two different profiles are used in function of the final use of the building. Therefore, different demand profiles are proposed for residential and commercial buildings. The following figures present

their corresponding DHW distribution and occupation during weekdays and weekends. Regarding the residential buildings, see Fig. IX-2:



(a) (b)

Fig. IX-2. (a) SH and (b) DHW profiles in residential buildings in District 1

And for the commercial buildings (See Fig. IX-3):

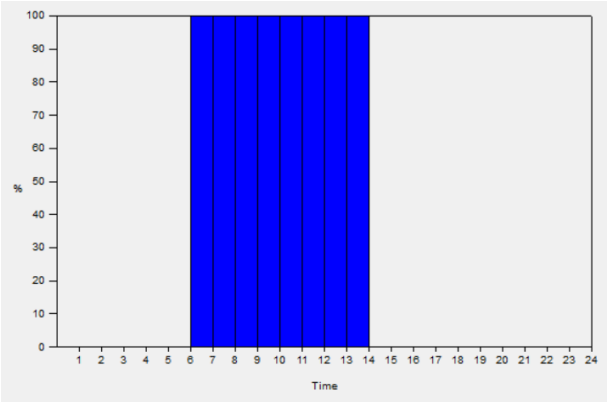
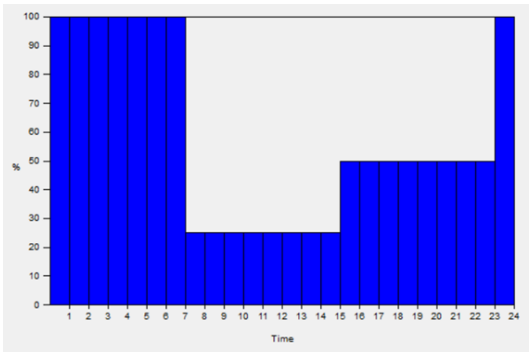
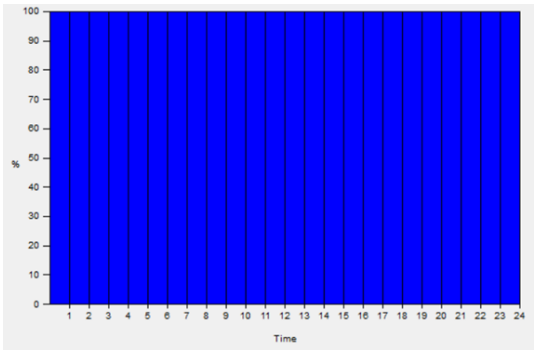


Fig. IX-3. SH profiles in commercial buildings in District 1

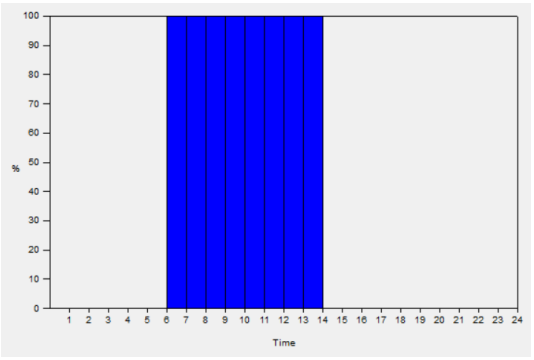
The occupational behavior inside the buildings is shown in the following images in Fig. IX-4:



(a)



(b)



(c)

Fig. IX-4. Occupation in (a) weekdays in residential buildings; (b) weekends in residential buildings and (c) weekdays and Saturdays in commercial buildings.

There is no occupation on Sundays in the commercial buildings. It is supposed that these buildings are going be unoccupied on holidays.

As there is no information of the size of the network, it has been considered 1km pipelines and 15% of heat losses, as it has been used in different references [134].

3.1.2. District 2

District 2 is supposed to be a small-to-medium size DH network, where 100 buildings are connected to the heating system. Among these buildings, three different buildings´ uses are included so that similarity to reality is provided to the system. Among these buildings, 88 buildings are devoted to residential purposes, two of them are educational

buildings and the rest of the buildings are commercial buildings[7]. As there is no information of the size of the network, it has been considered 10km pipelines and 15% of heat losses, as it has been used in based on previous literature references [134].

The useful areas of these buildings to be heated are simulated with a random function:

- **Residential buildings** will comprise heating areas between 100 and 2000 m^2, considering individual and multi-storey buildings.
- **Educational Buildings** will comprise heating areas between 2000 and 4000 m^2.
- **Commercial Buildings** will comprise heating areas between 100 and 1000 m^2. These building could be used as offices, shops, etc.

As for the thermal transmittance of the buildings, two types of buildings are included: (i) Old buildings with high thermal transmittance (4 W/m^2K) and (ii) New or refurbished building with a low thermal transmittance (1.2 W/ W/m^2K).

Regarding the demand to be covered in the buildings, in general, DH network will cover the demand for space-heating and domestic hot water. However, some of the buildings are supposed to include additional and independent heat sources to cover the demand for DHW: heat pumps or other sources. Moreover, some of the buildings do not have DHW demand in summer (from 15th July to 15th September).

Finally, in function of the type of the building use, a night setback is included from 23pm to 5am and from 6am to 8am. All this information is summarized in the following table (Table IX-2).

[7] Note that this distribution is arbitrary. This distribution is based on the buildings´ share of Tartu´s DH network.

Table IX-2. Buildings characteristics for demand calculation in District 2.

	Residential	Educational	Commercial
Number of Buildings	88	2	10
Thermal Transmittances [W/m^2·K]	1.2 & 4	1.2 & 4	1.2 & 4
Heated Areas [m^2]	[100-2000]	[2000-4000]	[100-1000]
SH Setbacks	23pm-5am	23pm-5am	23pm-5am
DHW by the DH network	78/88	1/2	10/10
DHW Summer	68/88	1/2	7/10

The SH demand in these buildings is obtained using the heating degree days method (HDD method) using the following equations that are obtained from [134].

$$SH_{Building} = U_{building} \cdot A_{building} \cdot HDD = U_{building} \cdot A_{building} \cdot (T_{HDD} - T_{OUT}) \quad \text{Eq. (26)}$$

$$DHW_{Building} = \frac{\frac{SH_{Year}}{1 - HW_{Y\%}} \cdot DHW_{Y\%}}{h} \quad \text{Eq. (27)}$$

For the calculation of HDD demand in buildings, T_{HDD} varies from old building to new or refurbished buildings. Whereas for old buildings where the $U_{building}$ is 4 W/m^{2}K, this temperature is set at 15°C, this temperature is set at 10°C for new or refurbished buildings. This temperature change simulates the setpoint difference between buildings.

Finally, and in order to minimize the linearity of the demand in the buildings and increase the realism of the simulations, the instant hourly demand obtained from all the steps above is multiplied by a random factor from -50% to 50% (PF), following the next equation:

$$Q_{hour} = Q_{hour} + (Q_{hour} - Mean(Q) \cdot (1 + \frac{PF}{100}) \quad \text{Eq. (28)}$$

3.2. Demand Prediction

Previous paragraphs have shown the two methods for calculating the energy demand in buildings, shaping the demand of two districts. Using the calculated demand in these districts, part of the data will be used to train the models and the other set of the data will be used to study the efficiency of these algorithms. Three different models will be tested to predict the demand:

Model 1: Temperature based algorithm (Baseline)
Model 2: The Q-T Algorithm explained in Chapter V [2]
Model 3: The advanced ML model developed in Chapter VIII.

The temperature-based algorithm (Model 1) is commonly used for the prediction of energy demand in districts, and it is based on a simple linear regression against climatic variables in the location. Since the demands to be predicted are simulated (not real energy demands), the Temperature (Model 1) based algorithm is supposed to perform better than with real conditions. Focusing on Model 3, in this chapter the model that obtained best prediction results in Chapter VIII will be applied. Therefore, a multi-step method combining K-means for identifying patterns, K-NN for classification of the patterns and Extreme Gradient Boosting (XGB) for forecasting the demand will be used.

3.3. Heat Production Management Algorithm

This final section of the methodology starts with the predicted demand in the buildings and embrace the heat production process for the district. The objective of a DH network is to supply enough energy to all the buildings every moment. Thus, the predicted demand must be supplied from the network to the buildings. In this context and for this case study we are going to simulate the simplest case in which the energy is produced in a unique generation plant. Thus, it is not necessary to define priorities among different generation plants. In cases where there is more than one heat generation

plant, we need to define which production mix is the optimal every moment. In those cases, we need to consider the following factors:

- Energy **generation type**: The size of the plant and the type of technology used for heat generation will determine the role of this plan in the generation-mix: base production plants, medium production plants and peak production systems. Small and flexible generation plans are usually considered as peak producers (for example, small solar thermal plants of small gas boilers). On the other hand, technologies such as large CHP plants are usually more appropriate to fulfill the base demand.
- Energy **production price**: The instant heat generation costs will determine which of the generation plant(s), among the same type of plants, would supply heat to the network. So that the minimal production cost is ensured every moment.

In the case study presented in this chapter, a unique generation plant has been considered so that there is no need to couple the production with other generation plants. This task is usually cost-based and does involve market study.

For our case study, a medium size non-condensing gas boiler has been considered. This boiler will be different in District 1 and District 2 since the size of the networks is different. The Heating Only Boiler (HOB) in District 1 is sized as a 120 KW condensing boiler. In the second one, since the maximum energy demand in one hour is slightly above 12 MWh, the power of this gas boiler has been considered to be 13MW with a nominal efficiency ($\eta_{nominal}$) of 98%. In other words, 1MWh of gas energy is converted into 0.98MWh of useful heat for the network.

The nominal efficiency usually decreases when the outlet temperature increases and when the inlet temperature decreases.

This case study is just a theoretical simulation of a network in order to compare the efficiency of demand prediction algorithms. So, the gas boiler is supposed to work in

partial load almost all the time. Almost all the energy production technologies reduce their efficiency when working at partial load, so in terms of efficiency we prefer the system to work full load every moment. However, this case will generate an excess heat supply to the network that must be disposed or expelled in an energy storage system. It is known that the boiler would stop automatically when there is overheating in the supply temperature of the DH network and the real economic savings will be lower than the resulting from this analysis.

For the specific case of gas boilers, the evolution of partial load efficiency reduction factor ($F_{thermal}$) is shown in Fig. IX-5 and the equation that determines the production efficiency of this gas boilers is presented in the following equation:

$$\eta_{thermal} = \eta_{nominal} \cdot F_{thermal} \qquad \text{Eq. (29)}$$

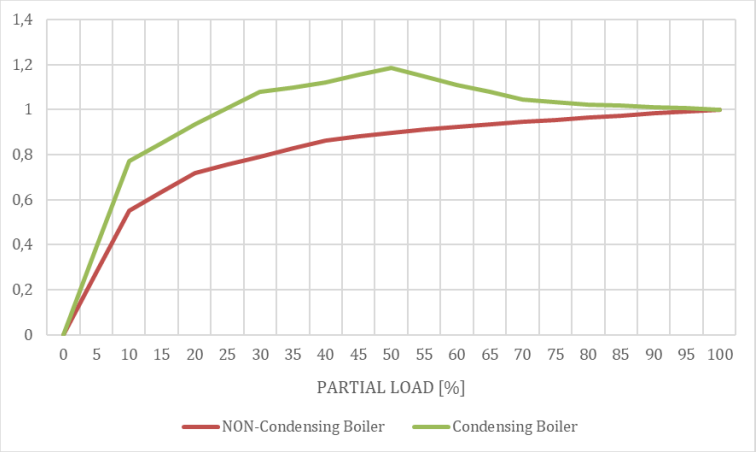


Fig. IX-5. Efficiency reduction factor of thermal efficiency in gas boilers.

All the pipelines that form the heating network contributes with a thermal inertia that could be used to balance small demand variations. All in all, the DH management strategy is based on the following points.

In the first hour of the simulated heat production, the energy produced by the gas boiler is the same than the demand plus heat losses. The production algorithm used for the three predictions algorithms is the same and it is based on the *variability* concept:

$$Variability\ (t) = Demand\ (t+1) - Demand\ (t) \qquad \text{Eq. (30)}$$

Where *t* refers to a specific hour.

Thus, the production is governed by the following points:

- If *Variability(t) > 0* ← The demand for the next hour is supposed to increase. We may anticipate that demand increase by increasing the production.
- If *Variability(t) < 0* ← The demand for the next hour is supposed to decrease. We van anticipate that demand decrease by decreasing the instant production. The rest of the instant demand is covered by the thermal inertia of the network.
- If *Variability(t) = 0* ← The demand for the next is the same. We continue to produce using the same strategy than this hour.

This variability concept enables the production system to anticipate the trend of the demand and supply the exact energy in the correct moment. If the heat source would have been a CHP (or another big installation), the variability should have been calculated with more than one hour frequency. Gas boilers can change their working regime varying the operating frequency.

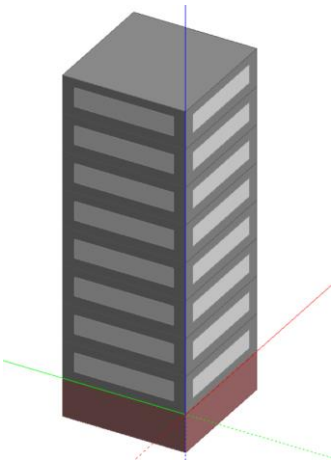
4. Results

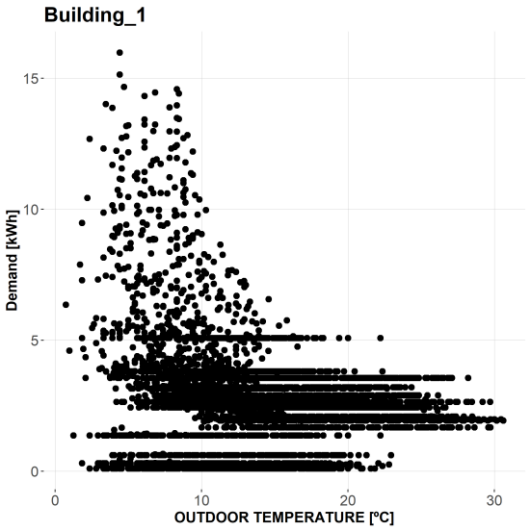
This section will follow the same structure than methodology section, so that the tracking of the results is connected to the explained methods. Therefore, this section will firstly present the demand of the buildings obtained for the two calculation methods. Then, the predictive accuracy of the forecasting models will be analyzed to finalize with the economic benefit/assessment of each district using the management algorithm explained in Section 3.3.

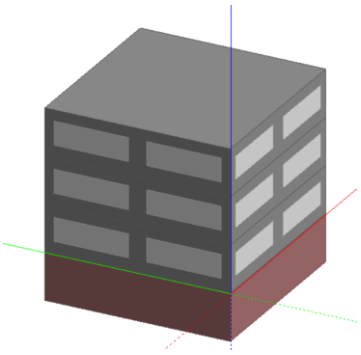
4.1. Districts´ Description

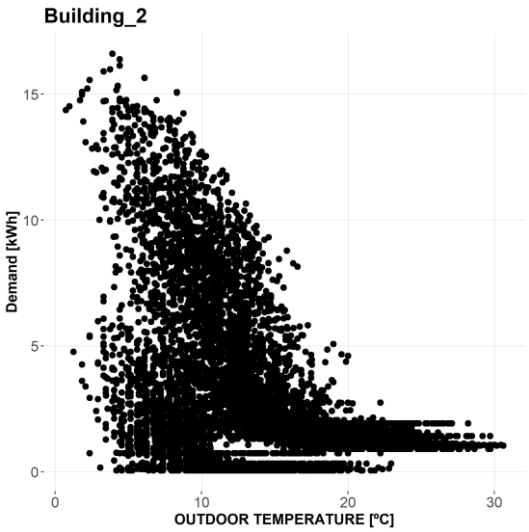
4.1.1. District 1

For the calculation of the demand in the four buildings, Design Builder was used with the parameters defined in Section 3.1.1. The setpoint temperature for the SH demand was defined at 20°C in all the buildings. The following images (Fig. IX-6) show the general dimension of the building accompanied by the total demand in that building. The heating demand corresponds with the total energy requirements against the outdoor temperature.









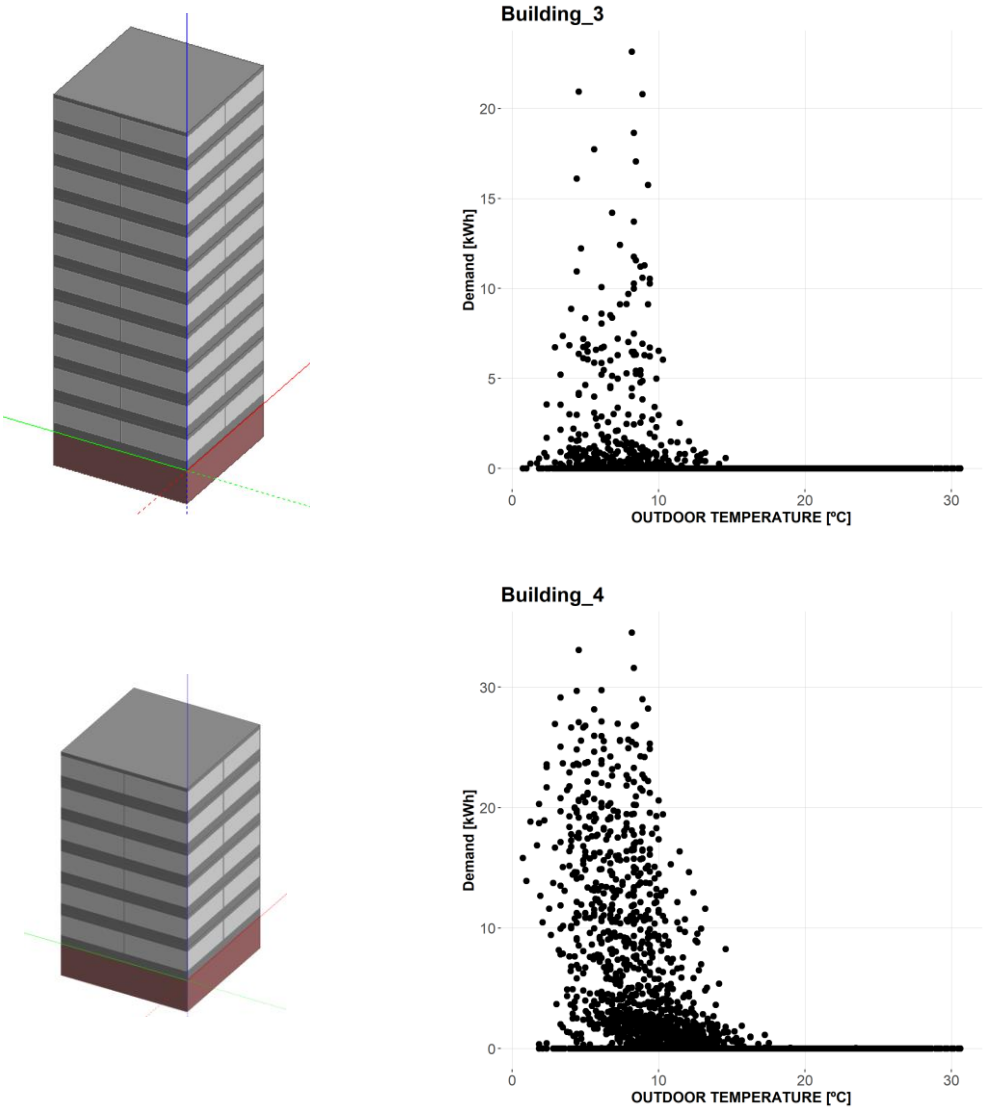


Fig. IX-6. From top to the bottom: Energy Demand (SH+DHW) of Building_1, Building_2, Building_3 and Building_4 of District_1

As it can be observed from Fig. IX-6, while Building_1 and Building_2 (residential dwellings) do present demand for DHW and SH, commercial buildings only require SH demand. This is why in moment with high outdoor temperature, the total demand in those building is zero.

4.1.2. District 2

On the other hand, District_2 is formed by 100 buildings and consequently it is not possible to present the results in all the buildings. As there are different building

typologies (see Table IX-2), this chapter will present results for some of the different typologies.

- **Building A:** Residential Building that is built under old CTE specifications (U = 4 W/m^{2}K) with no DHW demand from the network. See Fig. IX-7a.
- **Building B:** New or refurbished Building (U = 1.2 W/m^{2}K). See Fig. IX-7b.
- **Building C:** Educational Building. See Fig. IX-7c.

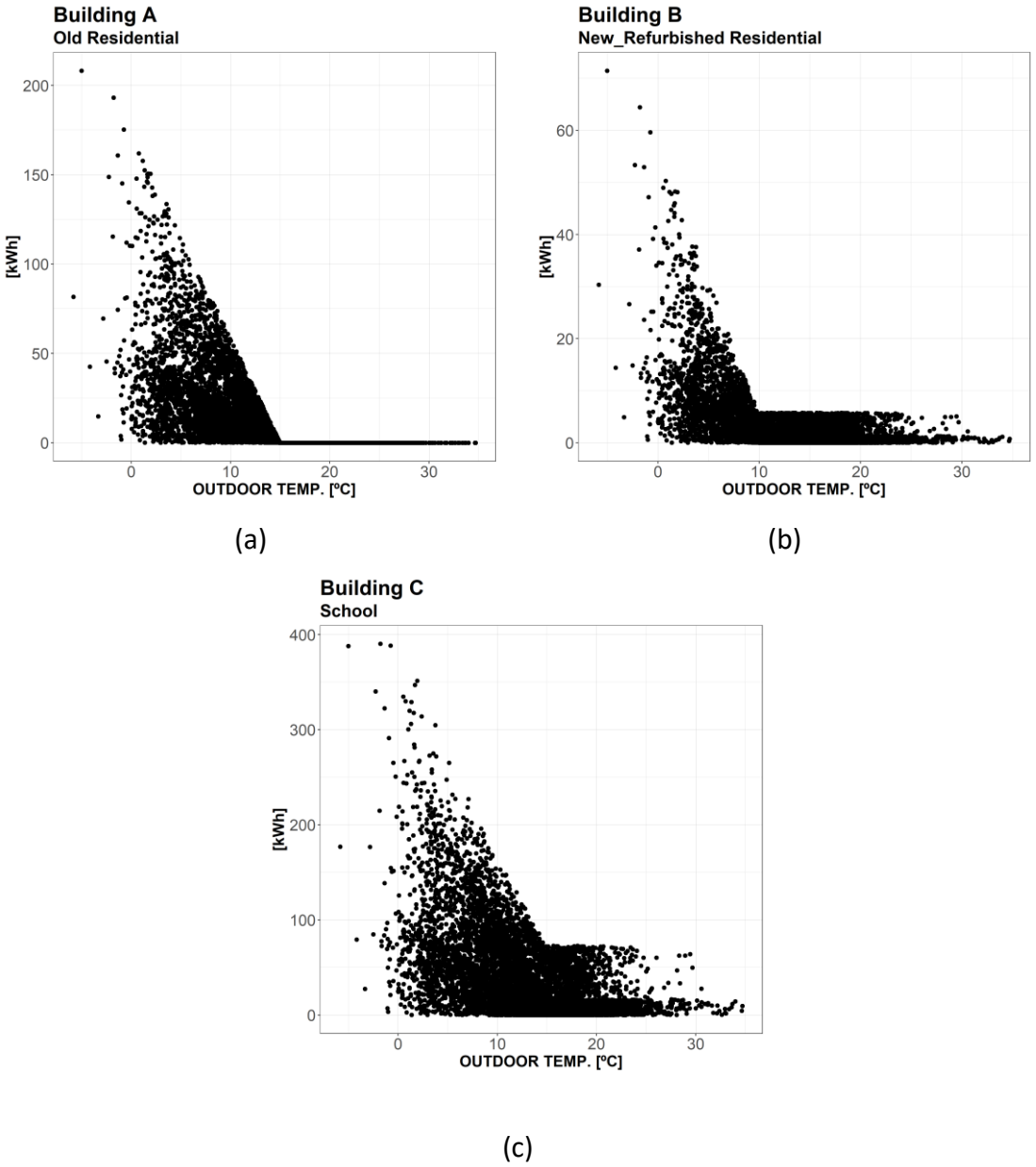


Fig. IX-7. Heating demand against outdoor temperature for (a) Building A: Old Residential; (b) Building B: New Building and (c) Building C: School

Fig. IX-7 shows the difference between the heating demands in the three buildings selected for the study. While Building B is modelled with a low total transmittance ($U_{BUILDING\_B}$ = 1.2 W/m^{2}K), Building A and Building C are modelled with a higher transmittance. Thus, the heating demand density (kWh/m^2) in Building B is lower than the rest of the buildings. The same would happen with the cooling demand, however, this is not part of the study since the network is only supposed to cover heating

demands: space heating and domestic hot water. Building A does not present DHW demand as it is observed on the right side of Fig. IX-7a. This is caused because the building may have another heat source only for DHW demand. For every building is observed that SH demand presents, at least, two curves corresponding to night setback and "normal" heating. Finally, comparing Building B and Building C, the DHW demand in Building C is higher than Building B. The reason for that phenomenon is that this demand is simulated as a percentage of the SH demand. Thus, when increasing the total SH demand, DHW is also increased in the percentage.

Additionally, the total energy demand in District 2 is shown in Fig. IX-8.

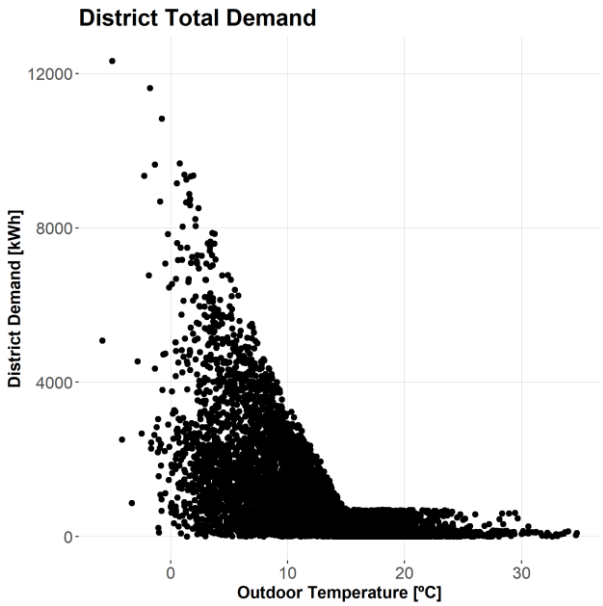


Fig. IX-8. Total hourly demand against outdoor temperature in District 2

4.2. Demand Forecasting Results

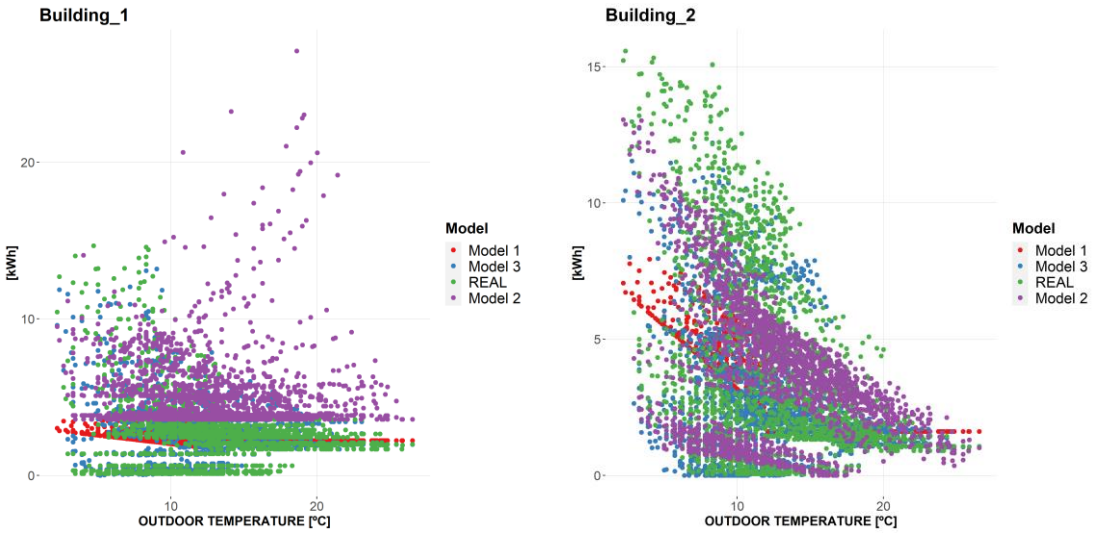
The next step in this chapter is to analyze the prediction accuracy of the energy forecasting models. In this chapter, the forecasting results for the different models are compared against the Baseline Scenario using error metrics. Then, Section 4.3 will translate this prediction accuracy difference to economic benefits for implementing the advanced model or Model 3 in this chapter.

The aim of this chapter is to develop and use the ML model to manage a DH network. Consequently, we need some data to train the model and the other part of the data to test the results. In this case, we used 75% of the data to train the model and the 25% for testing. Since the data follows a time-based structure, we used the first 274 days of the data (75% of 365) for training the models and the rest of the days in the year to test the efficiency of the model.

4.2.1. District 1

This district is supposed to be composed by four buildings described in 3.1.1. The first step for the energy management of a district is to obtain the demand forecasting for the following hours, so that the heat production can match the expected demand in the district every moment. In this paragraph we are going to present the forecasting results obtained with three models for the four buildings independently.

Fig. IX-9 shows the real demand (green points) and the prediction results for Model 1, Model 2 and Model 3 against the outdoor temperature. It is shown for the four buildings in the network.



(a) (b)

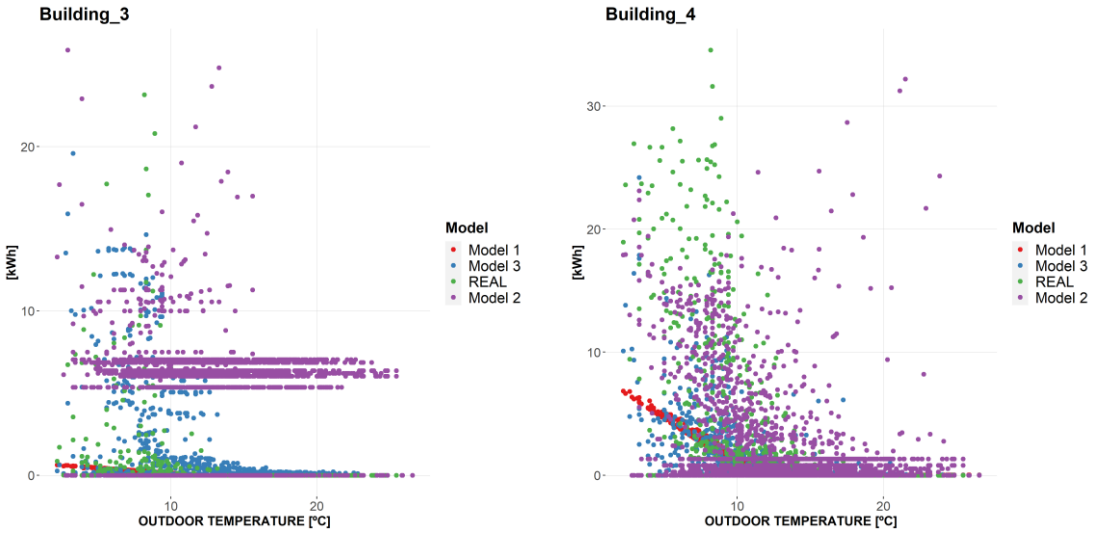


Fig. IX-9. Heating demand Forecasting results against TOUT in (a) Building_1, (b) Building_2, (c) Building_3 and (d) Building_4.

From Fig. IX-9 is observed that Q-T algorithm (or Model 2 in this section) only obtains accurate predictions for Building_2 and Building_4, while the errors in the other buildings are very relevant. Due to the very low linearity of the demand resulting from the simulations in Design Builder, the Model 1 is not capable of accurately characterizing the demand and the unique model that is able to characterize the demand is our Model 3.

Additionally, the MAPE values obtained in each building are summarized in Table IX-3. The results shown in this table confirms that Model 3 is the best among these models.

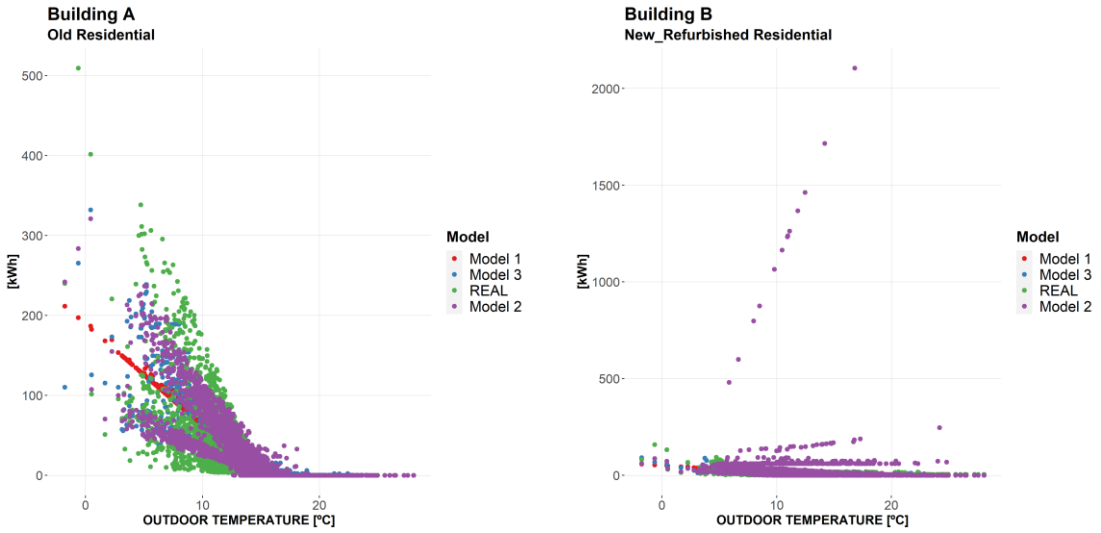
Table IX-3. MAPE values [%] for the predictions in the four buildings of District 1.

MAPE [%]	Building_1	Building_2	Building_3	Building_4
Model 1	2.969	2.901	NA	NA
Model 2	5.694	0.897	NA	NA
Model 3	0.361	0.734	NA	NA

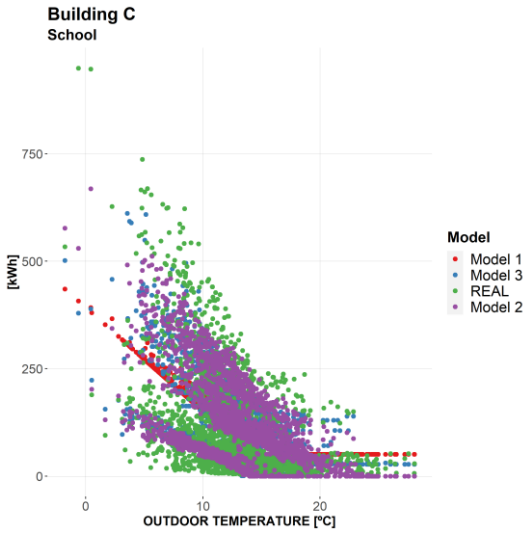
Model 2 (or the Q-T algorithm) presents problems when the buildings have low transmittance values and consequently low SH demand. This model is built in order to identify the difference between SH and DHW demand. When the demand for SH is relatively low and usually matches the DHW demand, the model is not able to differently characterize three two demands and consequently, results on horizontal (Q_{REF}) results even in moments with SH demand.

4.2.2. District 2

On the other hand, this network connects 100 buildings previously described in Table IX-2. As it not viable to show the prediction results for all the buildings, this section will only be focused on the three buildings analyzed in the previous chapter, so that the demand of all the district will be analyzed in the paragraph for the economic assessment. Therefore, the following figure (Fig. IX-10) shows the demand (green points) and the prediction results for Model 1, Model 2 and Model 3 against the outdoor temperature. This figure shows the results for the three particular buildings previously analyzed.



(a) (b)



(c)

Fig. IX-10. Predictions and real demand against outdoor temperature for (a) Building A, (b) Building B and (c) Building C.

Q-T algorithm´s accuracy in Building B is very low and does not allow to analyze the efficiency of the rest of the models. Consequently, Fig. IX-11 shows the same predictions but removing the results obtained with this model (Model 2).

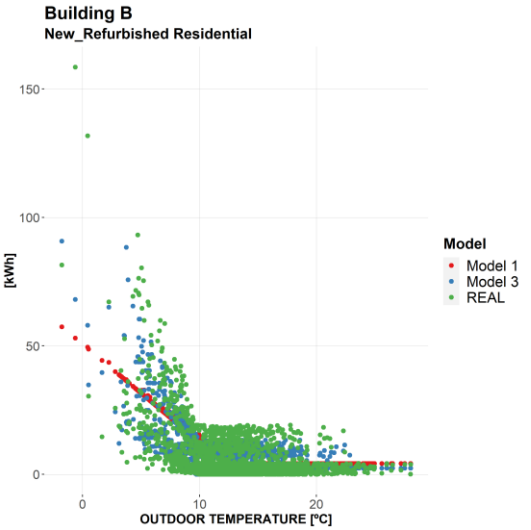
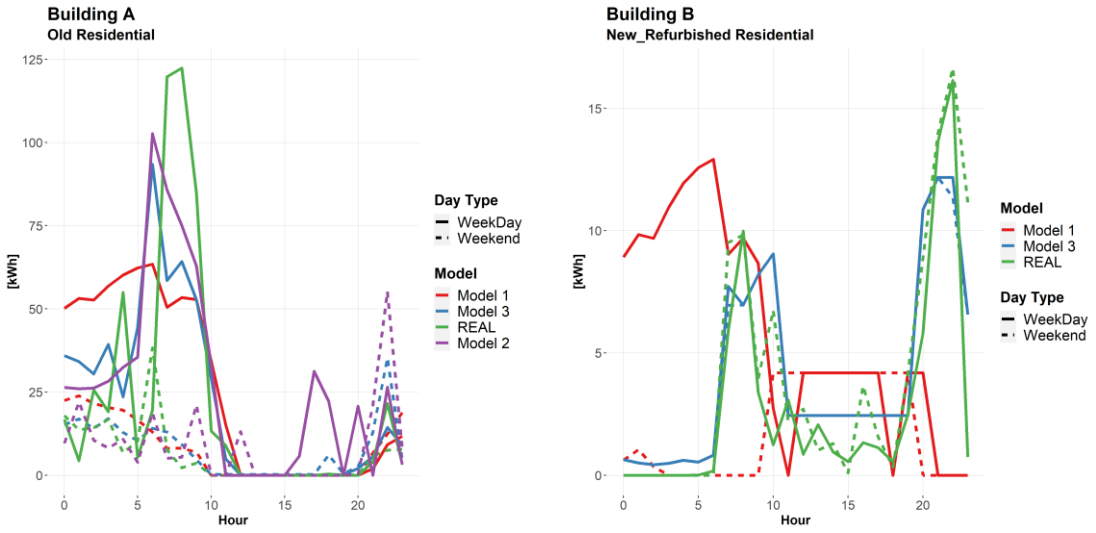


Fig. IX-11. Predictions and real demand against outdoor temperature without Q-T algorithm (Model 2) in Building B

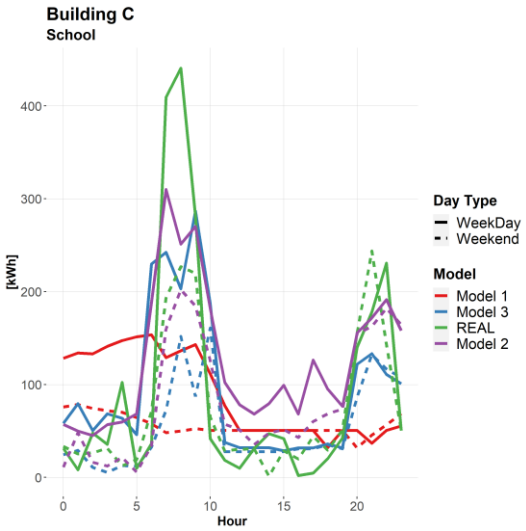
Once the heating profiles are characterized, Fig. IX-12 presents the results for the demand prediction in building scale. The figures show two images for each building. The predictions obtained from Q-T algorithm present large errors, and it does not allow to evaluate the efficiency of the Advanced Model. This is why that in each building, the figure of the right presents the three models and the figure on the right only show the results for the temperature-based model and the advanced model.

Fig. IX-12 shows the predictions profiles for two types of day:

- **Weekend**: This day corresponds with the 15th of December, Sunday and not business day.
- **Weekday**: This day corresponds with the 17th of December, that in this case corresponds with a Tuesday and business day.



(a) (b)



(c)

Fig. IX-12. Heating demand profiles and predictions for (a) Building A: Old Residential; (b) Building B: New Building and (c) Building C: School

In general, the results from the predictions in the building scale show that the Advanced Model (Model 3) is the one with highest prediction accuracy. Q-T algorithms show bad results, especially for high temperature hours. On the other hand, the temperature-based algorithm is able to follow the trend of the demand, but it shows two main problems: (i) it is no able to model the peak demands of the morning and (ii) it overpredicts when the demand is low in the mid-afternoon hours.

4.3. Economic Assessment

Finally, this section will translate the forecasting results to the energy management task and consequently, we will analyze the effect of increasing or reducing energy forecasting accuracy on the economic assessment of the overall system. Therefore, we will use the Model 1 as the baseline for all the calculations and the economic savings resulting from a more accurate energy forecasting will be presented. The energy management algorithm used is the method explained in Section 3.3. Following the same structure than the other sections of the chapter, we will divide the results by the two districts modelled for this purpose.

4.3.1. District 1

Despite being a small DH network, the total energy demand is a key factor for sizing the heat production unit and for the calculations of the energy management. For this reason, Fig. IX-13 presents the predictions results and the real demand of the four building that form the district against the outdoor temperature.

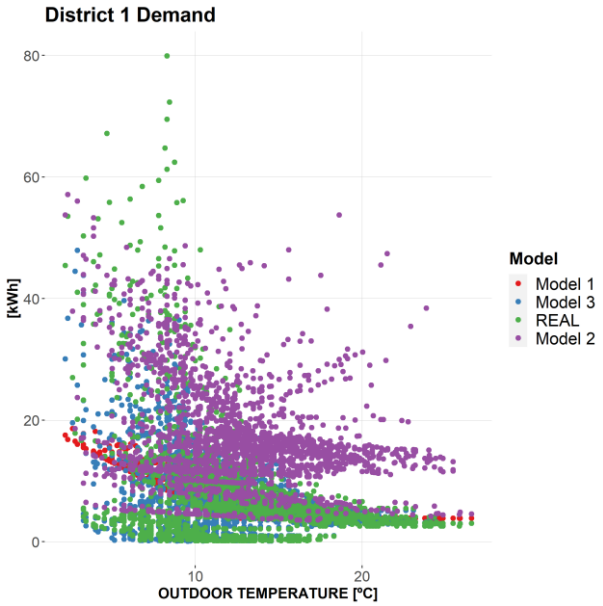
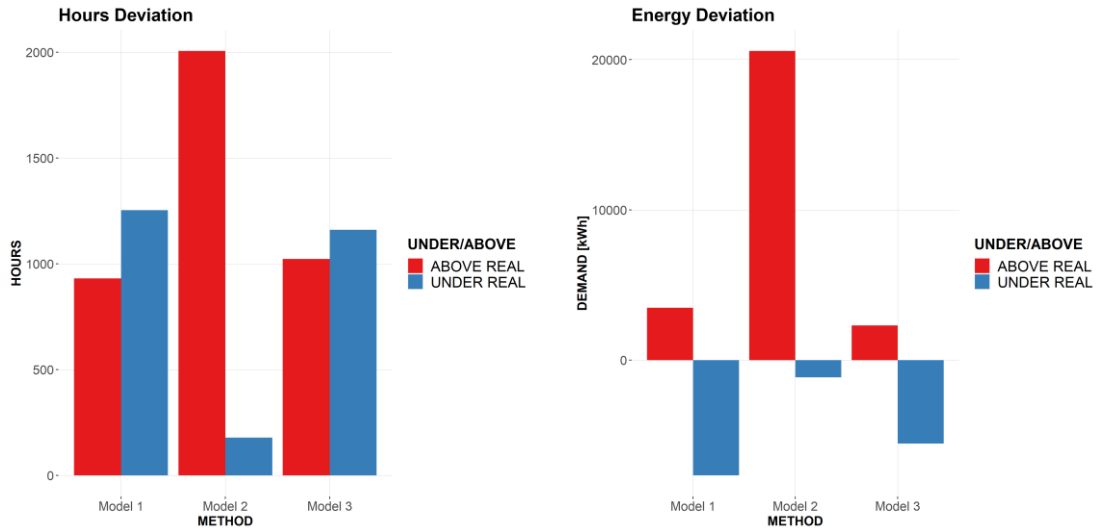


Fig. IX-13. Total demand prediction against outdoor temperature in District 1.

From previous figure is observed that Model 3 is the most accurate one. When showing the sum of the demand of the district, the inaccuracies shown by the Q-T algorithm (or Model 2 in this chapter) are "hidden", although it presents low accuracy for low demand zone. The baseline is only capable to predict low-demand moments with certain accuracy and big differences can be observed between real peak demands and the energy forecasted by the Model 1.

For the management of the network and to ensure the thermal comfort inside the buildings, it is important to analyze the hourly matching between the demand and the energy that is transferred to the buildings. Therefore, Fig. IX-14a analyzes the number of hours when the energy forecasting of the district is above and under the real demand,

while Fig. IX-14b presents the same results but translated to energy units. Therefore, Fig. IX-14b quantifies the energy overproduction and the underproduction moments.



(a) (b)

Fig. IX-14. (a) Number of hours deviations and (b) energy deviation in District 1

When observing Fig. IX-14a, there are few differences between Model 1 and Model 3, while Model 2 overpredicts more than the other models. Comparing the Baseline (Model 1) with our advanced model, Model 3 shows some more time with overprediction and less time under the real demand. This difference is not noticeable in energy units. Thus, Model 3 presents better results in all the cases. Whereas it enables to reduce the overproduced energy, it is also capable to reduce the underproduction moments. This is positive for energy management, since there is no necessity to include additional heating sources (or additional storage) to reach peak demands and this energy difference could be satisfied by the thermal inertia of the network.

Finally, the energy production costs of the network are shown in Fig. IX-15.

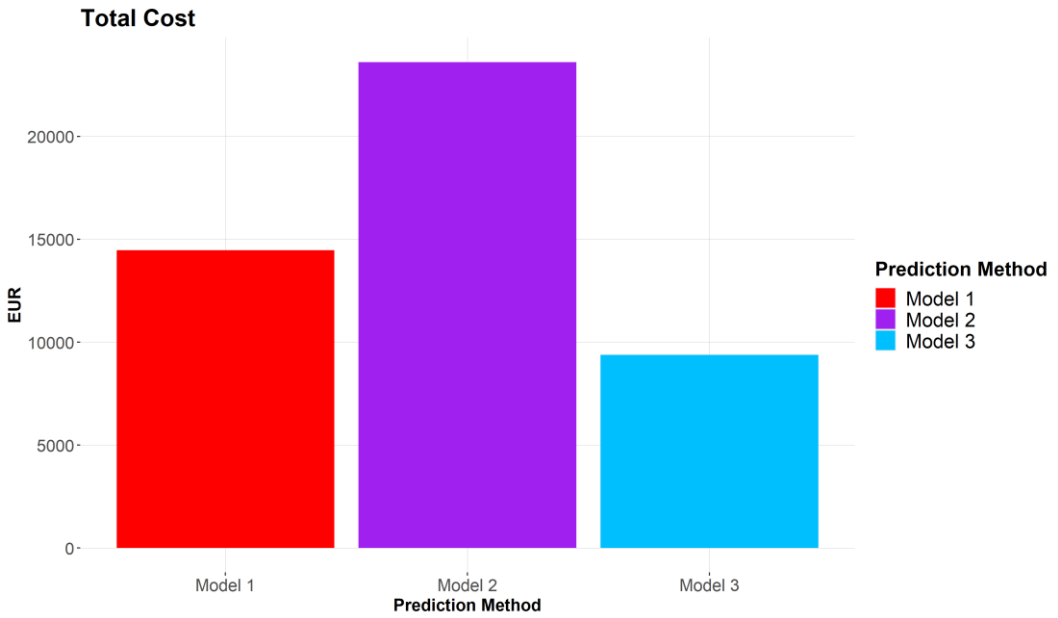


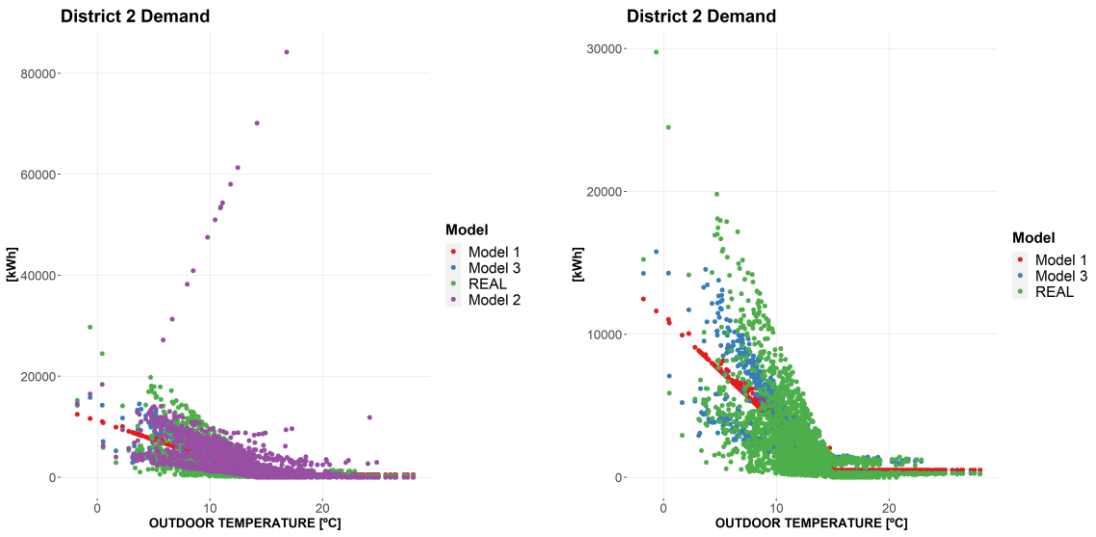
Fig. IX-15. Cost Comparison between forecasting Models in District 1

Before analyzing the results in Fig. IX-15, it is important to remind that the algorithm used for energy management in the production unit is the same for all the models. The costs that are shown in Fig. IX-15 and the way they are calculated are explained in section 3.3 and consists in the natural gas costs for fueling the boiler. The costs for maintenance or initial investment are not included since they will be very similar in all the cases. Undoubtedly, the results must be interpreted quantitatively, since the algorithm used for energy management is a simplified version of real methods.

The biggest difference is observed between Q-T algorithm (Model 2) and the ML model (Model 3). While the costs for the 25% of the year in Model 2 reach 23612.80 euros, Model 3 is capable to reduce these costs below 10000 euros. Moreover, the difference between Baseline (Model 1) and Model 3 is around 5000 euros.

4.3.2. District 2

On the other hand, the district simulated by the simplified method explained before is much bigger than District 1 and it is expected that the difference in the results will also be large. First, and following same results´ structure in both networks, Fig. IX-16 presents the prediction results for all the district.

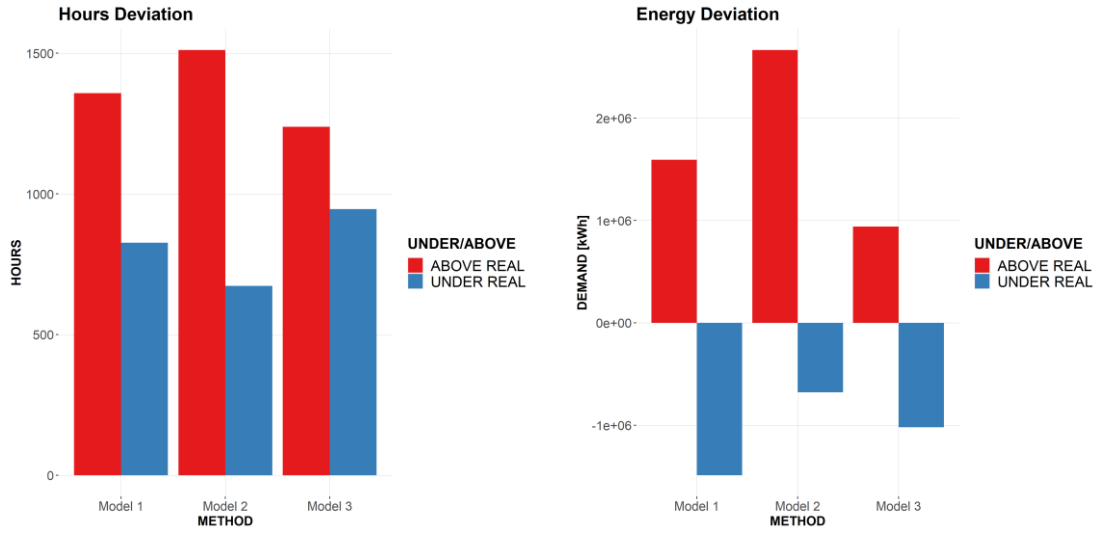


(a) (b)

Fig. IX-16. Total demand prediction against outdoor temperature in District 2.

Previous figure is divided into results with and without Model 2, due to its low accuracy in some moments. This low accuracy of Q-T algorithm is caused by the high variability included in the demand simulation and the difficulty to calculate the QREF of the model in such cases. Thus, Fig. IX-16b presents the same demand predictions but removing the predictions from the Q-T algorithm. This figure enables to visualize the real demand (simulation) in green and the predictions of Model 1 (temperature-based) in red and Model 3 (advanced) in blue. Since the Baseline algorithm is a multi-section linear regression, the model cannot follow the great variability included in the demand, especially in cold moments (high SH demand). As it was expected, the best prediction accuracy is obtained using the advance model (Model 3), combining supervised and unsupervised learning techniques.

Besides, Fig. IX-17 presents the dispersion of the predictions against the real demand. Therefore, Fig. IX-17a shows the number of hours in which the forecasting results are above and under the real demand, while Fig. IX-17a quantifies the overestimated and the underestimated energy.



(a) (b)

Fig. IX-17. (a) Number of hours deviations and (b) energy deviation in District 2.

In this district, Q-T algorithm (Model 2) is the model with lowest number of prediction hours under the real demand and the model with lowest energy under real demand. However, the opposite occurs for energy above the real demand, and it becomes the model with the highest number of hours and energy over the real demand. Thus, it can be concluded that Q-T algorithm (or Model 2) has an overestimation trend in most cases. Regarding Model 3, this model is capable to reduce both, the number of hours below and above the real demand, resulting in the highest accuracy of the energy forecasting in the district.

These forecasting results are then transferred to an economic study of the costs regarding the energy production in function of the estimations made by the three models. In this line, Fig. IX-18 shows a comparison between the total costs associated to District 2. Note that these costs only represent the 25% of the year since the model is tested against 25% of the days.

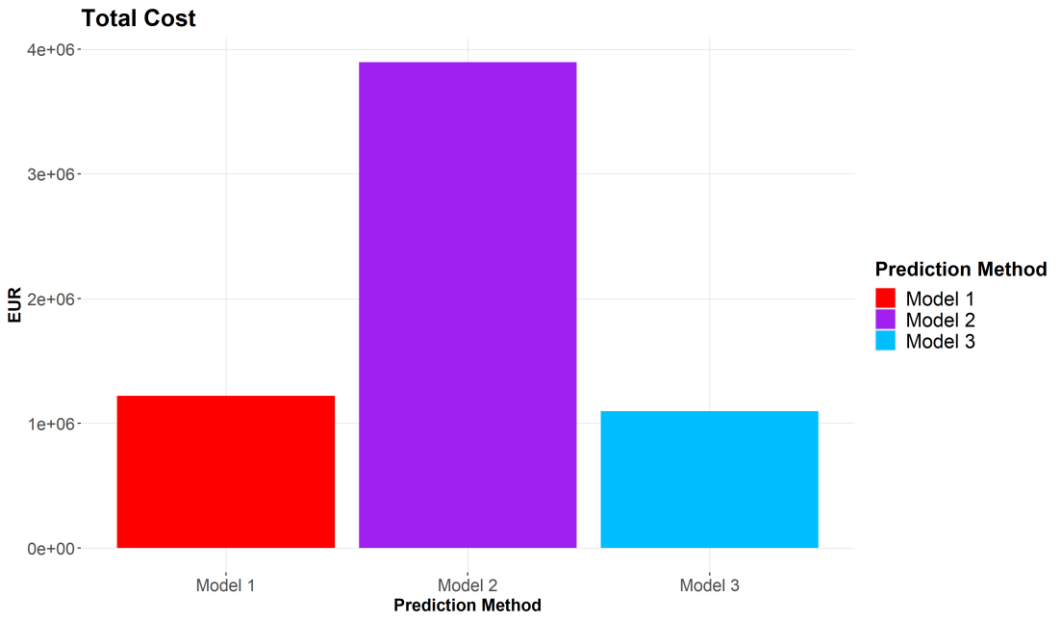


Fig. IX-18. Cost Comparison between forecasting Models in District 2.

District 2 presents similar trend than in the analogous figure for District 1 regarding energy costs. Since the size of the network is much bigger than District 1, the total costs associated to the network will also increase. Nevertheless, Model 2 requires the maximum costs since it is the model with highest overprediction. In this case the energy productions costs reach $3.9 \cdot 10^6$ €. On the other side, Model 3 requires $1.09 \cdot 10^6$ € and Model 1 requires $1.22 \cdot 10^6$ €. Thus, Model 3 reaches an economic savings of 11% compared with Baseline Scenario, which is a very significant value, regarding that the unique difference between the scenarios is the forecasting accuracy of the models used for this purpose.

5. Discussion & Conclusions

This report is focused on analyzing the results from the application of the self-developed algorithm for the management of the demand in a whole district-heating network. The so-called Model 3 throughout this chapter. The rest of the chapters were focused on analyzing and developing the optimal solution for each of the steps in the algorithm. The developed algorithm was optimized using real data from a DH network in Tartu. In these

chapters it was confirmed that the developed model was able to obtain high prediction accuracy, with R^2 values above 0.9 in certain buildings in hourly frequency predictions.

In this chapter two new districts have been modelled using firstly a district modelled using DESING BUILDER and secondly, using a simplified method based on Heating Degree Days theory. The simulated districts are conformed by 4 and 100 different buildings, respectively, including residential, commercial, and educational buildings. The districts are now located in Bilbao (Spain), which corresponds with a C_{fB} classification (oceanic climate) in the Köppen-Geiger [84] climatic classification. Consequently, the climatic severity of the location is also changed in order to study this effect on the models, since the unique predictors used predictive purposes are climatic and calendar-based variables.

For the energy management of these networks and with the objective of simplifying the energy production process, a centralized energy production system has been modelled in which all the energy for the district is produced by a single heating only boiler (HOB). It is true that the centralized model opposes the philosophy of modern district-heating network in which the energy is produced in small, decentralized and low-grade energy sources, such as ST energy or waste heat streams. However, for the comparison of the predictive efficiency of the self-developed model and the currently used prediction models, the energy production system is not relevant.

To avoid lengthening this chapter, the following points will summarize the most important conclusions from the application of the self-developed algorithm to manage the energy production in a district-heating network:

- Buildings modelled by DESING BUILDER software (District 1) turn out to lower demands than buildings modelled by the simplified method. This demand reduction is caused because the software includes the solar gains and other thermal gains inside the buildings, while the simplified method only models the demand using climatic conditions (HDD).

- Contrastingly to the first point, the dispersion of the demand by DESING BUILDER are higher than in District 2, even though including the PF factor for increasing the randomness of the demand.
- In general, predictions in individual buildings obtain better accuracy results in District 2 than in District 1, caused by the higher linearity and lower dispersion of the demands.
- Q-T algorithm overpredicts in every building and it is not capable to get accurate predictive results. Thus, this model is not considered for the comparison with the self-developed algorithm that combined supervised and unsupervised machine-learning models.
- In comparison with the results obtained in the DH in Tartu, both models, Model 2 (or Q-T algorithm) and Model 3, obtain lower accuracy in the district located in Bilbao. The milder climate with cooler temperatures impacts on the SH demand and reduced the correlation between the climatic conditions and the real demand.
- Model 3 is able to reduce the hours in which the model overpredicts the real demand compared with the Baseline in District 1. Nevertheless, the opposite trend is observed in District 2.
- In terms of energy, the advanced model presents less energy over-produced and less energy underproduced in comparison with the temperature-based model. This means that the prediction accuracy of this models is better. This happens in both districts.
- A better prediction accuracy results in economic savings in the energy production process. In this case, the economic savings reach the 10% in comparison with the baseline. The energy production mix and the type of energy produced in the system will determine the exact savings in each district-heating, but a better predictive accuracy of the demand will always ensure better economic performance in the network. Some energy sources (such as solar thermal energy or other process-based heats streams) produce energy

continuously and regardless of the current demand in the district. In these cases, it is necessary to design a dynamic heat storage system that controls the energy flows. The better the demand is predicted; the better optimization of the heat storage size can be obtained. In this type of districts, it will be necessary to include an additional peak energy plant that fulfills the demand not covered by the renewable sources.

- The boiler control system will not produce more heat than the one is supplied to the buildings plus the losses. Otherwise, the DH grid would be overheated until evaporation of the water would occur and the grid would be broken. Imagine we introduce to the grid all the energy of Model 2. Consumption + losses consume what Model 3 says. The difference of the heat injected to the grid would be used to overheat the water within the tubes of the grid until it would be evaporated. The boiler will not allow it.

Chapter X

Conclusions, Contributions and Future Work

Abstract

This last chapter gathers the main conclusion and contributions of the PhD Thesis. Additionally, it presents the potential future research lines that derive from this dissertation and summarizes the dissemination activities, including international journals or conferences, carried out for diffusion of the results.

Resumen

Este último capítulo recoge las principales conclusiones y aportaciones de la Tesis Doctoral. Además, presenta las posibles líneas de investigación futuras que se derivan de esta tesis y resume las actividades de divulgación, incluidas revistas o congresos internacionales, realizadas para la difusión de los resultados.

Chapter X Conclusions, Contributions and Future Work

This final chapter of the dissertation presents a brief summary of the contributions that these studies have made, showing the overall conclusions of the work and analyzing the potential directions of the works that the dissertation has led.

Therefore, this chapter contains the following sections:

- **Main Contributions & General Conclusions:** Main Contributions of the works concerning the current state of the art.
- **Dissemination/Diffusion of the Results**: Including conferences, journal articles and other dissemination activities.
- **Future Directions**.

1. Main Contributions & General Conclusions

The main goal of this thesis has been to explore the usability of Machine-Learning algorithms in a DH context for different purposes in a building-scale and evaluate the efficiency of these black-box models for the energy management in a district-scale. It can be concluded that this PhD Thesis validates the possibility of using ML models in this context and that the performance metrics obtained by these models overperforms the metrics of current networks.

Each of the chapters have presented the partial conclusions related to the analysis carried out in each section of the Thesis. Therefore, the main contributions and conclusion of this Thesis can be summarized in the following bullet points:

- The Thesis gathers the most recent literature review on:
    - Unsupervised learning applied to energy demand in buildings.
    - Supervised learning applied to real energy demand data from buildings connected to DH networks.

Regarding the real case of the DH in Tartu:

- It presents a novel data-driven model, the so-called Q-T algorithm for energy predictions of buildings. This model is based on a multi-variable regression model divided by a previous Decision-Tree analysis. Similar studies have always divided the demand data by a specific outdoor temperature, but this way, there is part of the data that is never well characterized. As the heating demand is function of the external climatic variables and the calendar attributes, there are moments that the demand does not follow only climatic-dependence. It is demonstrated that dividing data by a specific demand, Q_{REF} in Q-T algorithm, works better and the model is more widely applicable to all type of buildings.
- The Thesis study a wide range of unsupervised algorithms for the identification of heating demand patterns in buildings. This study concluded that K-means algorithm is the one with highest performance metrics. For this purpose, several

CVIs have been used to analyze the "quality" of clusters or heating patterns in all the buildings of the case study.

- The Dissertation identifies several heat demand patterns for the different buildings using K-means algorithm. The final use of the buildings, e.g., residential buildings, commercial, etc. completely determines the heating patterns in the buildings. Synergies between buildings have been found.
- Other heating demand patterns have been identified by the supervised Decision-Trees based on calendar variables (these demand patterns can be found in some of the buildings only):
    - **Night Setback**. It can be used by the DH operator to reduce energy production in periods when a low heat load is expected, regardless of the climate conditions. A reduction of the heat load is identified between 3AM to 5AM in some of the buildings.
    - **Weekday-Weekend Patterns**. The lower or non-occupancy of the buildings in weekend days cause that the heat demand at the these days is lower than on weekdays in some of the buildings. This behavior is independent from the climatic conditions.
    - **Seasonal Patterns**. Despite being relatively low external temperatures at some moments of the summer, the monitored heat energy demand does not correspond to expectations for similar climatic conditions outside this season. This divergence could be motivated by a reduction of the heat load by the DH operator in this period. The methodology developed for the identification of this period is explained in Chapter XI.
- CART models enabled the qualitative characterization of the clusters/heat demand patterns. This study demonstrated that this algorithm is effective for cluster prediction. We can conclude that CART model without hourly temperatures used as predictors is the optimal CART for this purpose. Therefore, simple models are the most appropriate to avoid bias in testing. Besides, among the predictor variables, seasonality (summer and rest of the year), day of the

week and mean temperature are the variables that most affect the clustering process.

- As a further conclusion derived from the previous point, the classification model with highest accuracy results to be k-NN algorithm. Even though this model does not enable the authors to visualize the variables that are determining the clusters, better accuracy results than CARTs have been obtained. K-NN was used for developing the second step in the forecasting analysis.
- The prediction methodology that includes unsupervised cluster analysis improves the efficiency for prediction in building scale compared against Q-T algorithm or other regression models. In the four buildings where this study was focused, the R^2 value was improved, somehow, compared with the obtained with the Q-T algorithm. This efficiency gain, comparing use or not using clusters, ranges between 2% to around 50% (MVLR in Building 10949), always maintaining a reasonable computation time.
- Computation cost is a key variable for the operation of large DHN, where the hourly demand of several buildings is characterized. For high frequency predictions, such as hourly or sub-hourly forecasting, the response time of the model needs to be as fast as possible, so that the prediction of hundreds or thousands of buildings is feasible within the prediction interval. The Q-T algorithm enabled to discover the correlation between calendar variables and the instant demand, and it achieved remarkable prediction results. However, the large computation time of this model made necessary to analyze alternative ML models.
- Regarding other forecasting models developed along the Thesis, XGB-based method is the one with best prediction results, regardless the number of clusters and the classification model. Moreover, these accurate predictions are obtained with the lowest computation time among all the models simulated. XGB is followed, in terms of prediction accuracy, by SVR and RF, respectively. Therefore, the multistep method presented in this paper using extreme gradient

boosting as the predictive algorithm becomes a promising alternative to the most common operation algorithms used in the current DHN.

As for the case-study of the DH in Bilbao:

- When the models developed for the DH in Tartu are transferred to other networks in warmer climates (Bilbao, Spain), the forecasting results show lower accuracy than in the building located in Tartu (Estonia).
- Buildings modelled by DESING BUILDER software turn out to present lower demands than buildings modelled by other simplified methods. This demand reduction is caused because the software includes the solar gains and other thermal gains inside the buildings, while the simplified method only models the demand using climatic conditions (HDD). Contrastingly to the first point, the dispersion of the demand by DESING BUILDER are higher than other, even though including the PF factor for increasing the randomness of the demand.
- Predictions in individual buildings obtain better accuracy results when higher linearity of the demand is higher, and the dispersion of the demand is lower.
- Q-T algorithm overpredicts in every building and it is not capable to get accurate predictive results in warm climates. The milder climate with cooler temperatures impacts on the SH demand and reduced the correlation between the climatic conditions and the real demand.
- The model using a combination of unsupervised and supervised analysis is able to reduce the hours in which the model overpredicts the real demand compared with other models. In terms of energy, this model presents less energy over-produced and less energy underproduced in comparison with the temperature-based model. This means that the prediction accuracy of this models is better.
- A better prediction accuracy of the demand in a DH network results in economic savings in the energy production process. In the case analyzed in the last chapter of the Thesis, the economic savings reach the 10% in comparison with the baseline. The energy production mix and the type of energy produced in the

system will determine the exact savings in each district-heating, but a better predictive accuracy of the demand will always ensure better economic performance of the energy network. Some energy sources (such as solar thermal energy or other process-based heats streams) produce energy continuously regardless of the current demand in the district. In these cases, it is necessary to design a dynamic heat storage system that controls these energy flows from the production source to the final users. The better the demand is predicted; the better optimization of the heat storage size can be obtained. In this type of districts, it will be necessary to include an additional peak energy plant that fulfills the demand not covered by RES.

2. Future Directions

Although the PhD Thesis finishes here, the research work will still continue. Two main directions have been identified:

1. Application of Deep-Learning (Neural Networks) algorithms for modelling the heating demand of buildings.
2. Industrialization of the models developed in the dissertation.

Each of the work lines are described in the following paragraphs.

2.1. Application of Deep-Learning

This dissertation has been limited to the application of machine learning algorithm for various applications regarding heating energy demand in buildings connected to DH networks. The reason for the application of ML models (and not Deep-Learning) is the proximity of these models to physical models, such as traditional white-box models.

The results of this Thesis conclude that ML models could be successfully used for the management of real DH networks, obtaining in most cases, better results than the current models. However, there is still room for improvement in all directions: performance metrics, speed of training the models, etc. Thus, an option could be the artificial neural networks.

Deep-Learning algorithms, and in particular, neural networks are computing systems inspired by the biological neural networks that constitute animal brains. Artificial neural networks (ANNs) are comprised of a node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network.

Regarding the investigation lines that we are proposing as future directions of data-driven models are the following:

- Self-Organizing Maps (also known as Kohonen Maps) (SOM) for Anomaly Detection or Fault Identification [142]. Self-Organizing Maps or Kohonen's map is a type of artificial neural networks introduced by Teuvo Kohonen in the 1980s. The goal of the technique is to reduce dimensions and detect features. The maps help to visualize high-dimensional data. It represents the multidimensional data in a two- dimensional space using the self-organizing neural networks. The technique is used for data mining, face recognition, pattern recognition, speech analysis, industrial and medical diagnostics, anomalies detection.

  SOMs have been traditionally used in other fields such as bank faults detection or medical diagnostics. This model could be applied to energy demand anomaly identification, fault detection in the network or energy profile pattern recognition. An example of this type of network is shown in Fig. X-1.

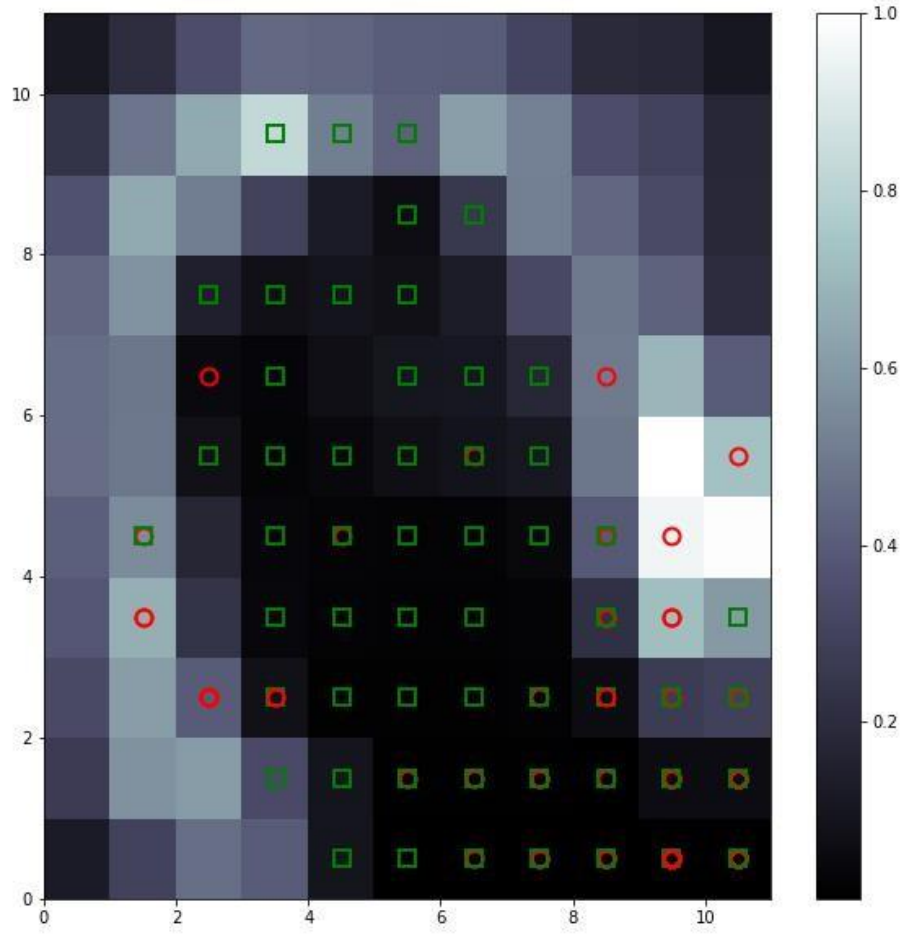


Fig. X-1. Example of SOM network. Developed by Mikel Lumbreras

- Long-Short-Term Memory (LSTM) for energy prediction [143]. In Chapter VIII we have studied the use of several ML models for energy prediction. Another option is the use of LSTM. The LSTM model is a special form of the recurrent neural network (RNN). This model conserves long-term memory by using memory units that can update the previous hidden state. It provides feedback at each neuron. We may compare the efficiency metrics and the time required for this optimized model and compare against the results in Chapter VIII.
- Use of other Neural Networks for energy prediction: Recurrent Neural Networks, Multilayer perceptron, Restricted Boltzmann machines, etc. The application method is the same than the analyzed in the previous point.

2.2. Industrialization of the Models

Finally, the other research line resulting from this Thesis is the industrialization or the production process of the models developed within this dissertation. Therefore, it is known that development time required in ML projects is small compared against the total time required for the real implementation of these algorithms. Chapter IX has studied the theoretical applicability of the models in different networks´ conditions. However, the research carried out in this Thesis only reaches the deployment in simulated cases. It would be interesting to advance to a real application demonstration of the models. For this purpose, it is necessary:

- Define and I&T structure for data storage and access.
- Define the Pre-processing activities for the data storage.
- Define the requirements for a Cloud or On-Premise infrastructure for the models.

3. Dissemination/Diffusion of the Results

Throughout the dissertation, we have already presented some of the diffusion ways in terms of articles published in international journals. In addition, at the time of writing these lines, another article is under the process of being published.

At the time of writing these lines, the following contributions to the dissemination of the results have been carried out, divided by international journals and conferences.

3.1. International Journals

The following articles are directly related to specific studied explained in the dissertation:

- M. Lumbreras *et al.*, “Data driven model for heat load prediction in buildings connected to District Heating by using smart heat meters,” *Energy*, vol. 239, p. 122318, Jan. 2022, doi: 10.1016/J.ENERGY.2021.122318.

- M. Lumbreras, G. Diarce, K. Martin, R. Garay-Martinez, and B. Arregi, "Unsupervised recognition and prediction of daily patterns in heating loads in buildings," *Journal of Building Engineering*, vol. 65, p. 105732, Apr. 2023, doi: 10.1016/J.JOBE.2022.105732.

The following article is under process of publication:

- Advanced Heat-Load Prediction Models in Buildings Combining Supervised & Unsupervised Learning.

Other articles published in the field of DH networks:

- M. Lumbreras and R. Garay, "Energy & economic assessment of façade-integrated solar thermal systems combined with ultra-low temperature district-heating," Renew Energy, vol. 159, pp. 1000–1014, Oct. 2020, doi: 10.1016/J.RENENE.2020.06.019.
- M. Lumbreras, G. Diarce, K. Martin-Escudero, A. Campos-Celador, and P. Larrinaga, "Design of district heating networks in built environments using GIS: A case study in Vitoria-Gasteiz, Spain," *J Clean Prod*, vol. 349, p. 131491, May 2022, doi: 10.1016/J.JCLEPRO.2022.131491.

The first page of all these publications is shown in Appendix.

3.2. International Conferences

SPLITECH, CISBAT, NSB, DECARBONIATION, RELATED (Roma)

- M. Lumbreras, R. Garay, and A. G. Marijuan, "Energy meters in District-Heating Substations for Heat Demand Characterization and Prediction Using Machine-Learning Techniques," *IOP Conf Ser Earth Environ Sci*, vol. 588, no. 3, p. 032007, Nov. 2020, doi: 10.1088/1755-1315/588/3/032007. [136]
- A. G. Marijuan, R. Garay, M. Lumbreras, L. Vladic, and R. Savić, "District Heating De-Carbonisation in Belgrade. Multi-Year transition plan," *IOP Conf Ser Earth Environ Sci*, vol. 588, no. 5, p. 052034, Nov. 2020, doi: 10.1088/1755-1315/588/5/052034. [137]

- M. Lumbreras, K. Martin-Escudero, G. Diarce, R. Garay-Martinez, and R. Mulero, “Unsupervised Clustering for Pattern Recognition of Heating Energy Demand in Buildings Connected to District-Heating Network,” *2021 6th International Conference on Smart and Sustainable Technologies (SpliTech)*, pp. 1–5, 2021, doi: 10.23919/SpliTech52315.2021.9566420. [138]
- R. Garay-Martinez, B. Arregi, M. Lumbreras, B. Zurro, J. M. Gonzalez, and J. L. Hernandez, “Data driven process for the energy assessment of building envelope retrofits,” *E3S Web of Conferences*, vol. 172, p. 25001, Jun. 2020, doi: 10.1051/e3sconf/202017225001. [139]
- A. G. Marijuan, R. Garay, M. Lumbreras, V. Sánchez, O. Macias, and J. P. S. de Rozas, “RELaTED Project: New Developments on Ultra-Low Temperature District Heating Networks,” in *The 8th Annual International Sustainable Places Conference (SP2020) Proceedings*, Dec. 2020, p. 8. doi: 10.3390/proceedings2020065008. [140]

3.3. National Conferences

- EESAP 2021 (Bilbao). Mikel Lumbreras, Koldobika Martin-Escudero, Gonzalo Diarce, Roberto Garay-Martinez, “Data-Driven Analysis of Heating Demand in Buildings Connected to District-Heating: Pattern Recognition and Demand Prediction”. [141]

Chapter XI

Appendix

Chapter XI Appendix

1. Publications` First Page

This section presents the different publications related with this dissertation and where the author is the corresponding author of the article:

Mikel Lumbreras, Roberto Garay, Energy & economic assessment of façade-integrated solar thermal systems combined with ultra-low temperature district-heating, Renewable Energy, Volume 159, 2020, Pages 1000-1014, ISSN 0960-1481

DOI: https://doi.org/10.1016/j.renene.2020.06.019.

An update to this article is included at the end

Renewable Energy 159 (2020) 1000–1014

Contents lists available at ScienceDirect

Renewable Energy

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

Energy & economic assessment of façade-integrated solar thermal systems combined with ultra-low temperature district-heating

Mikel Lumbreras *, Roberto Garay

TECNALIA, Basque Research and Technology Alliance (BRTA), Bizkaia Science and Technology Park, Astondo bidea 700, Derio, Spain

ARTICLE INFO

Article history:
Received 9 January 2020
Received in revised form
25 May 2020
Accepted 3 June 2020
Available online 8 June 2020

Keywords:
Solar district heating
Renewable energy sources
Building integrated solar thermal
4th generation district-heating
Smart energy systems

ABSTRACT

This paper conducts an energy and economic assessment of District Heating (DH) integrated Solar Thermal (ST) systems. An implementation with building-integrated ST collectors coupled to a Low Temperature District Heating (LTDH) system is studied, with special focus on unglazed collectors. ST heat is exploited in the building through direct use, while excess heat is delivered to the network. A novel control strategy for heat flows in the system is proposed.

A meta-analysis of several DH configurations, interconnection schemes and installed ST capacity is performed in three different climates: Sevilla (Spain), Bordeaux (France) & Copenhagen (Denmark). Heat loads corresponding to buildings with various insulation levels and domestic hot water loads are assessed in hourly simulations.

The proposed interconnection concept provides a variety of connection modes to the DH network, allowing up to a 50% increase in the provision of solar heat compared to an isolated ST system. Positive Return of Investment (ROI) for such a setup is achieved in 22% of the studied cases. The DH network is found to be a suitable heat sink in up to 25% of the buildings with ST systems installed.

1. Introduction

In the European Union (EU), 40% of the total energy consumption is performed in buildings [1]. European regulations ([2,3]) are promoting an overall energy performance improvement of 20% in buildings. For achieving this objective, the reduction in energy demand and the increase of renewable energy production in buildings ([4,5]) are the major routes.

In consolidated urban areas, most buildings are already constructed and the potential for reducing their heat loads and incorporating renewable energy sources (RES) is limited due to local constraints. District Heating (DH) networks allow the integration of buildings into greater energy systems, levelling various load profiles and optimizing use of renewable heat [6]. Existing buildings with higher heat loads can take advantage of excess heat productions in Solar Thermal fields integrated in neighbouring buildings.

DH is a highly efficient heat supply technology, identified as a key technology for the de-carbonisation of heat supply in Europe [7]. DH networks cover 13% of the heating energy to buildings in the EU [8], but are still highly dependent on fossil fuels [9], which are used for the production of 70% (Europe) to 90% (worldwide) of the heat.

There is a trend for the reduction of supply temperature levels in DHs towards the so-called 4th Generation of DH ([10–12]). Temperature reduction allows for a substantial increase in the use of RES. When considering the use of local RES in DH, the main sources are Solar Thermal (ST) systems and waste heat from industrial and commercial buildings [13]. In Ref. [14] the possibility of using industrial waste heat in LTDH in China was studied and it was concluded that LTDH increased both the waste heat production rate in industrial facilities and their thermal energy efficiency. In Ref. [15] the potential of Data Centres (DC) as waste heat streams for DH is studied, estimating operational cost savings in the range of 0.6–7.3% for a case study in Finland. In Sweden, so-called Open District Heating™ systems [16] have been introduced, where third parties are allowed to sell excess heat to the DH network. In Ref. [17] the feasibility of Large Solar Thermal (LST) combined with seasonal heat storage is validated for systems with collector surfaces in the range of 150 000 to 650 000 m^2 connected to the DH.

Solar Energy is the largest available RES in Earth [18]. Considering this, building energy codes in developed countries (such as

* Corresponding author.
E-mail addresses: mikel.lumbreras@tecnalia.com (M. Lumbreras), roberto.garay@tecnalia.com (R. Garay).

https://doi.org/10.1016/j.renene.2020.06.019

Mikel Lumbreras, Gonzalo Diarce, Koldobika Martin-Escudero, Alvaro Campos-Celador, Pello Larrinaga, Design of district heating networks in built environments using GIS: A case study in Vitoria-Gasteiz, Spain, Journal of Cleaner Production, Volume 349, 2022, 131491, ISSN 0959-6526,

DOI: https://doi.org/10.1016/j.jclepro.2022.131491.

Journal of Cleaner Production 349 (2022) 131491

Contents lists available at ScienceDirect

Journal of Cleaner Production

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

Design of district heating networks in built environments using GIS: A case study in Vitoria-Gasteiz, Spain

Mikel Lumbreras [a,*], Gonzalo Diarce [a], Koldobika Martin-Escudero [a], Alvaro Campos-Celador [b], Pello Larrinaga [a]

[a] ENEDI Research Group, Energy Engineering Department, Faculty of Engineering of Bilbao, University of the Basque Country (UPV/EHU), Pza. Ingeniero Torres Quevedo 1, Bilbao, 48013, Spain
[b] ENEDI Research Group, Energy Engineering Department, Faculty of Engineering of Bilbao, University of the Basque Country (UPV/EHU), Avda. Otaola 26, Eibar, 20600, Spain

ARTICLE INFO

Handling Editor: Mingzhou Jin

Keywords:
GIS
Industrial waste heat
Data-driven model
LiDAR
District heating

ABSTRACT

The efficient integration of high levels of industrial waste heat in low temperature district-heating networks is a promising technique that requires specific methodologies for its satisfactory implementation. This paper presents a novel methodology for assessing the energy and economic feasibility of new district-heating networks in existing urban areas for the integration of industrial waste heat sources. The methodology consists in an innovative multistep procedure using geographic information systems and data analysis tools, combining georeferenced data about buildings, industries and roads. The spatial distribution of the analysis area is divided into smaller buffers and grids, as a result, the routing design of the pipelines that makes up the district-heating topology is obtained under several assumptions. The methodology provides the most suitable area choice for the deployment of a district-heating, also implemented with a multi-step algorithm for routing the pipelines of the network. This methodology is applied to a particular case study located in Vitoria-Gasteiz (northern Spain). Different configurations for the district heating network are obtained with lengths of the network varying from 8 to 27 km. Payback values near to six years are achieved in most of the district-heating network configurations. The maximum payback period obtained within the configurations is 8.5 years. An economic sensitivity analysis is presented for the proposed optimal district-heating network configuration. The proposed methodology could be replicated for different case studies as long as the input data is available to the user.

1. Introduction

Energy consumption in buildings currently accounts for around 40% of the total energy consumption in the European Union (EU) (Pérez-Lombard et al., 2008). In the particular case of residential buildings, 57% of the total final energy consumption is used for space heating and 25% for domestic hot water (DHW) (Balaras et al., 2005). More than 50% of this energy consumption is nowadays fulfilled with natural gas and electricity (European Commission, 2019). Therefore, the implementation of alternative energy sources in buildings is vital to maintain a sustainable environment in cities and achieve the objectives of carbon neutral environment for 2050 (European Commission, 2017; Zhang et al., 2020) in the EU.

District Heating (DH) networks can provide an efficient alternative to individual installations in densely populated urban areas (Christian Holmstedt Hansen, 2018). Today, DH networks provide more than 13% of the heating energy to buildings in the EU (Lund, 2007). Besides, new DH networks enable the connection of low grade and decentralized renewable energy sources (Lund et al., 2018; Wen et al., 2021; Yılmaz Balaman and Selim, 2016), which offers the possibility to employ waste heat streams as a sustainable heat source.

There are different sources of waste heat available in cities or near them. Industrial residual thermal energy —commonly referred to as industrial waste heat (IWH) — is one of the most common. Recent studies show that the amount of IWH currently rejected to the environment accounts for 20–50% of the industrial energy consumption across EU (Brueckner et al., 2014). From that, 18–30% of it could be re-used in a technically feasible way (Brueckner et al., 2014; Vance et al., 2019). Since the emission temperature of around 65% of this IWH remains below 200 °C (Ankur Kapil, Igor Bulatov, Robin Smith, 2017), its reutilization for electricity production is hindered; however, these

* Corresponding author.
E-mail address: mikel.lumbreras@ehu.eus (M. Lumbreras).

https://doi.org/10.1016/j.jclepro.2022.131491
Received 18 August 2021; Received in revised form 8 February 2022; Accepted 20 March 2022
Available online 22 March 2022

Mikel Lumbreras, Roberto Garay-Martinez, Beñat Arregi, Koldobika Martin-Escudero, Gonzalo Diarce, Margus Raud, Indrek Hagu, Data driven model for heat load prediction in buildings connected to District Heating by using smart heat meters, Energy, Volume 239, Part D, 2022, 122318, ISSN 0360-5442,

DOI: https://doi.org/10.1016/j.energy.2021.122318.

Energy 239 (2022) 122318

Contents lists available at ScienceDirect

Energy

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

Data driven model for heat load prediction in buildings connected to District Heating by using smart heat meters

Mikel Lumbreras [a, *], Roberto Garay-Martinez [b], Beñat Arregi [b], Koldobika Martin-Escudero [a], Gonzalo Diarce [a], Margus Raud [c], Indrek Hagu [c]

[a] ENEDI Research Group, Department of Energy Engineering, Faculty of Engineering of Bilbao, University of the Basque Country UPV/EHU, Pza. Ingeniero Torres Quevedo 1, Bilbao, 48013, Spain
[b] TECNALIA, Basque Research and Technology Alliance (BRTA), Bizkaia Science and Technology Park, Astondo Bidea 700, Derio, Spain
[c] GREN Eesti, Turu 18, Tartu, Estonia

ARTICLE INFO

Article history:
Received 14 May 2021
Received in revised form
6 October 2021
Accepted 9 October 2021
Available online 12 October 2021

Keywords:
Load forecasting
Heat meters
Data-driven model
Building
District Heating

ABSTRACT

An accurate characterization and prediction of heat loads in buildings connected to a District Heating (DH) network is crucial for the effective operation of these systems. The high variability of the heat production process of DH networks with low supply temperatures and derived from the incorporation of different heat sources increases the need for heat demand prediction models. This paper presents a novel data-driven model for the characterization and prediction of heating demand in buildings connected to a DH network.

This model is built on the so-called Q-algorithm and fed with real data from 42 smart energy meters located in 42 buildings connected to the DH in Tartu (Estonia). These meters deliver heat consumption data with a 1-h frequency. Heat load profiles are analysed, and a model based on supervised clustering methods in combination with multiple variable regression is proposed. The model makes use of four climatic variables, including outdoor ambient temperature, global solar radiation and wind speed and direction, combined with time factors and data from smart meters. The model is designed for deployment over large sets of the building stock, and thus aims to forecast heat load regardless of the construction characteristics or final use of the building. The low computational cost required by this algorithm enables its integration into machines with no special requirements due to the equations governing the model.

The data-driven model is evaluated both statistically and from an engineering or energetic point of view. R^2 values from 0.70 to 0.99 are obtained for daily data resolution and R^2 values up to 0.95 for hourly data resolution. Hourly results are very promising for more than 90% of the buildings under study.

1. Introduction

Energy consumption in buildings accounts for up to 40% of the total energy consumption in the European Union (EU) [1]. Considering this, increasing energy efficiency in buildings is one of the key targets of the EU strategy for the de-carbonisation of the economy ([2,3]).

Current District Heating (DH) networks are responsible for covering around 13% of the total thermal energy demand in the EU [4]. The evolution of DH networks over the years has been reducing supply temperatures, originally in the range of 80 °C and over, with the progressive implementation of the so-called 4th Generation District Heating (4GDH) ([5,6]) or Ultra Low Temperature (ULT) DH networks, which supply heat at temperatures around 45 °C. This has enabled an increased integration of low grade energy sources such as solar thermal (ST) systems [7] or waste heat (WH) streams ([8–10]) in the heat network.

The increasingly important role of renewable energy sources in 4GDH increases the variability of the heat generation profile in the heat production facilities. This requires the introduction of energy generation flexibility techniques to adapt heat production and demand in the network. To do so, accurate characterization methods for heat loads are required, so the available energy sources can be correctly managed with respect to such external variables as

* Corresponding author.
E-mail address: mikel.lumbreras@ehu.eus (M. Lumbreras).

https://doi.org/10.1016/j.energy.2021.122318

Mikel Lumbreras, Gonzalo Diarce, Koldobika Martin, Roberto Garay-Martinez, Beñat Arregi, Unsupervised recognition and prediction of daily patterns in heating loads in buildings, Journal of Building Engineering, Volume 65, 2023, 105732, ISSN 2352-7102

DOI: https://doi.org/10.1016/j.jobe.2022.105732.

Journal of Building Engineering 65 (2023) 105732

Contents lists available at ScienceDirect

Journal of Building Engineering

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

Unsupervised recognition and prediction of daily patterns in heating loads in buildings

Mikel Lumbreras [a,*], Gonzalo Diarce [a], Koldobika Martin [a], Roberto Garay-Martinez [b], Beñat Arregi [c]

[a] ENEDI Research Group, Energy Engineering Department, Faculty of Engineering of Bilbao, University of the Basque Country (UPV/EHU), Pza. Ingeniero Torres Quevedo 1, Bilbao, 48013, Spain
[b] Institute of Technology, Faculty of Engineering, University of Deusto, Av. Universidades, 24, 48007, Bilbao, Spain
[c] TECNALIA, Basque Research and Technology Alliance (BRTA), Bizkaia Science and Technology Park, Astondo Bidea 700, Derio, Spain

ARTICLE INFO

Keywords:
Pattern recognition
Unsupervised clustering
Heating loads
Daily profiles

ABSTRACT

This paper presents a multistep methodology combining unsupervised and supervised learning techniques for the identification of the daily heating energy consumption patterns in buildings. The relevant number of typical profiles is obtained through unsupervised clustering processes. Then Classification and Regression Trees are used to predict the profile type corresponding to external variables, including calendar and climatic variables, from any given day. The methodology is tested with a variety of datasets for three different buildings with different uses connected to the district heating network in Tartu (Estonia). The three buildings under analysis present different energy behaviors (residential, kindergarten and commercial buildings). The paper shows that unsupervised clustering is effective for pattern recognition since the results from the classification and regression trees match the results from the unsupervised clustering. Three main patterns have been identified in each building, seasonality and daily mean temperature being the variables that have the greatest effect. The results concluded that the best classification accuracy is obtained with a small number of clusters with a classification accuracy from 0.7 to 0.85, approximately.

Nomenclature

Acronyms

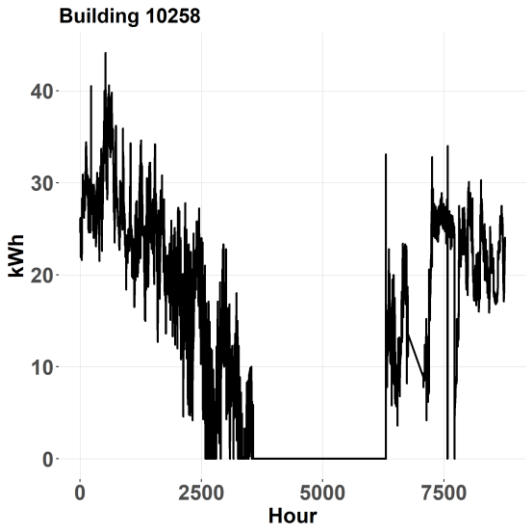
CART	Classification & Regression trees
CVI	Cluster Validation Index
DBSCAN	Density Based Clustering
DH	District-Heating
DHW	Domestic Hot Water
DS	Dataset
EU	European Union
RES	Renewable Energy Source

* Corresponding author.
E-mail address: mikel.lumbreras@ehu.eus (M. Lumbreras).

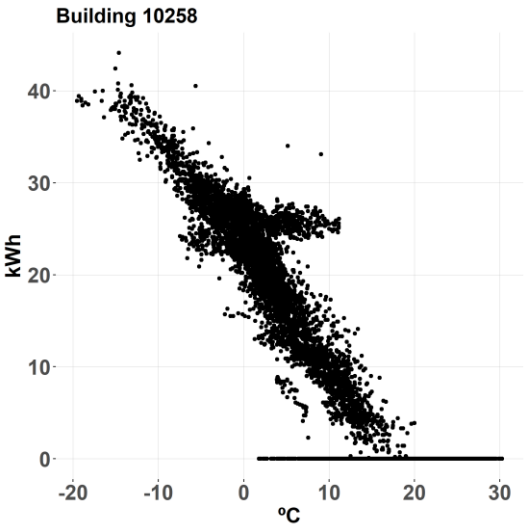
https://doi.org/10.1016/j.jobe.2022.105732
Received 11 August 2022; Received in revised form 5 December 2022; Accepted 10 December 2022
Available online 16 December 2022

2. Buildings´ Demand Profiles

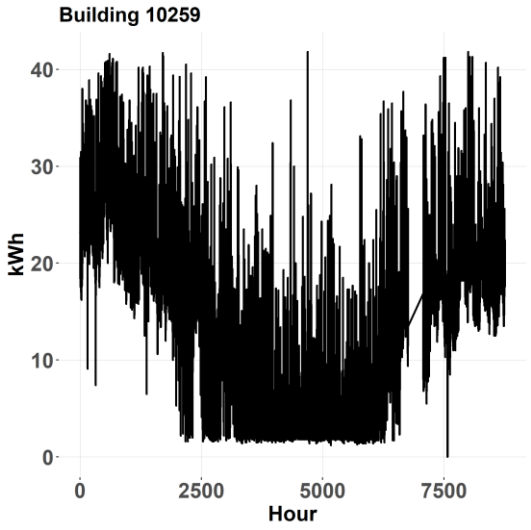
The following figures´ group present the total demand of all the buildings under study.



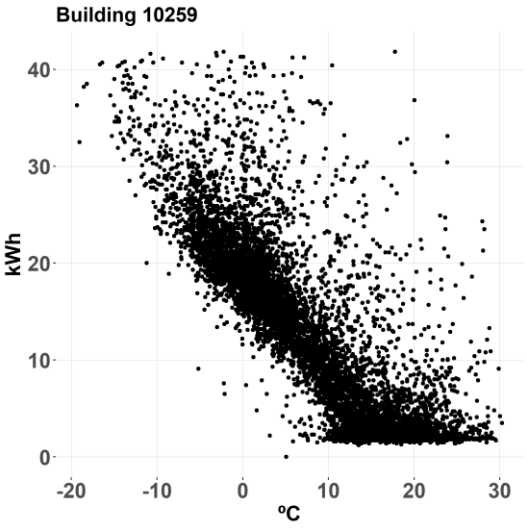
(a)



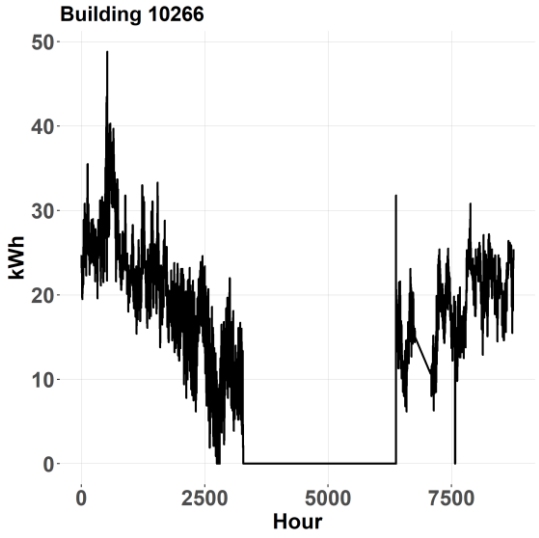
(b)



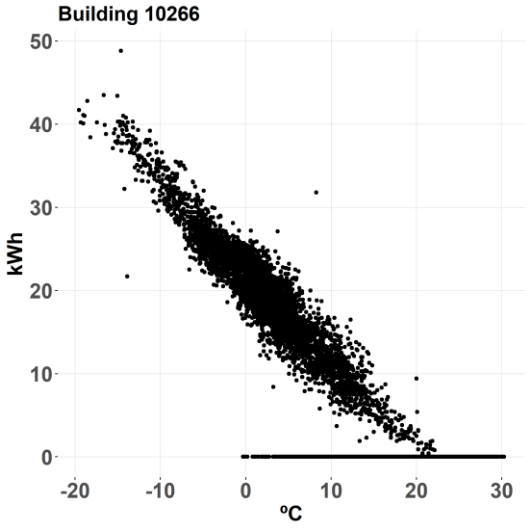
(a)



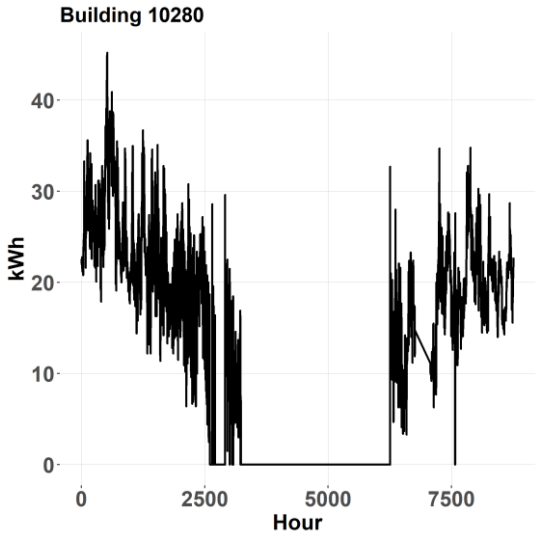
(b)


Building 10266
kWh
Hour
0
2500
5000
7500
0
10
20
30
40
50

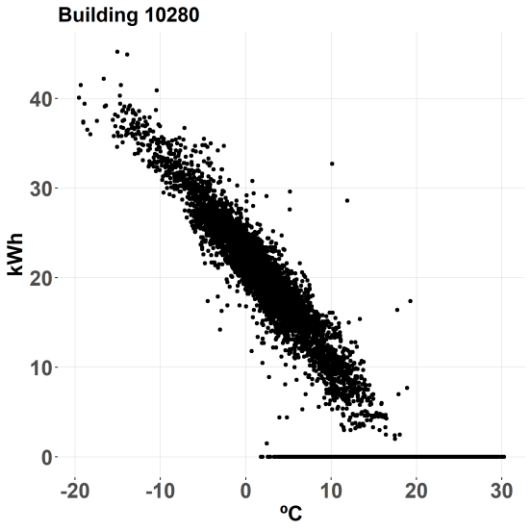
(a)


Building 10266
kWh
°C
-20
-10
0
10
20
30
0
10
20
30
40
50

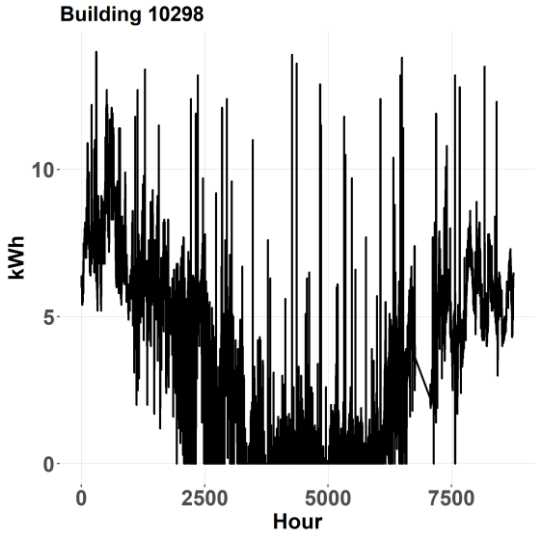
(b)


Building 10280
kWh
Hour
0
2500
5000
7500
0
10
20
30
40

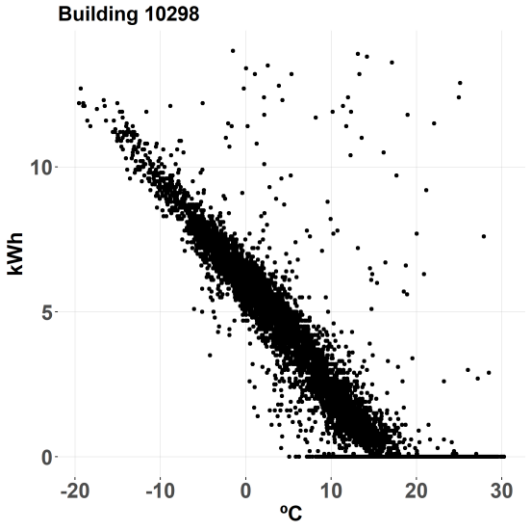
(a)


Building 10280
kWh
°C
-20
-10
0
10
20
30
0
10
20
30
40

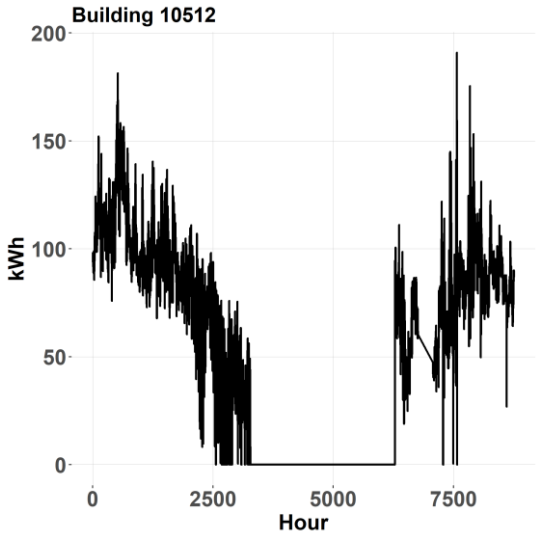
(b)


Building 10298
kWh
10
5
0
0
2500
5000
7500
Hour

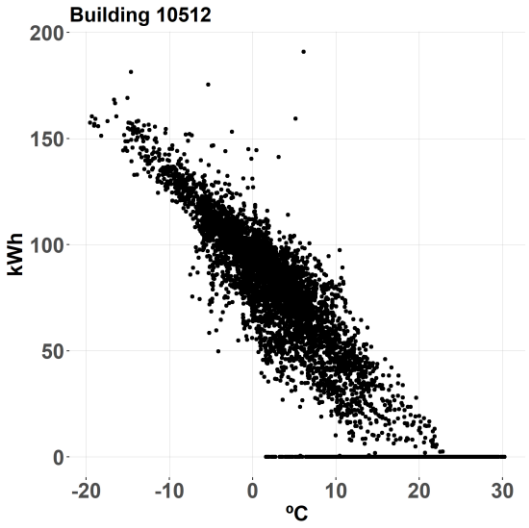
(a)


Building 10298
kWh
10
5
0
-20
-10
0
10
20
30
ºC

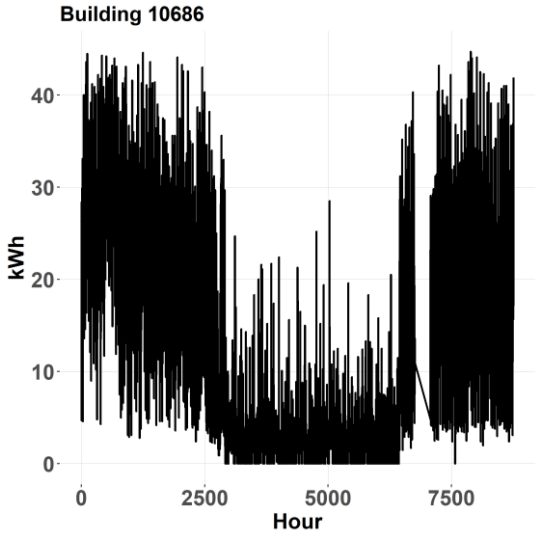
(b)


Building 10512
kWh
200
150
100
50
0
0
2500
5000
7500
Hour

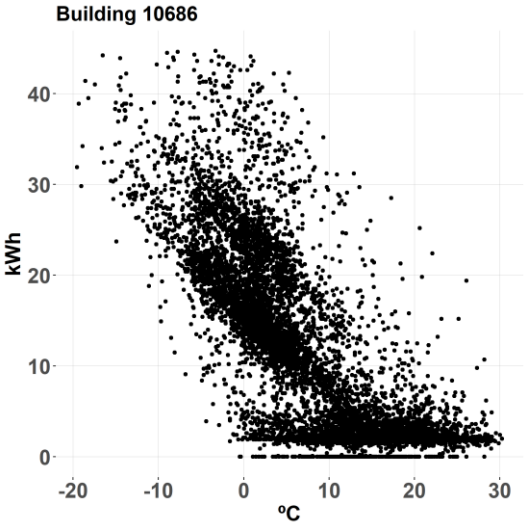
(a)


Building 10512
kWh
200
150
100
50
0
-20
-10
0
10
20
30
ºC

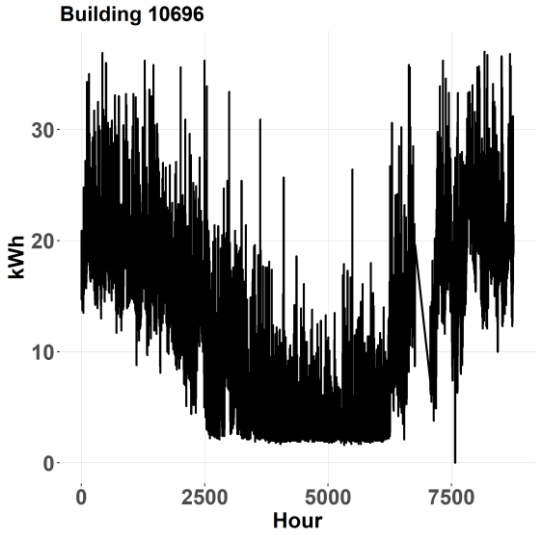
(b)


Building 10686
kWh
Hour
0
10
20
30
40
0
2500
5000
7500

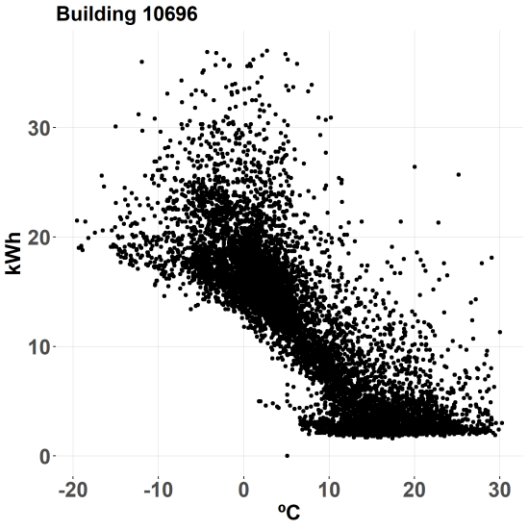
(a)


Building 10686
kWh
°C
0
10
20
30
40
-20
-10
0
10
20
30

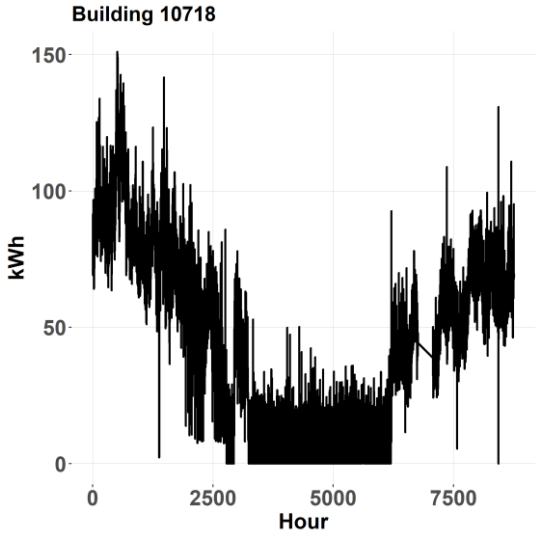
(b)


Building 10696
kWh
Hour
0
10
20
30
0
2500
5000
7500

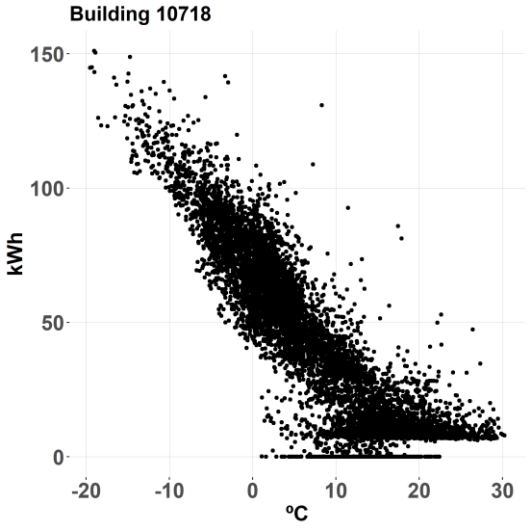
(a)


Building 10696
kWh
°C
0
10
20
30
-20
-10
0
10
20
30

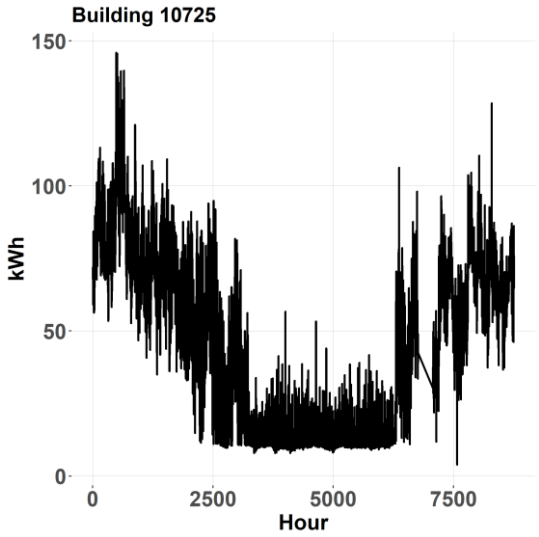
(b)


Building 10718
150
100
50
0
kWh
0
2500
5000
7500
Hour

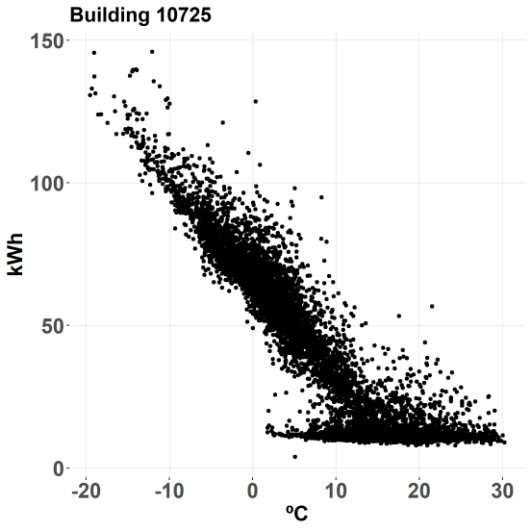
(a)


Building 10718
150
100
50
0
kWh
-20
-10
0
10
20
30
ºC

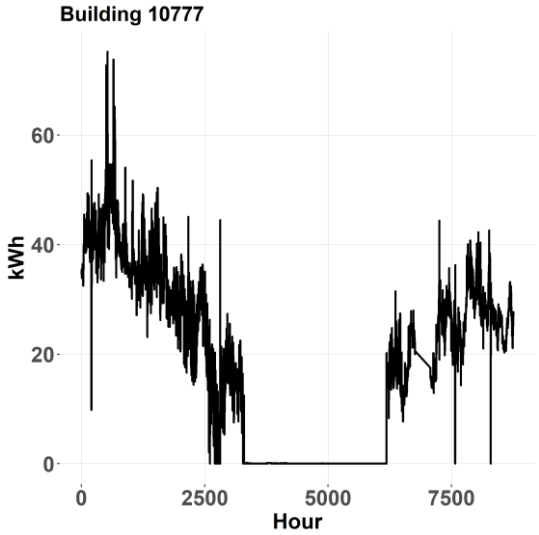
(b)


Building 10725
150
100
50
0
kWh
0
2500
5000
7500
Hour

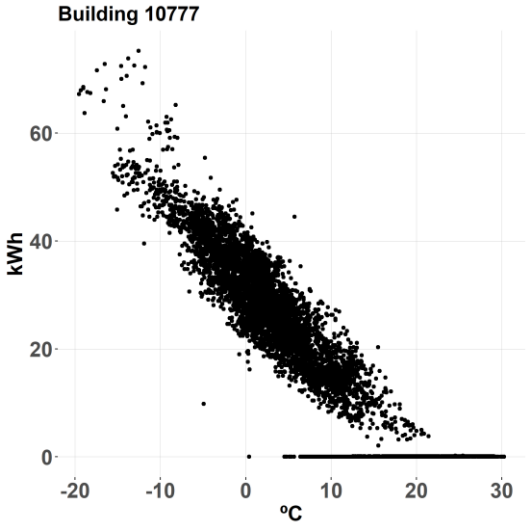
(a)


Building 10725
150
100
50
0
kWh
-20
-10
0
10
20
30
ºC

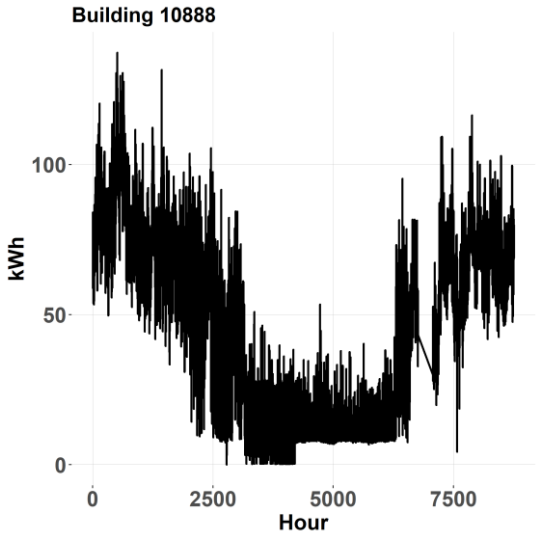
(b)


Building 10777
kWh
0
20
40
60
0
2500
5000
7500
Hour

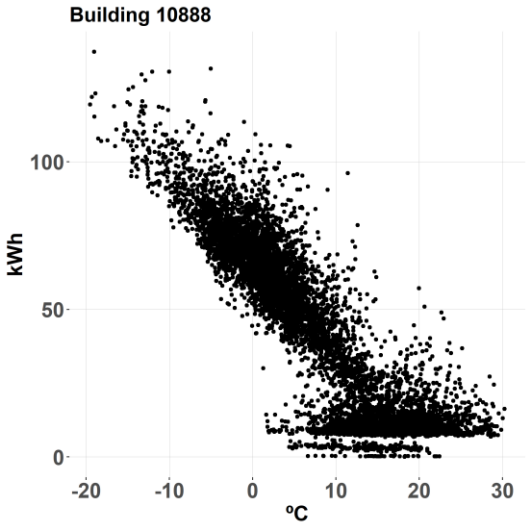
(a)


Building 10777
kWh
0
20
40
60
-20
-10
0
10
20
30
ºC

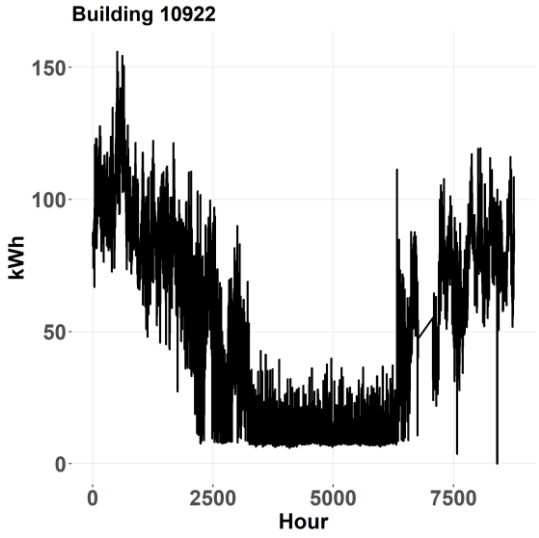
(b)


Building 10888
kWh
0
50
100
0
2500
5000
7500
Hour

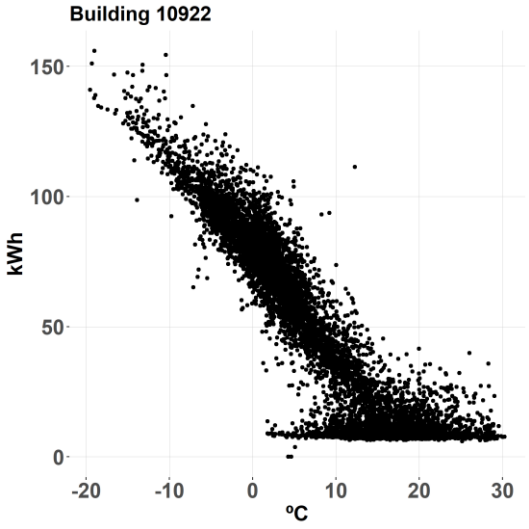
(a)


Building 10888
kWh
0
50
100
-20
-10
0
10
20
30
ºC

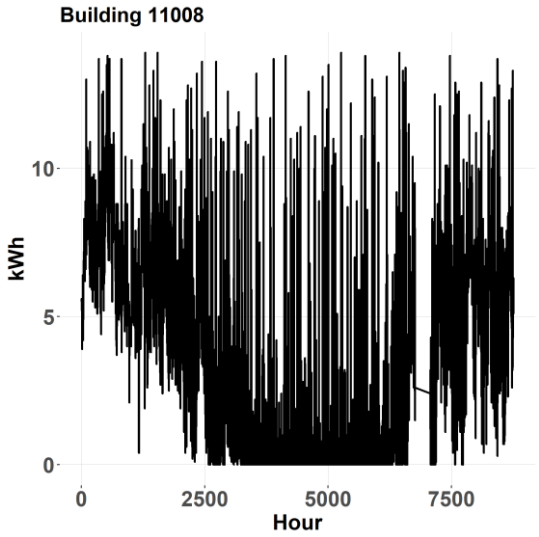
(b)


Building 10922
150
100
50
0
kWh
0
2500
5000
7500
Hour

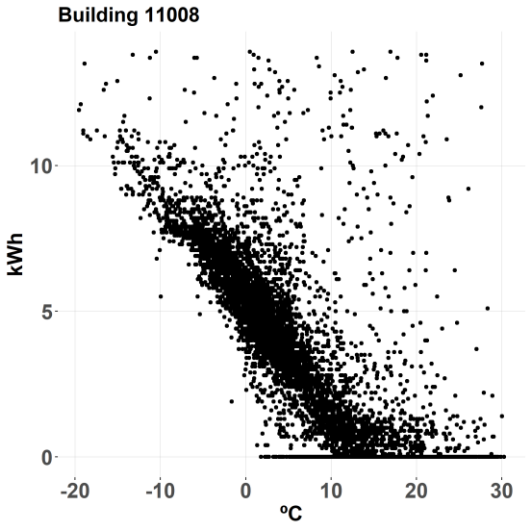
(a)


Building 10922
150
100
50
0
kWh
-20
-10
0
10
20
30
ºC

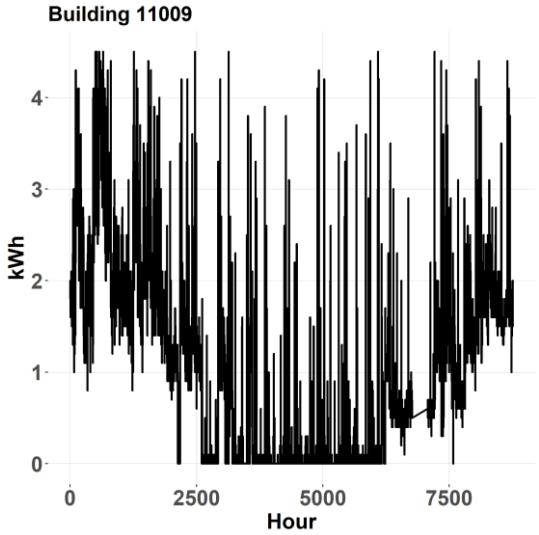
(b)


Building 11008
10
5
0
kWh
0
2500
5000
7500
Hour

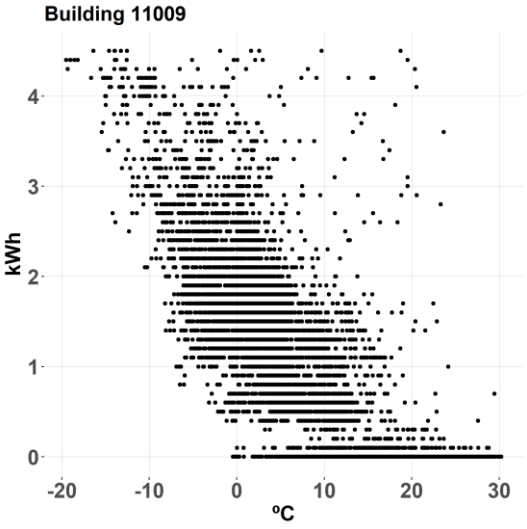
(a)


Building 11008
10
5
0
kWh
-20
-10
0
10
20
30
ºC

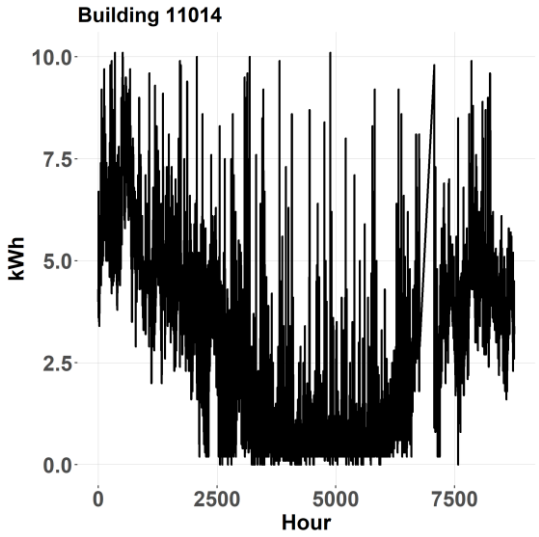
(b)


Building 11009
kWh
Hour
0
2500
5000
7500
0
1
2
3
4

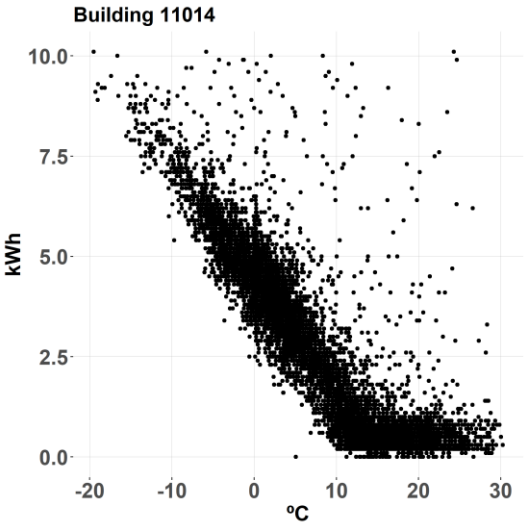
(a)


Building 11009
kWh
ºC
-20
-10
0
10
20
30
0
1
2
3
4

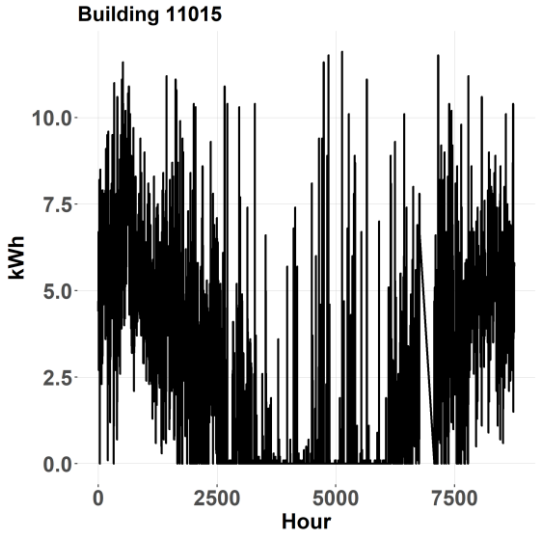
(b)


Building 11014
kWh
Hour
0
2500
5000
7500
0.0
2.5
5.0
7.5
10.0

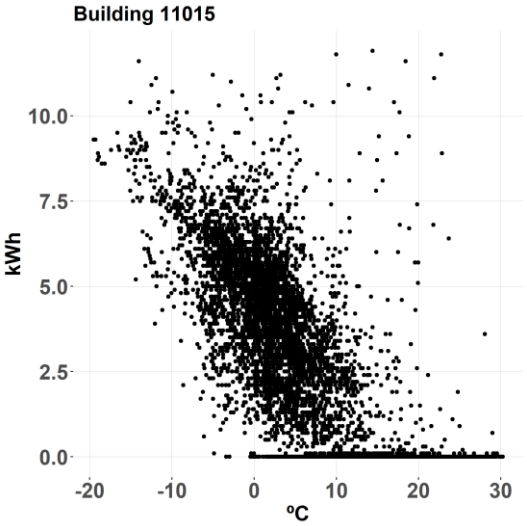
(a)


Building 11014
kWh
ºC
-20
-10
0
10
20
30
0.0
2.5
5.0
7.5
10.0

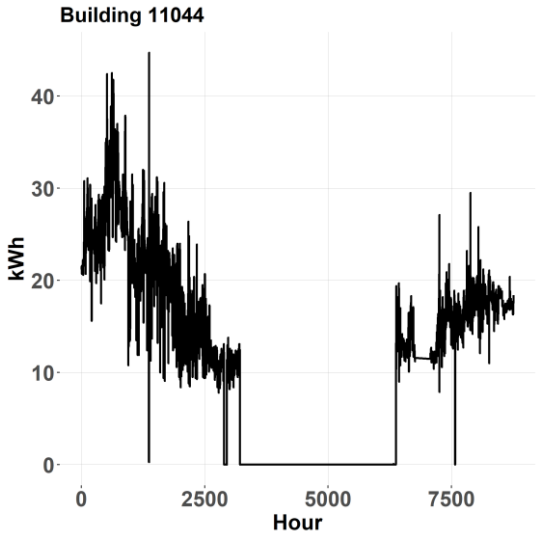
(b)


Building 11015
kWh
10.0
7.5
5.0
2.5
0.0
0
2500
5000
7500
Hour

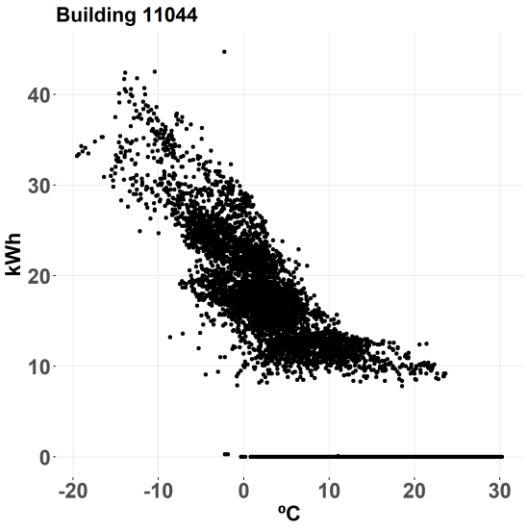
(a)


Building 11015
kWh
10.0
7.5
5.0
2.5
0.0
-20
-10
0
10
20
30
ºC

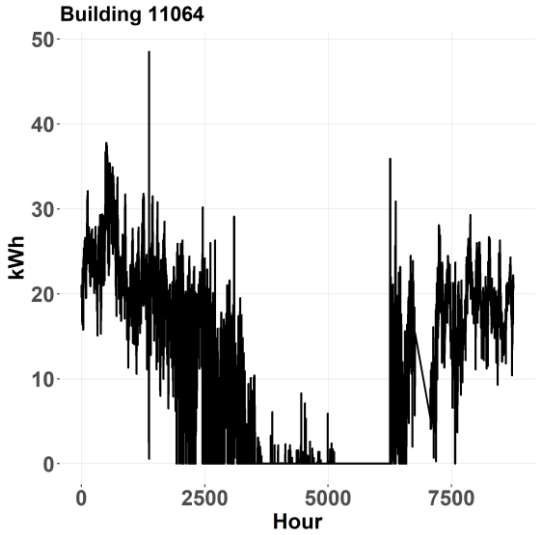
(b)


Building 11044
kWh
40
30
20
10
0
0
2500
5000
7500
Hour

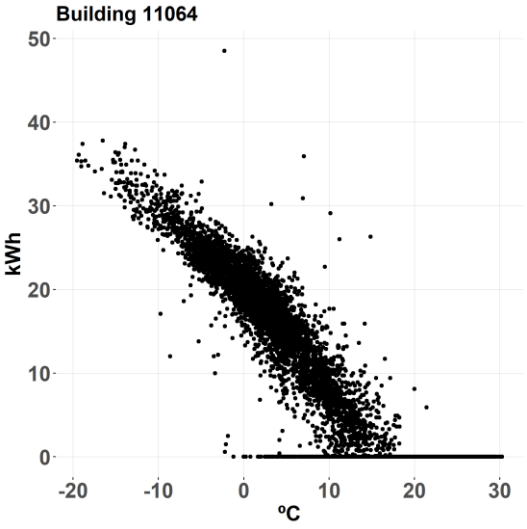
(a)


Building 11044
kWh
40
30
20
10
0
-20
-10
0
10
20
30
ºC

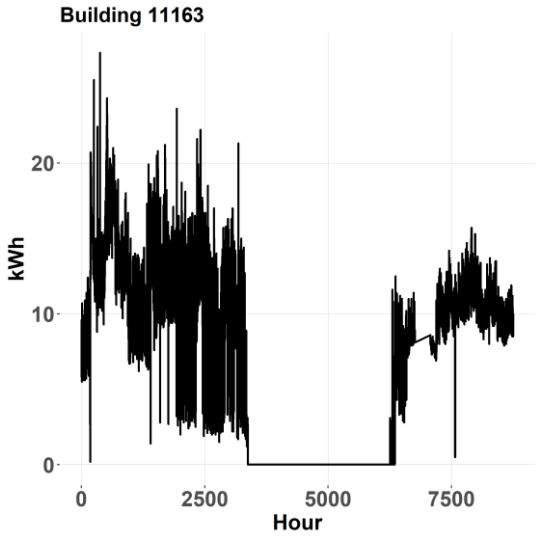
(b)


Building 11064
kWh
Hour
0
2500
5000
7500
0
10
20
30
40
50

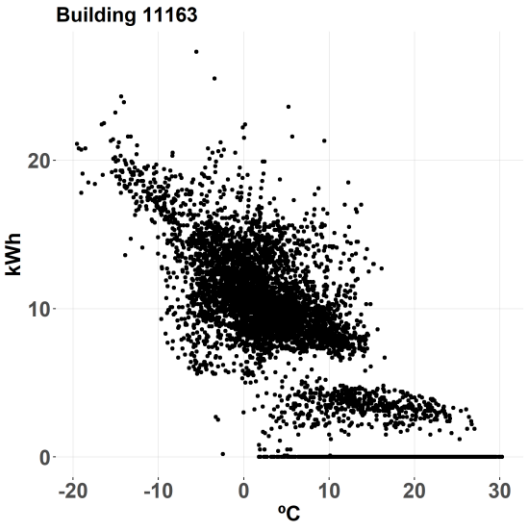
(a)


Building 11064
kWh
°C
-20
-10
0
10
20
30
0
10
20
30
40
50

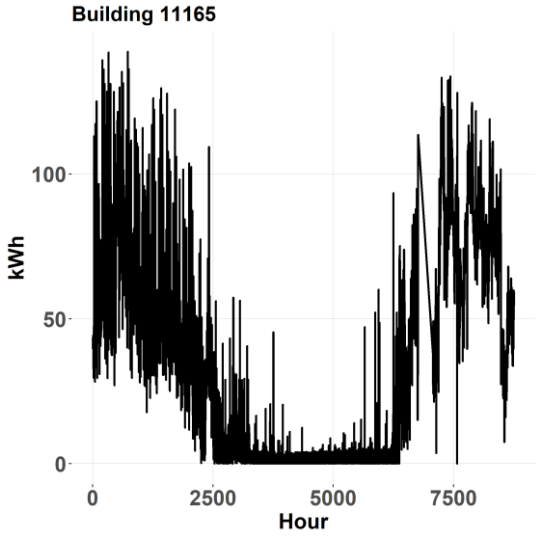
(b)


Building 11163
kWh
Hour
0
2500
5000
7500
0
10
20

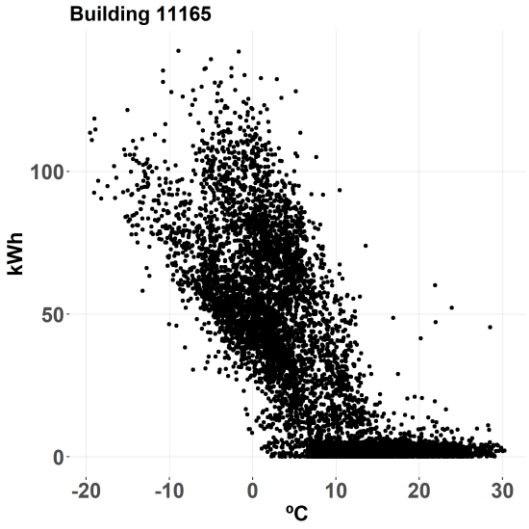
(a)


Building 11163
kWh
°C
-20
-10
0
10
20
30
0
10
20

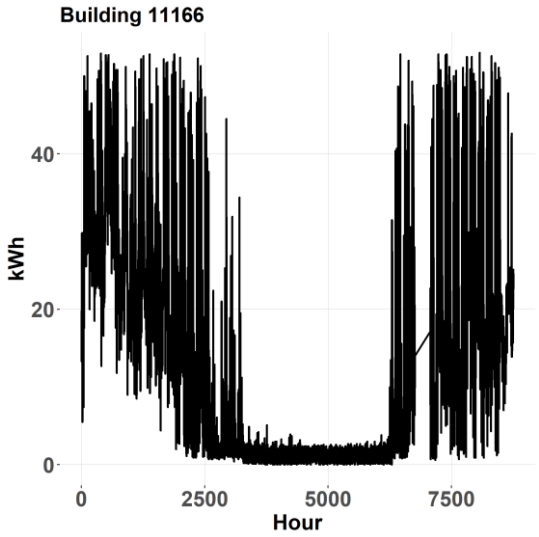
(b)


Building 11165
kWh
100
50
0
0
2500
5000
7500
Hour

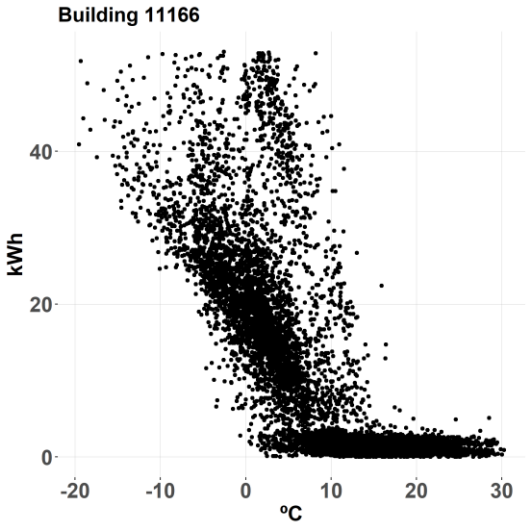
(a)


Building 11165
kWh
100
50
0
-20
-10
0
10
20
30
ºC

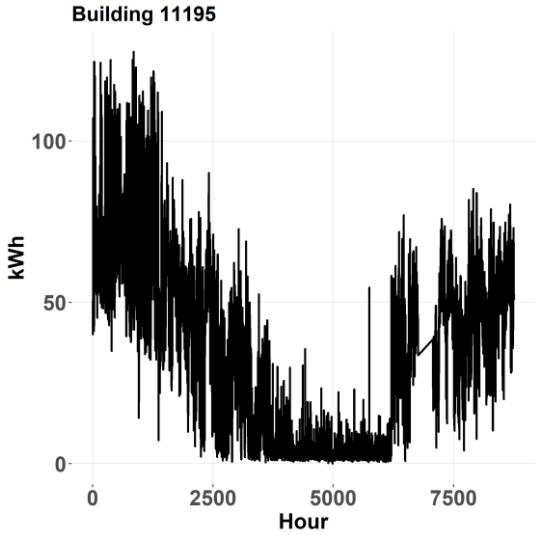
(b)


Building 11166
kWh
40
20
0
0
2500
5000
7500
Hour

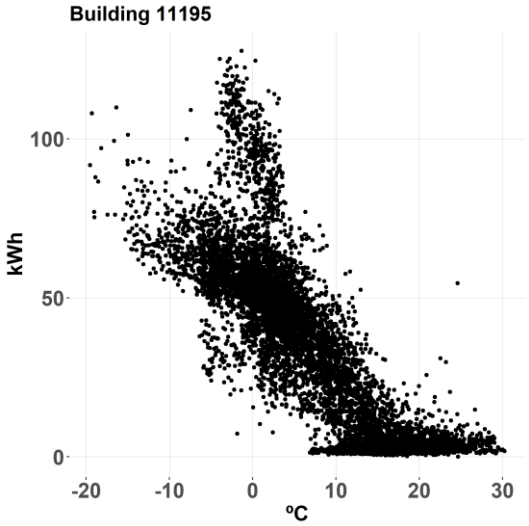
(a)


Building 11166
kWh
40
20
0
-20
-10
0
10
20
30
ºC

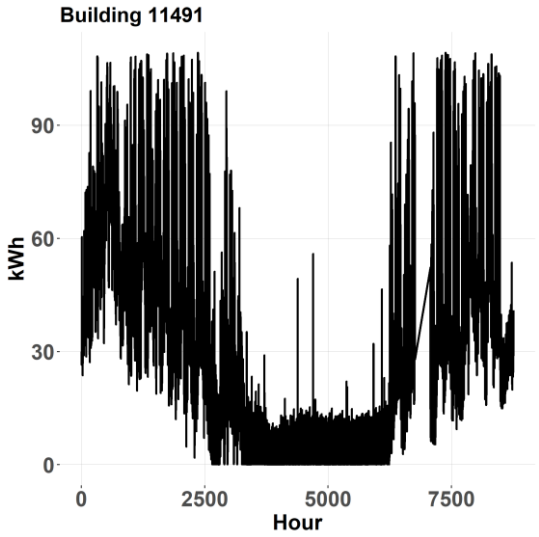
(b)


Building 11195
kWh
100
50
0
0
2500
5000
7500
Hour

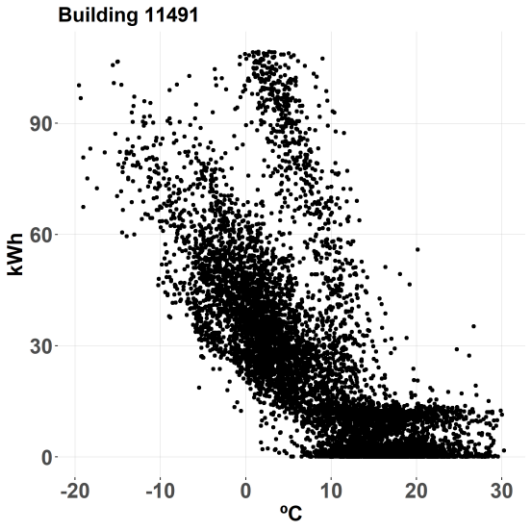
(a)


Building 11195
kWh
100
50
0
-20
-10
0
10
20
30
ºC

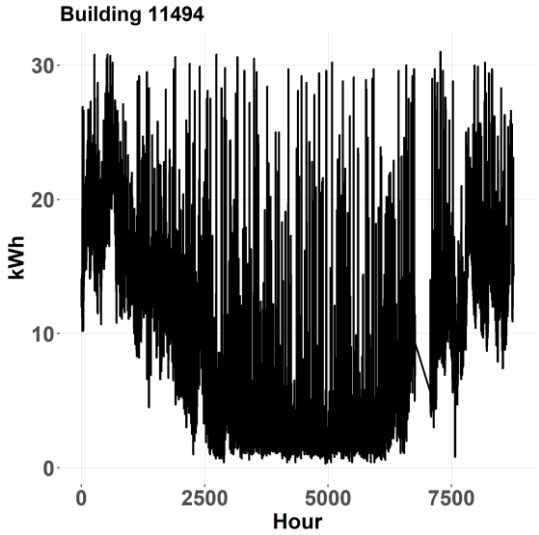
(b)


Building 11491
kWh
90
60
30
0
0
2500
5000
7500
Hour

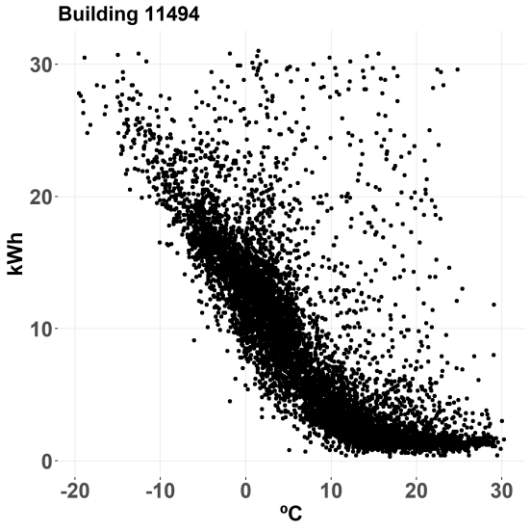
(a)


Building 11491
kWh
90
60
30
0
-20
-10
0
10
20
30
ºC

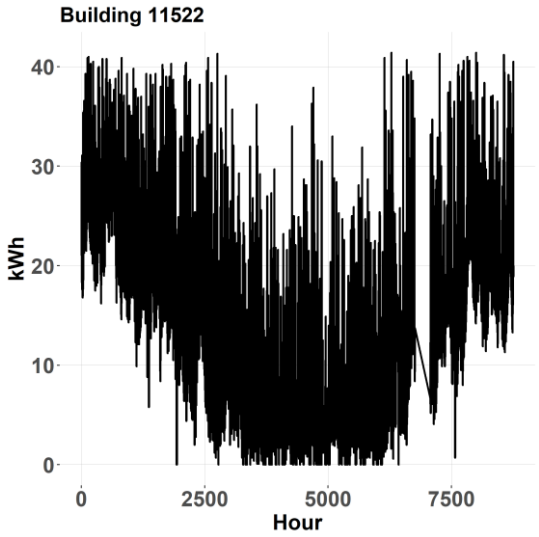
(b)


Building 11494
kWh
Hour
0
2500
5000
7500
0
10
20
30

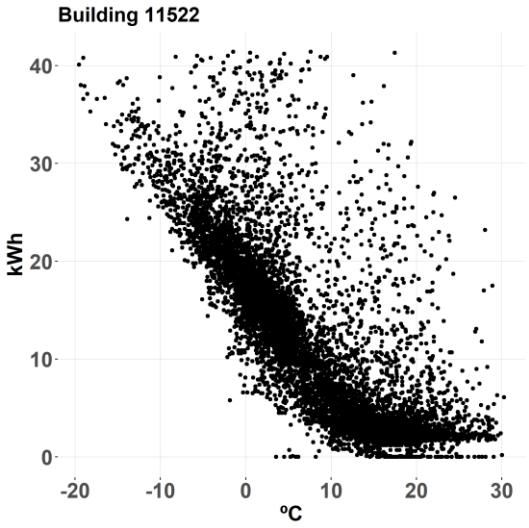
(a)


Building 11494
kWh
°C
-20
-10
0
10
20
30
0
10
20
30

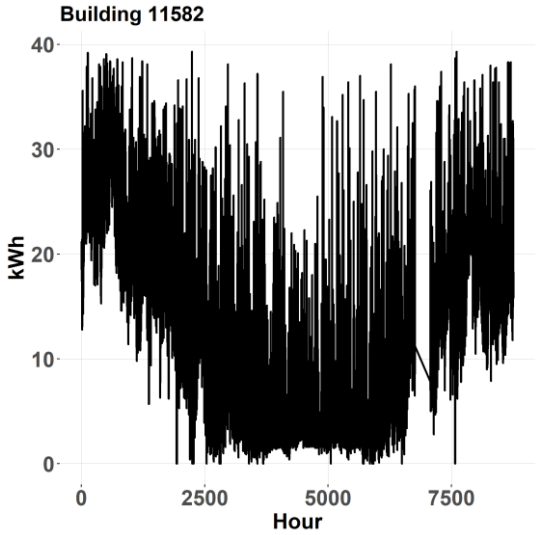
(b)


Building 11522
kWh
Hour
0
2500
5000
7500
0
10
20
30
40

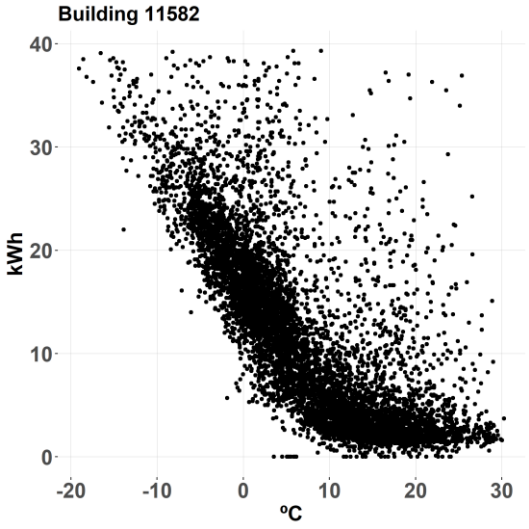
(a)


Building 11522
kWh
°C
-20
-10
0
10
20
30
0
10
20
30
40

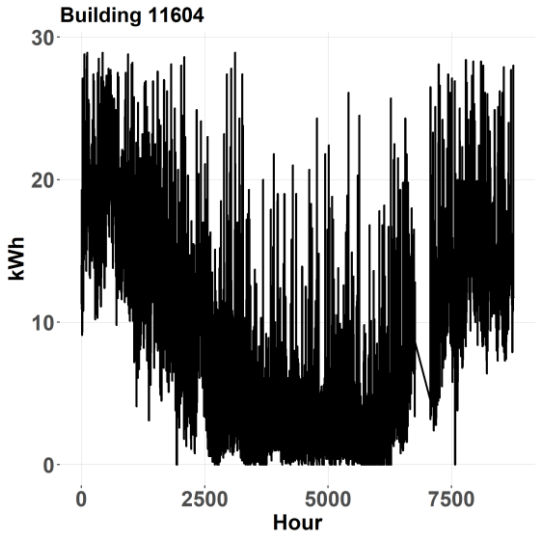
(b)


Building 11582
kWh
Hour
0
2500
5000
7500
0
10
20
30
40

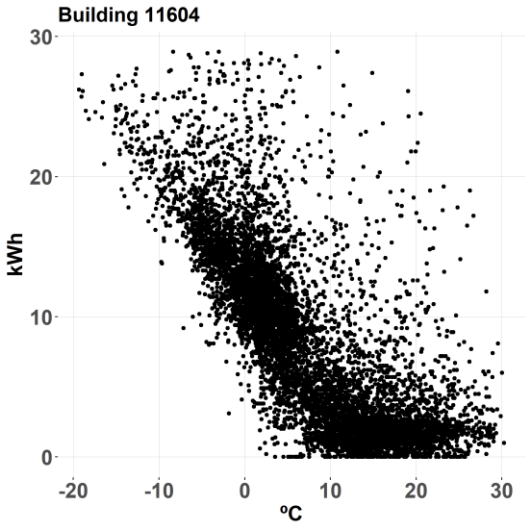
(a)


Building 11582
kWh
°C
-20
-10
0
10
20
30
0
10
20
30
40

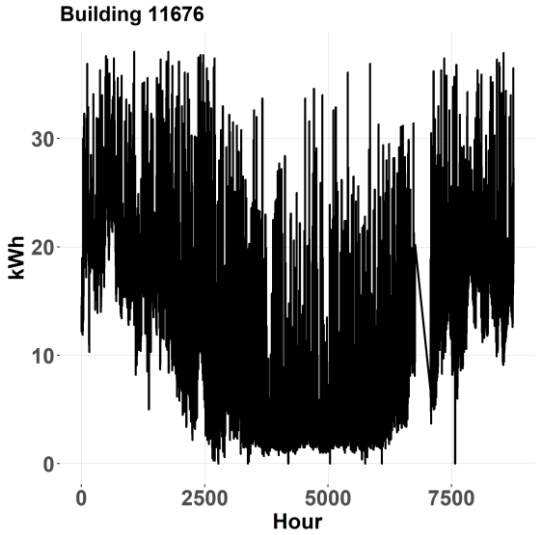
(b)


Building 11604
kWh
Hour
0
2500
5000
7500
0
10
20
30

(a)


Building 11604
kWh
°C
-20
-10
0
10
20
30
0
10
20
30

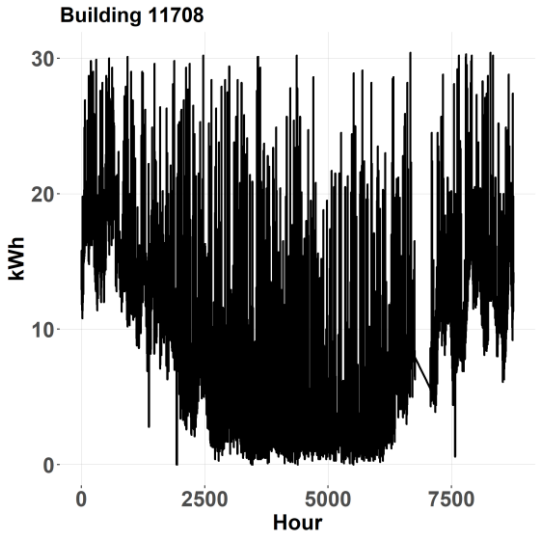
(b)


Building 11676
kWh
Hour
0
10
20
30
0
2500
5000
7500

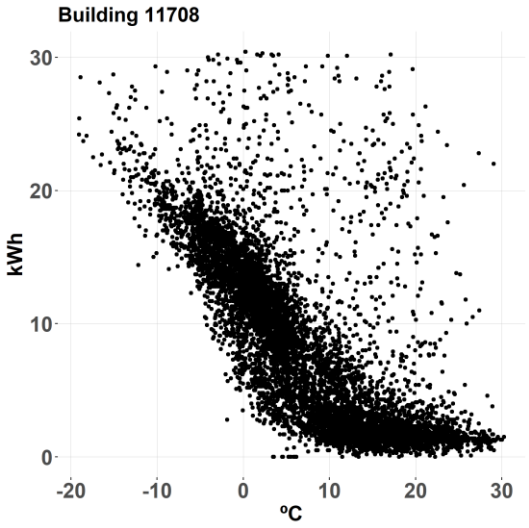
(a)


Building 11676
kWh
°C
0
10
20
30
-20
-10
0
10
20
30

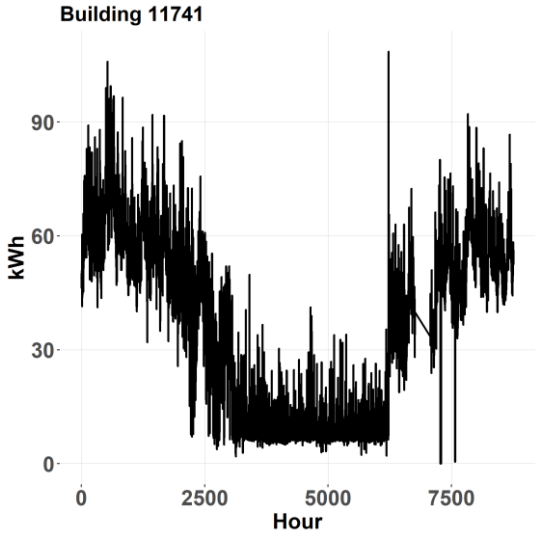
(b)


Building 11708
kWh
Hour
0
10
20
30
0
2500
5000
7500

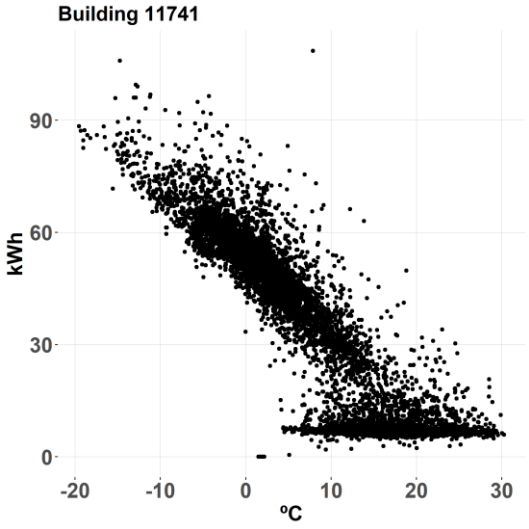
(a)


Building 11708
kWh
°C
0
10
20
30
-20
-10
0
10
20
30

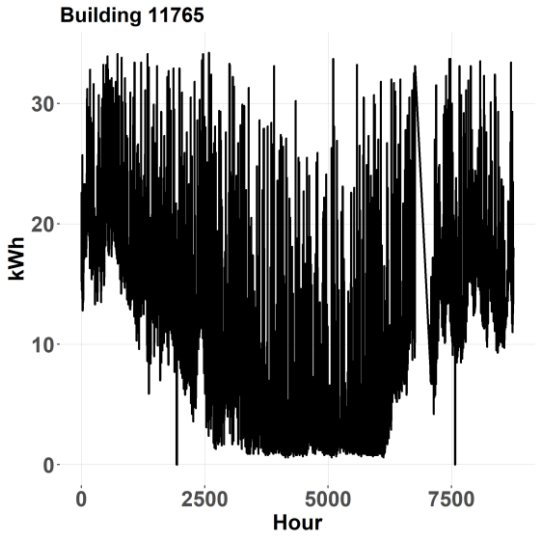
(b)


Building 11741
kWh
90
60
30
0
0
2500
5000
7500
Hour

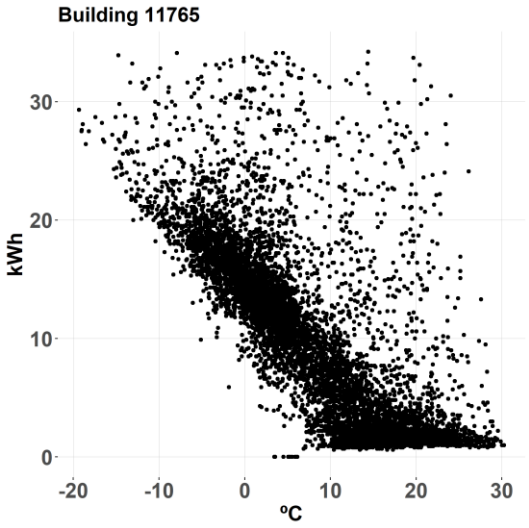
(a)


Building 11741
kWh
90
60
30
0
-20
-10
0
10
20
30
ºC

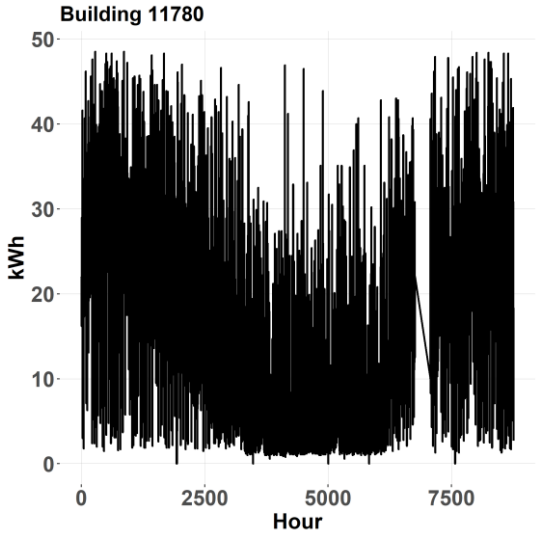
(b)


Building 11765
kWh
30
20
10
0
0
2500
5000
7500
Hour

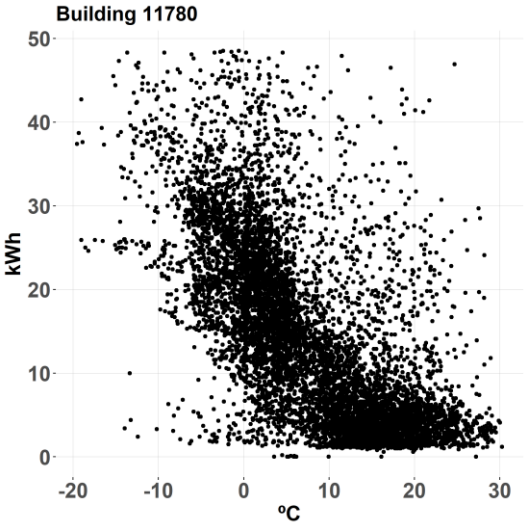
(a)


Building 11765
kWh
30
20
10
0
-20
-10
0
10
20
30
ºC

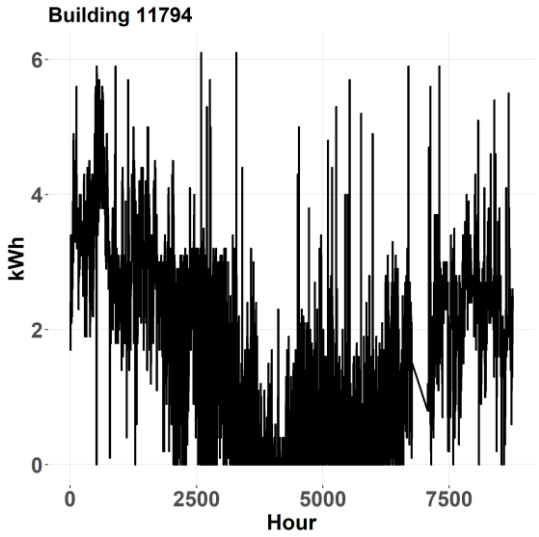
(b)


Building 11780
kWh
Hour
0
2500
5000
7500
0
10
20
30
40
50

(a)


Building 11780
kWh
ºC
-20
-10
0
10
20
30
0
10
20
30
40
50

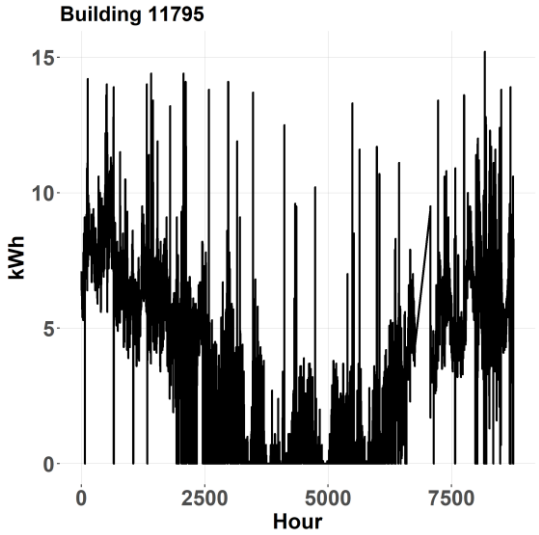
(b)


Building 11794
kWh
Hour
0
2500
5000
7500
0
2
4
6

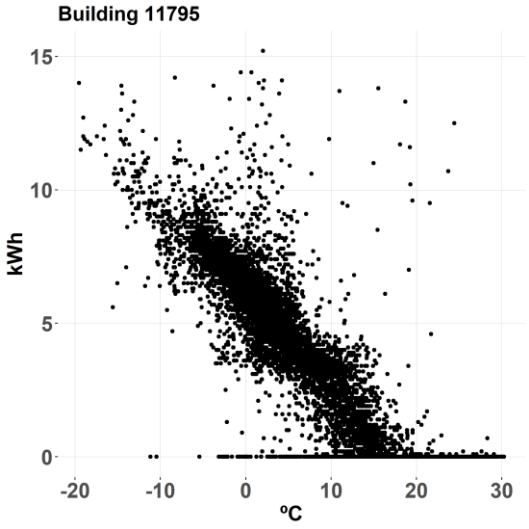
(a)


Building 11794
kWh
ºC
-20
-10
0
10
20
30
0
2
4
6

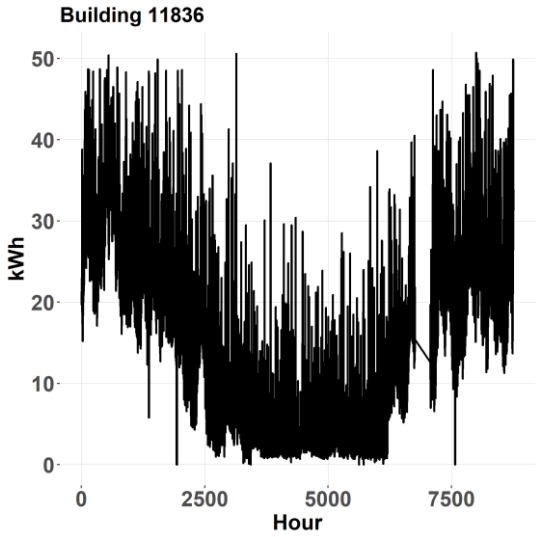
(b)


Building 11795
kWh
Hour
0
2500
5000
7500
0
5
10
15

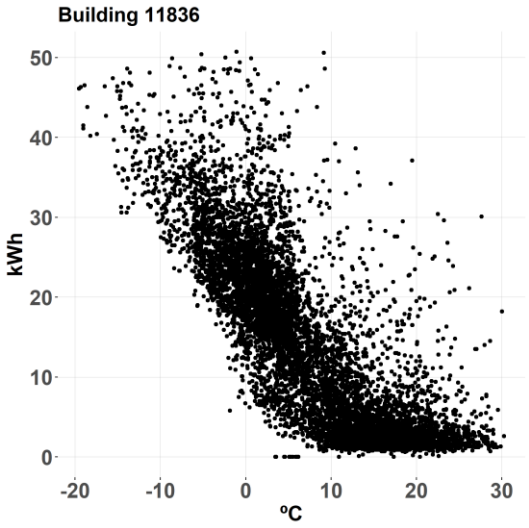
(a)


Building 11795
kWh
°C
-20
-10
0
10
20
30
0
5
10
15

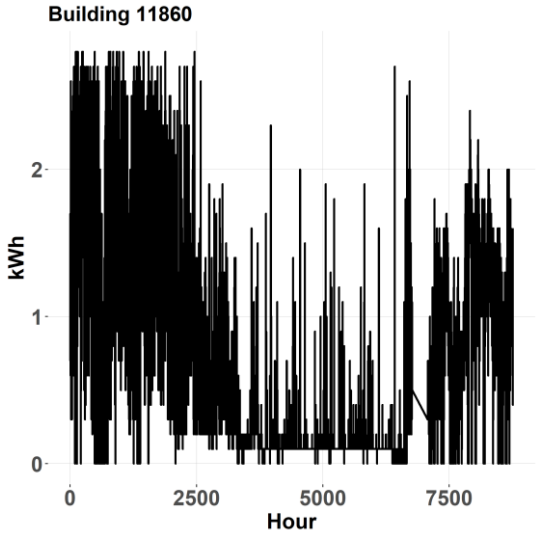
(b)


Building 11836
kWh
Hour
0
2500
5000
7500
0
10
20
30
40
50

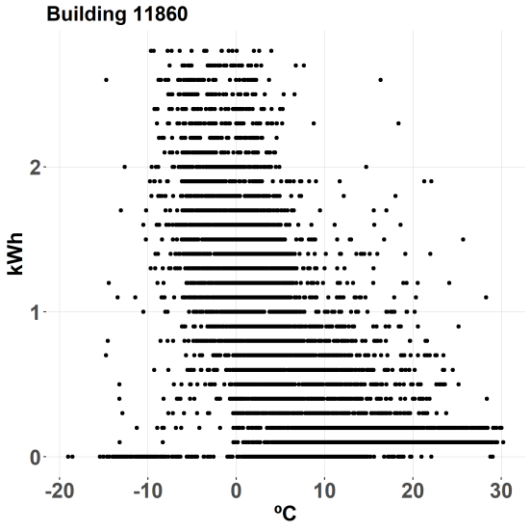
(a)


Building 11836
kWh
°C
-20
-10
0
10
20
30
0
10
20
30
40
50

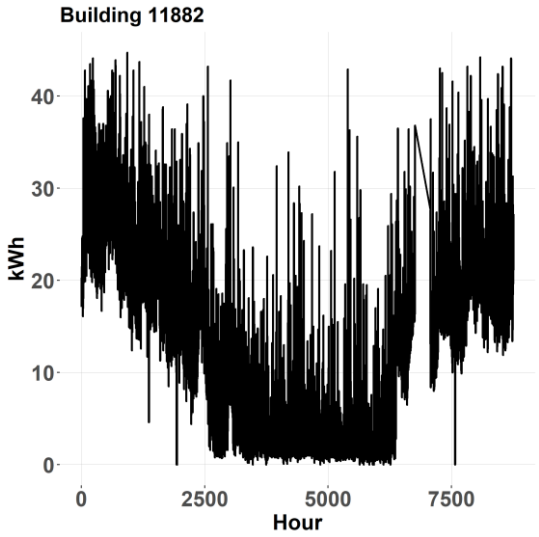
(b)


Building 11860
kWh
0
1
2
0
2500
5000
7500
Hour

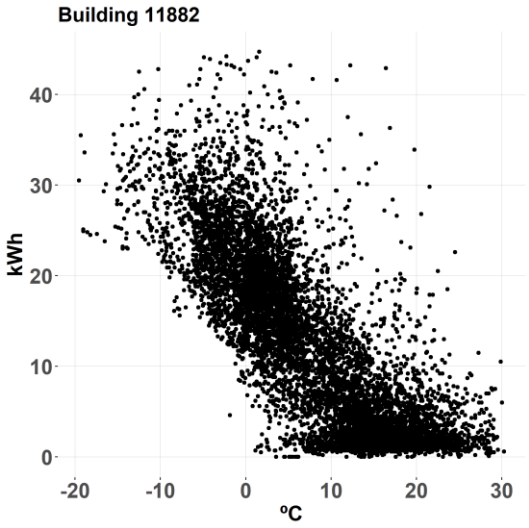
(a)


Building 11860
kWh
0
1
2
-20
-10
0
10
20
30
ºC

(b)


Building 11882
kWh
0
10
20
30
40
0
2500
5000
7500
Hour

(a)


Building 11882
kWh
0
10
20
30
40
-20
-10
0
10
20
30
ºC

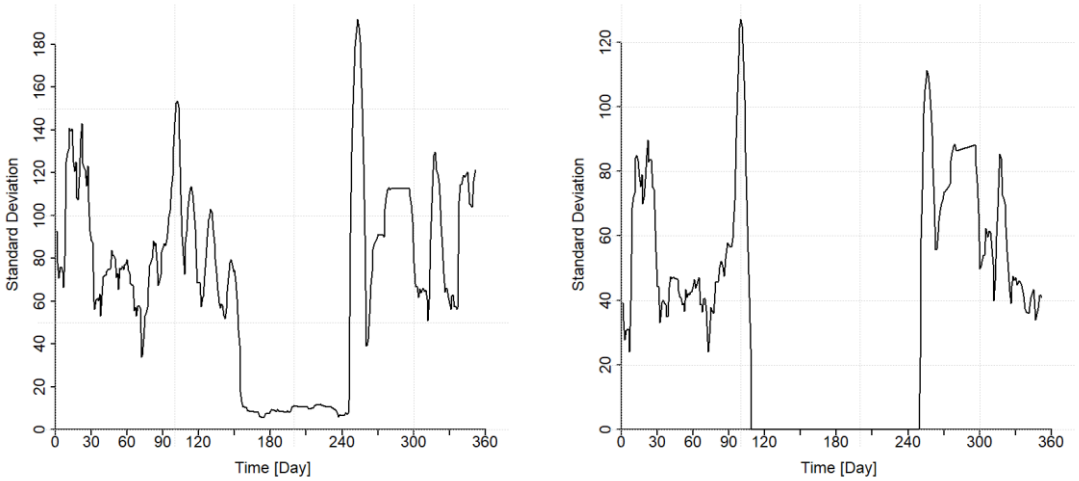
(b)

3. Summer Period Identification Methodology

Buildings under study show two different heating demand profiles in function of the season of the year: SUMMER and REST (these days does not correspond with the exact days of the summer, but it is way to differentiate the two periods). In the summer period, the heating demand does not respond only to climatic conditions, remaining relatively constant over different periods. On the other side, energetic demand in days classified as REST (of the year) is completely dependent mainly in the climatic variables of the moment.

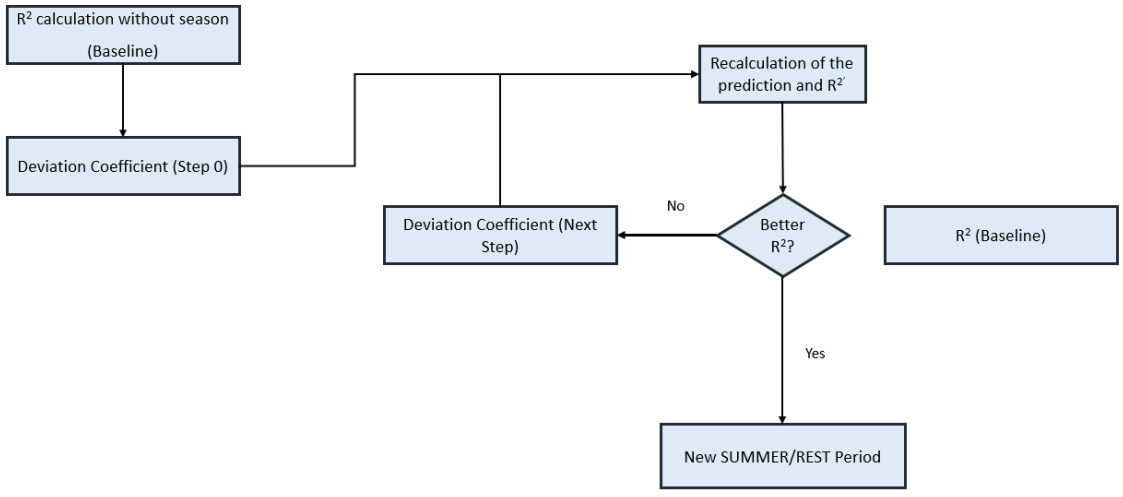
As a result of an exhaust study of the profiles of raw demand data, it is found that standard deviation of the heating demand in winter time is much higher than the deviation in summer time. Different strategies have been tried to identify these two periods: moving average of the errors, variaiblity of the errors etc. However, the strategy that better reaches the objective of this identification is the use of Standar Deviation (SD) of the demand.

Thus, with a daily frequency, from n=1 to N (n = days), days have been gruped by 15 consecutive days. For n = 1, a first group of days is composed by n = [1,16] days. For n = 2, this group is composed by n = [2,17] and so on. For each of this group, SD is calculated, resulting a value that represents the varibility of the demand in that period. If this methodology is applied to the whole year, the resulting curve is shown in the following figure.



As it can be observed in previous figure, it is possible to identify the summer period as the period in which the standard deviation is constant and minimal. In the building of the left image, the minimal standard deviation is not equal to zero in summer, because there is DHW demand. In contrast, the building on the right figure shows no standard deviation in summer, since all the demand in this period is equal to cero.

The following equations show the algorithm used for this summer/winter pattern:

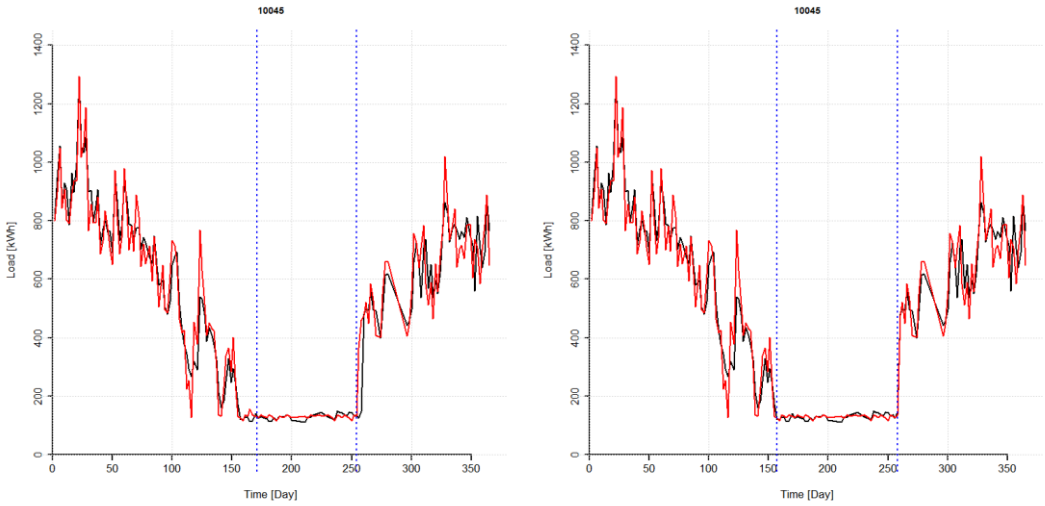


Where:

- *Deviation Coefficient* is a vector from 1.2 up to 4 by 0.1, used in the iterative process.

Therefore, all the values of the SD vector are compared against the minimum value of the SD vector. It is known that the demand in summer period is relatively constant compared with the demand in the winter. This difference is shown in SD vector, and it is used for saving as a TRUE/FALSE command. With each of the values of the *Deviation Coefficient* a new prediction is made, recalculating then the R^2 of the prediction and compared with the previous value. The maximum R^2 for value the corresponding summer/winter days will determine the final classification.

So, in the following images, the evolution of the algorithm by means of the different deviation coefficients used and how the model fits automatically and adapts to the best modelling of the summer/winter periods is shown.



The days that best divide summer and winter time are shown on the right image. This process is replicable for all the buildings in the district. Moreover, some buildings do not have the differentiation between summer and winter. In these cases, all the days are considered as REST.

References

[1] GREN Eesti, “GREN Eesti,” *https://gren.com/ee/*, 2021.

[2] M. Lumbreras *et al.*, “Data driven model for heat load prediction in buildings connected to District Heating by using smart heat meters,” *Energy*, vol. 239, p. 122318, Jan. 2022, doi: 10.1016/J.ENERGY.2021.122318.

[3] M. Lumbreras, G. Diarce, K. Martin, R. Garay-Martinez, and B. Arregi, “Unsupervised recognition and prediction of daily patterns in heating loads in buildings,” *Journal of Building Engineering*, vol. 65, p. 105732, Apr. 2023, doi: 10.1016/J.JOBE.2022.105732.

[4] Energética en la Edificación, “ENEDI Group,” 2023. https://www.enedi.es/

[5] EHU, “ENEDI (Energética en la Edificación),” 2022. https://www.ehu.eus/es/web/enedi/home

[6] FUNDACION TECNALIA RESEARCH & INNOVATION, “RELaTED-REnewable Low TEmperature District,” 2017. https://cordis.europa.eu/project/id/768567/es

[7] European Commission, “Going climate-neutral by 2050: A strategic long-term vision for a prosperous, modern, competitive and climate-neutral EU economy,” *European Commission*, pp. 1–20, 2019.

[8] L. Pérez-Lombard, J. Ortiz, and C. Pout, “A review on buildings energy demand information,” *Energy Build*, vol. 40, no. 3, pp. 394–398, Jan. 2008, doi: 10.1016/J.ENBUILD.2007.03.007.

[9] International Energy Agency, “Global Energy & CO_2 Status Report 2019,” 2019.

[10] European Commision, *Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on energy efficiency, amending Directives 2009/125/EC and 2010/30/EU and repealing Directives 2004/8/EC and 2006/32/EC Text with EEA relevance OJ L 315*. 2012, pp. 1–56.

[11] European Comission, *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 efficienc*. 2018.

[12] H. Averfalk and S. Werner, "Economic benefits of fourth generation district heating," *Energy*, vol. 193, p. 116727, Feb. 2020, doi: 10.1016/J.ENERGY.2019.116727.

[13] H. Lund *et al.*, "4th Generation District Heating (4GDH): Integrating smart thermal grids into future sustainable energy systems," *Energy*, vol. 68, pp. 1–11, Apr. 2014, doi: 10.1016/J.ENERGY.2014.02.089.

[14] S. Werner, "International review of district heating and cooling," *Energy*, vol. 137, pp. 617–631, Oct. 2017, doi: 10.1016/J.ENERGY.2017.04.045.

[15] H. Lund *et al.*, "The status of 4th generation district heating: Research and results," *Energy*, vol. 164, pp. 147–159, Dec. 2018, doi: 10.1016/J.ENERGY.2018.08.206.

[16] J. von Rhein, G. P. Henze, N. Long, and Y. Fu, "Development of a topology analysis tool for fifth-generation district heating and cooling networks," *Energy Convers Manag*, vol. 196, pp. 705–716, Sep. 2019, doi: 10.1016/J.ENCONMAN.2019.05.066.

[17] H. Lund *et al.*, "Perspectives on fourth and fifth generation district heating," *Energy*, vol. 227, p. 120520, Jul. 2021, doi: 10.1016/J.ENERGY.2021.120520.

[18] M. Lumbreras and R. Garay, "Energy & economic assessment of façade-integrated solar thermal systems combined with ultra-low temperature district-heating," *Renew Energy*, vol. 159, pp. 1000–1014, Oct. 2020, doi: 10.1016/J.RENENE.2020.06.019.

[19] S. Moser and S. Lassacher, "External use of industrial waste heat - An analysis of existing implementations in Austria," *J Clean Prod*, vol. 264, p. 121531, Aug. 2020, doi: 10.1016/J.JCLEPRO.2020.121531.

[20] J. Pelda, F. Stelter, and S. Holler, “Potential of integrating industrial waste heat and solar thermal energy into district heating networks in Germany,” *Energy*, vol. 203, p. 117812, Jul. 2020, doi: 10.1016/J.ENERGY.2020.117812.

[21] M. Lumbreras, R. Garay, and V. S. Zabala, “Triple function substation and high-efficiency micro booster heat pump for Ultra Low Temperature District Heating,” *IOP Conf. Ser.: Mater. Sci. Eng*, vol. 609, 2019, doi: 10.1088/1757-899X/609/5/052008.

[22] ACCADEMIA EUROPEA DI BOLZANO, “FLEXYNETS,” 2015. https://cordis.europa.eu/project/id/649820

[23] J. Song, F. Wallin, and H. Li, “A Dynamic Pricing Mechanism for District Heating – Based on a levelized cost of heat and prediction of total heat demand,” 2017. doi: 978-91-7673-408-7.

[24] Eurostat, “Land prices vary considerably between and within Member States,” 2018.

[25] J. McCarthy, “WHAT IS ARTIFICIAL INTELLIGENCE?,” *Formal Standford*, 2007, [Online]. Available: https://www-formal.stanford.edu/jmc/whatisai.pdf

[26] G. W. G. W. I. I. Liu X, “Benchmarking smart meter data analytics,” in *In Proc of the 18th international conference on extending database technology*, 2015, pp. 385–396.

[27] S. Darby, “Smart metering: what potential for householder engagement?,” *Building Research & Information*, vol. 38, no. 5, pp. 442–457, Oct. 2010, doi: 10.1080/09613218.2010.492660.

[28] M. Vesterlund, A. Toffolo, and J. Dahl, “Optimization of multi-source complex district heating network, a case study,” *Energy*, vol. 126, pp. 53–63, May 2017, doi: 10.1016/J.ENERGY.2017.03.018.

[29] K. Lichtenegger, D. Wöss, C. Halmdienst, E. Höftberger, C. Schmidl, and T. Pröll, “Intelligent heat networks: First results of an energy-information-cost-model,”

Sustainable Energy, Grids and Networks, vol. 11, pp. 1–12, Sep. 2017, doi: 10.1016/J.SEGAN.2017.05.001.

[30] M. Lumbreras, G. Diarce, K. Martin-Escudero, A. Campos-Celador, and P. Larrinaga, "Design of district heating networks in built environments using GIS: A case study in Vitoria-Gasteiz, Spain," *J Clean Prod*, vol. 349, p. 131491, May 2022, doi: 10.1016/J.JCLEPRO.2022.131491.

[31] U.S. Department of Energy, "EnergyPlus TM." 2018. [Online]. Available: https://energyplus.net/

[32] S. A. et al Klein, "TRNSYS 18: A Transient System Simulation Program." Solar Energy Laboratory, University of Wisconsin, Madison, USA, 2017. [Online]. Available: http://sel.me.wisc.edu/trnsys.

[33] P. A. Strachan and L. Vandaele, "Case studies of outdoor testing and analysis of building components," *Build Environ*, vol. 43, no. 2, pp. 129–142, Feb. 2008, doi: 10.1016/j.buildenv.2006.10.043.

[34] H. Madsen and J. Holst, "Estimation of continuous-time models for the heat dynamics of a building," *Energy Build*, vol. 22, no. 1, pp. 67–79, Mar. 1995, doi: 10.1016/0378-7788(94)00904-X.

[35] K. K. Andersen, H. Madsen, and L. H. Hansen, "Modelling the heat dynamics of a building using stochastic differential equations," *Energy Build*, vol. 31, no. 1, pp. 13–24, Jan. 2000, doi: 10.1016/S0378-7788(98)00069-3.

[36] P. Bacher and H. Madsen, "Identifying suitable models for the heat dynamics of buildings," *Energy Build*, vol. 43, no. 7, pp. 1511–1522, Jul. 2011, doi: 10.1016/j.enbuild.2011.02.005.

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

[38] D. E. Kissock, J. K.; Haberl, J. S.; Claridge, *Change-Point Linear and Multiple-Linear Inverse Building Energy Analysis Models*. Energy Systems Laboratory, Texas A&M University, 2002. [Online]. Available: https://hdl.handle.net/1969.1/2847

[39] C. C. Holt, "Forecasting seasonals and trends by exponentially weighted moving averages," *Int J Forecast*, vol. 20, no. 1, pp. 5–10, Jan. 2004, doi: 10.1016/J.IJFORECAST.2003.09.015.

[40] L. Ferbar Tratar, B. Mojškerc, and A. Toman, "Demand forecasting with four-parameter exponential smoothing," *Int J Prod Econ*, vol. 181, pp. 162–173, Nov. 2016, doi: 10.1016/J.IJPE.2016.08.004.

[41] L. Ferbar Tratar and E. Strmčnik, "The comparison of Holt–Winters method and Multiple regression method: A case study," *Energy*, vol. 109, pp. 266–276, Aug. 2016, doi: 10.1016/J.ENERGY.2016.04.115.

[42] Z. Verbai, Á. Lakatos, and F. Kalmár, "Prediction of energy demand for heating of residential buildings using variable degree day," *Energy*, vol. 76, pp. 780–787, Nov. 2014, doi: 10.1016/J.ENERGY.2014.08.075.

[43] A. Tureczek and P. Nielsen, "Structured Literature Review of Electricity Demand Classification Using Smart Meter Data," *Energies (Basel)*, vol. 10, no. 5, p. 584, Apr. 2017, doi: 10.3390/en10050584.

[44] Wernstedt F, Davidsson P, and Johansson C, "Demand side management in district heating systems," in *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, 2015, p. 272.

[45] F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," *Appl Energy*, vol. 141, pp. 190–199, Mar. 2015, doi: 10.1016/j.apenergy.2014.12.039.

[46] H. Y. Chang, J. A. Thomson, and X. Chen, "Microarray Analysis of Stem Cells and Differentiation," *Handbook of Stem Cells*, vol. 1, pp. 399–407, 2013, doi: 10.1016/B978-0-12-385942-6.00034-2.

[47] C. Madeira do Carmo and T. H. Christensen, “Cluster analysis of residential heat load profiles and the role of technical and household characteristics,” *Energy Build*, vol. 125, Aug. 2016, doi: 10.1016/j.enbuild.2016.04.079.

[48] J. J. López, J. A. Aguado, F. Martín, F. Muñoz, A. Rodríguez, and J. E. Ruiz, “Hopfield–K-Means clustering algorithm: A proposal for the segmentation of electricity customers,” *Electric Power Systems Research*, vol. 81, no. 2, pp. 716–724, Feb. 2011, doi: 10.1016/j.epsr.2010.10.036.

[49] A. Albert and R. Rajagopal, “Smart Meter Driven Segmentation: What Your Demand Says About You,” *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4019–4030, Nov. 2013, doi: 10.1109/TPWRS.2013.2266122.

[50] A. Ozawa, R. Furusato, and Y. Yoshida, “Determining the relationship between a household’s lifestyle and its electricity demand in Japan by analyzing measured electric load profiles,” *Energy Build*, vol. 119, pp. 200–210, May 2016, doi: 10.1016/j.enbuild.2016.03.047.

[51] M. N. Q. Macedo, J. J. M. Galo, L. A. L. de Almeida, and A. C. de C. Lima, “Demand side management using artificial neural networks in a smart grid environment,” *Renewable and Sustainable Energy Reviews*, vol. 41, pp. 128–133, Jan. 2015, doi: 10.1016/j.rser.2014.08.035.

[52] K. B. Lindberg, P. Seljom, H. Madsen, D. Fischer, and M. Korpås, “Long-term electricity load forecasting: Current and future trends,” *Util Policy*, vol. 58, pp. 102–119, Jun. 2019, doi: 10.1016/j.jup.2019.04.001.

[53] F. M. Andersen, H. V. Larsen, and T. K. Boomsma, “Long-term forecasting of hourly electricity load: Identification of demand profiles and segmentation of customers,” *Energy Convers Manag*, vol. 68, pp. 244–252, Apr. 2013, doi: 10.1016/j.enconman.2013.01.018.

[54] Y. Jang, E. Byon, E. Jahani, and K. Cetin, “On the long-term density prediction of peak electricity load with demand side management in buildings,” *Energy Build*, vol. 228, p. 110450, Dec. 2020, doi: 10.1016/J.ENBUILD.2020.110450.

[55] Y. Hu, J. Li, M. Hong, J. Ren, and Y. Man, “Industrial artificial intelligence based energy management system: Integrated framework for electricity load forecasting and fault prediction,” *Energy*, vol. 244, p. 123195, Apr. 2022, doi: 10.1016/J.ENERGY.2022.123195.

[56] C. H. Jin *et al.*, “A SOM clustering pattern sequence-based next symbol prediction method for day-ahead direct electricity load and price forecasting,” *Energy Convers Manag*, vol. 90, pp. 84–92, Jan. 2015, doi: 10.1016/J.ENCONMAN.2014.11.010.

[57] Z. Dong, J. Liu, B. Liu, K. Li, and X. Li, “Hourly energy demand prediction of an office building based on ensemble learning and energy demand pattern classification,” *Energy Build*, vol. 241, Jun. 2021, doi: 10.1016/j.enbuild.2021.110929.

[58] C. Fan, F. Xiao, Z. Li, and J. Wang, “Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review,” *Energy Build*, vol. 159, Jan. 2018, doi: 10.1016/j.enbuild.2017.11.008.

[59] K. Zhou, S. Yang, and Z. Shao, “Household monthly electricity demand pattern mining: A fuzzy clustering-based model and a case study,” *J Clean Prod*, vol. 141, Jan. 2017, doi: 10.1016/j.jclepro.2016.09.165.

[60] Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, and J. Li, “A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis,” *Energy and Built Environment*, vol. 1, no. 2, Apr. 2020, doi: 10.1016/j.enbenv.2019.11.003.

[61] J. Y. Park, X. Yang, C. Miller, P. Arjunan, and Z. Nagy, “Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings

using a large and diverse dataset," *Appl Energy*, vol. 236, Feb. 2019, doi: 10.1016/j.apenergy.2018.12.025.

[62] L. Wen, K. Zhou, and S. Yang, "A shape-based clustering method for pattern recognition of residential electricity demand," *J Clean Prod*, vol. 212, Mar. 2019, doi: 10.1016/j.jclepro.2018.12.067.

[63] Z. Ma, H. Li, Q. Sun, C. Wang, A. Yan, and F. Starfelt, "Statistical analysis of energy demand patterns on the heat demand of buildings in district heating systems," *Energy Build*, vol. 85, Dec. 2014, doi: 10.1016/j.enbuild.2014.09.048.

[64] H. Gadd and S. Werner, "Fault detection in district heating substations," *Appl Energy*, vol. 157, Nov. 2015, doi: 10.1016/j.apenergy.2015.07.061.

[65] P. Gianniou, X. Liu, A. Heller, P. S. Nielsen, and C. Rode, "Clustering-based analysis for residential district heating data," *Energy Convers Manag*, vol. 165, pp. 840–850, Jun. 2018, doi: 10.1016/J.ENCONMAN.2018.03.015.

[66] A. M. Tureczek, P. S. Nielsen, H. Madsen, and A. Brun, "Clustering district heat exchange stations using smart meter demand data," *Energy Build*, vol. 182, Jan. 2019, doi: 10.1016/j.enbuild.2018.10.009.

[67] E. Calikus, S. Nowaczyk, A. Sant'Anna, H. Gadd, and S. Werner, "A data-driven approach for discovering heat load patterns in district heating," *Appl Energy*, vol. 252, Oct. 2019, doi: 10.1016/j.apenergy.2019.113409.

[68] H. Johra, D. Leiria, P. Heiselberg, A. Marszal-Pomianowska, and T. Tvedebrink, "Treatment and analysis of smart energy meter data from a cluster of buildings connected to district heating: A Danish case," *E3S Web of Conferences*, vol. 172, pp. 2–9, 2020, doi: 10.1051/e3sconf/202017212004.

[69] X. Liu, Y. Ding, H. Tang, and F. Xiao, "A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity demand data," *Energy Build*, vol. 231, p. 110601, Jan. 2021, doi: 10.1016/J.ENBUILD.2020.110601.

[70] G. le Ray and P. Pinson, “Online adaptive clustering algorithm for load profiling,” *Sustainable Energy, Grids and Networks*, vol. 17, p. 100181, Mar. 2019, doi: 10.1016/j.segan.2018.100181.

[71] E. Calikus, S. Nowaczyk, A. Sant’Anna, H. Gadd, and S. Werner, “A data-driven approach for discovering heat load patterns in district heating,” *Appl Energy*, vol. 252, p. 113409, Oct. 2019, doi: 10.1016/j.apenergy.2019.113409.

[72] E. Dotzauer, “Simple model for prediction of loads in district-heating systems,” *Appl Energy*, vol. 73, no. 3–4, pp. 277–284, Nov. 2002, doi: 10.1016/S0306-2619(02)00078-8.

[73] A. J. Heller, “Heat-load modelling for large systems,” *Appl Energy*, vol. 72, no. 1, pp. 371–387, May 2002, doi: 10.1016/S0306-2619(02)00020-X.

[74] S. Grosswindhager, A. Voigt, and M. Kozek, “Online Short-Term Forecast of System Heat Load in District Heating Networks,” *In Proceedings of the 31st International Symposium on Forecasting*, no. 1, pp. 1–8, 2011, [Online]. Available: http://www.forecasters.org/submissions/GROSSWINDHAGERSTEFANISF2011.pdf

[75] S. Paudel *et al.*, “A relevant data selection method for energy demand prediction of low energy building based on support vector machine,” *Energy Build*, vol. 138, pp. 240–256, Mar. 2017, doi: 10.1016/j.enbuild.2016.11.009.

[76] M. Dahl, A. Brun, and G. B. Andresen, “Using ensemble weather predictions in district heating operation and load forecasting,” *Appl Energy*, vol. 193, pp. 455–465, May 2017, doi: 10.1016/J.APENERGY.2017.02.066.

[77] A. Sandberg, F. Wallin, H. Li, and M. Azaza, “An Analyze of Long-term Hourly District Heat Demand Forecasting of a Commercial Building Using Neural Networks,” *Energy Procedia*, vol. 105, pp. 3784–3790, May 2017, doi: 10.1016/J.EGYPRO.2017.03.884.

[78] L. Lei, W. Chen, B. Wu, C. Chen, and W. Liu, “A building energy demand prediction model based on rough set theory and deep learning algorithms,” *Energy Build*, vol. 240, p. 110886, Jun. 2021, doi: 10.1016/J.ENBUILD.2021.110886.

[79] P. Potočnik, P. Škerl, and E. Govekar, “Machine-learning-based multi-step heat demand forecasting in a district heating system,” *Energy Build*, vol. 233, p. 110673, Feb. 2021, doi: 10.1016/J.ENBUILD.2020.110673.

[80] J. Sauer, V. C. Mariani, L. dos Santos Coelho, M. H. D. M. Ribeiro, and M. Rampazzo, “Extreme gradient boosting model based on improved Jaya optimizer applied to forecasting energy demand in residential buildings,” *Evolving Systems*, vol. 13, no. 4, pp. 577–588, Aug. 2022, doi: 10.1007/s12530-021-09404-2.

[81] N. P. Sakkas and R. Abang, “Thermal load prediction of communal district heating systems by applying data-driven machine learning methods,” *Energy Reports*, vol. 8, pp. 1883–1895, Nov. 2022, doi: 10.1016/J.EGYR.2021.12.082.

[82] A. Zhao, L. Mi, X. Xue, J. Xi, and Y. Jiao, “Heating load prediction of residential district using hybrid model based on CNN,” *Energy Build*, vol. 266, p. 112122, Jul. 2022, doi: 10.1016/j.enbuild.2022.112122.

[83] University of Tartu, Institute of Physics, and Laboratory of Environmental Physics, “http://meteo.physic.ut.ee/?lang=en,” 2021.

[84] M. Kottek, J. Grieser, C. Beck, B. Rudolf, and F. Rubel, “World Map of the Köppen-Geiger climate classification updated,” *Meteorologische Zeitschrift*, vol. 15, no. 3, pp. 259–263, Jul. 2006, doi: 10.1127/0941-2948/2006/0130.

[85] W. Weibull, “Wide applicability,” *J Appl Mech*, vol. 103, pp. 293–297, 1951.

[86] Karmstrup, “https://www.kamstrup.com/en-us/heat-solutions/heat-meters/multical-603,” 2021.

[87] EN, *EN 1434-1:2015, Heat meters. Part 1: General requirements*. 2015.

[88] S. Hammarsten, “A critical appraisal of energy-signature models,” *Appl Energy*, vol. 26, no. 2, pp. 97–110, Jan. 1987, doi: 10.1016/0306-2619(87)90012-2.

[89] H. A. Nielsen and H. Madsen, “Modelling the heat demand in district heating systems using a grey-box approach,” *Energy Build*, vol. 38, no. 1, pp. 63–71, Jan. 2006, doi: 10.1016/J.ENBUILD.2005.05.002.

[90] K. M. Powell, A. Sriprasad, W. J. Cole, and T. F. Edgar, “Heating, cooling, and electrical load forecasting for a large-scale district energy system,” *Energy*, vol. 74, no. C, pp. 877–885, Sep. 2014, doi: 10.1016/J.ENERGY.2014.07.064.

[91] R Core Team, “R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing.” Austria, 2013.

[92] N. C. Schwertman, M. A. Owens, and R. Adnan, “A simple more general boxplot method for identifying outliers,” *Comput Stat Data Anal*, vol. 47, no. 1, pp. 165–174, Aug. 2004, doi: 10.1016/J.CSDA.2003.10.012.

[93] A. Li, M. Feng, Y. Li, and Z. Liu, “Application of Outlier Mining in Insider Identification Based on Boxplot Method,” *Procedia Comput Sci*, vol. 91, pp. 245–251, Jan. 2016, doi: 10.1016/J.PROCS.2016.07.069.

[94] M. Ester, H. Kriegel, X. Xu, and D. Miinchen, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” in *Proceedings of the 2nd ACM SIGKDD*, Portland, Oregon, 1996, pp. 226–231.

[95] M. Hashler, M. Piekenbrock, S. Arya, and D. Mount, “R, Package ’dbscan’ 2020 .” 2021.

[96] X. Liu, Y. Ding, H. Tang, and F. Xiao, “A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity demand data,” *Energy Build*, vol. 231, Jan. 2021, doi: 10.1016/j.enbuild.2020.110601.

[97] A. Rajabi *et al.*, "A pattern recognition methodology for analyzing residential customers load data and targeting demand response applications," *Energy Build*, vol. 203, Nov. 2019, doi: 10.1016/j.enbuild.2019.109455.

[98] S. Haben, C. Singleton, and P. Grindrod, "Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data," *IEEE Trans Smart Grid*, vol. 7, no. 1, Jan. 2016, doi: 10.1109/TSG.2015.2409786.

[99] M. Yang, C. Xi, J. Wang, Z. Feng, and S. Cao, "An interactive design framework for large-scale public buildings based on comfort and carbon abatement," *Energy Build*, vol. 279, p. 112679, Jan. 2023, doi: 10.1016/J.ENBUILD.2022.112679.

[100] K. Li, J. Zhang, X. Chen, and W. Xue, "Building's hourly electrical load prediction based on data clustering and ensemble learning strategy," *Energy Build*, vol. 261, p. 111943, Apr. 2022, doi: 10.1016/J.ENBUILD.2022.111943.

[101] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, p. Vol. 1, No. 14, pp. 281–297.

[102] K. Ghanem, "Towards More Accurate Clutering Method by Using Dynamic Time Warping," *International Journal of Data Mining & Knowledge Management Process*, vol. 3, no. 2, pp. 107–118, Mar. 2013, doi: 10.5121/ijdkp.2013.3207.

[103] A. Sarda, "dtwclust: Time Series Clustering Along with Optimizations for the Dynamic Time Warping Distance." 2022. [Online]. Available: https://cran.r-project.org/package=dtwclust

[104] R. Kalaba and E. Ruspini, "Identification of parameters in nonlinear boundary-value problems," *J Optim Theory Appl*, vol. 4, no. 6, pp. 371–377, Dec. 1969, doi: 10.1007/BF00927689.

[105] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, Jan. 1973, doi: 10.1080/01969727308546046.

[106] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Boston, MA: Springer US, 1981. doi: 10.1007/978-1-4757-0450-1.

[107] Z. Cebeci, F. Yildiz, A. T. Kavlak, C. Cebeci, and H. Onder, "Probabilistic and Possibilistic Cluster Analysis." 2020. [Online]. Available: https://cran.r-project.org/package=ppclust

[108] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 2nd Editio. San Diego: Academic Press, 2003.

[109] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J Comput Appl Math*, vol. 20, pp. 53–65, 1987.

[110] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans Pattern Anal Mach Intell*, vol. 2, pp. 224–227, 1979.

[111] J. Dunn, "Well separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, vol. 4, pp. 95–104, 1974.

[112] F. E. Harrell, "Evaluating the Yield of Medical Tests," *JAMA: The Journal of the American Medical Association*, vol. 247, no. 18, p. 2543, May 1982, doi: 10.1001/jama.1982.03320430047030.

[113] B. Desgraupes, "clusterCrit: Clustering Indices." 2018.

[114] J. R. Quinlan, "Simplifying decision trees," *Int J Man Mach Stud*, vol. 27, no. 3, pp. 221–234, Sep. 1987, doi: 10.1016/S0020-7373(87)80053-6.

[115] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93*, New York, New York, USA: ACM Press, 1993, pp. 207–216. doi: 10.1145/170035.170072.

[116] L. Breiman, JH. Friedman, RA. Olshen, and CJ. Stone, *Classification and Regression Trees*, Wadsworth Inc. 1984.

[117] L. Breiman, “Random forests,” *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[118] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach Learn*, vol. 20, pp. 273–297, 1995, [Online]. Available: https://doi.org/10.1007/BF00994018

[119] A. Capozzoli, M. S. Piscitelli, S. Brandi, D. Grassi, and G. Chicco, “Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings,” *Energy*, vol. 157, pp. 336–352, Aug. 2018, doi: 10.1016/j.energy.2018.05.127.

[120] J. L. Viegas, S. M. Vieira, R. Melício, V. M. F. Mendes, and J. M. C. Sousa, “Classification of new electricity customers based on surveys and smart metering data,” *Energy*, vol. 107, pp. 804–817, Jul. 2016, doi: 10.1016/j.energy.2016.04.065.

[121] D. Vercamer, B. Steurtewagen, D. van den Poel, and F. Vermeulen, “Predicting Consumer Load Profiles Using Commercial and Open Data,” *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3693–3701, Sep. 2016, doi: 10.1109/TPWRS.2015.2493083.

[122] V. Fabi, S. P. Corgnati, and R. K. Andersen, “Main physical environmental variables driving occupant behaviour with regard to natural ventilation,” in *Proceedings of the 5th International Building Physics Conference*, 2012.

[123] T. Therneau, B. Atkinson, B. Ripley, and M. B. Ripley, “R Package ’rpart’.” 2020.

[124] “class: Functions for Classification,” 2022. https://cran.r-project.org/package=class

[125] M. Majka, “naivebayes: High Performance Implementation of the Naive Bayes Algorithm.” 2020. [Online]. Available: https://cran.r-project.org/package=naivebayes

[126] D. Meyer, E. Dimitriadou, A. Weingessel, and F. Leisch, “e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.” 2022. [Online]. Available: https://cran.r-project.org/package=e1071

[127] A. Liaw and M. Wiener, “randomForest: Breiman and Cutler’s Random Forests for Classification and Regression.” pp. 18–22, 2002. [Online]. Available: https://cran.r-project.org/doc/Rnews/.

[128] F. M. Andersen, H. v. Larsen, and T. K. Boomsma, “Long-term forecasting of hourly electricity load: Identification of demand profiles and segmentation of customers,” *Energy Convers Manag*, vol. 68, pp. 244–252, Apr. 2013, doi: 10.1016/J.ENCONMAN.2013.01.018.

[129] S. Chen, Y. Ren, D. Friedrich, Z. Yu, and J. Yu, “Prediction of office building electricity demand using artificial neural network by splitting the time horizon for different occupancy rates,” *Energy and AI*, vol. 5, p. 100093, Sep. 2021, doi: 10.1016/J.EGYAI.2021.100093.

[130] T. Cholewa *et al.*, “On the short term forecasting of heat power for heating of building,” *J Clean Prod*, vol. 307, p. 127232, Jul. 2021, doi: 10.1016/J.JCLEPRO.2021.127232.

[131] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *KDD ’16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. doi: https://doi.org/10.1145/2939672.2939785.

[132] T. Chen, T. He, and M. Benesty, “xgboost: Extreme Gradient Boosting.” 2022. [Online]. Available: https://cran.r-project.org/package=xgboost

[133] U.S. DOE, “DESIGN BUILDER.”

[134] D. Conolly, D. Drysdale, K. Hansen, and T. Novosel, “Creating Hourly Profiles to Model both Demand and Supply Work Package 2 Background Report 2,” 2015.

[135] European Commission, “European-Comission PHOTOVOLTAIC GEOGRAPHICAL INFORMATION SYSTEM,” 2022. https://re.jrc.ec.europa.eu/pvg_tools/en/

[136] M. Lumbreras, R. Garay, and A. G. Marijuan, “Energy meters in District-Heating Substations for Heat Demand Characterization and Prediction Using Machine-Learning Techniques,” *IOP Conf Ser Earth Environ Sci*, vol. 588, no. 3, p. 032007, Nov. 2020, doi: 10.1088/1755-1315/588/3/032007.

[137] A. G. Marijuan, R. Garay, M. Lumbreras, L. Vladic, and R. Savić, “District Heating De-Carbonisation in Belgrade. Multi-Year transition plan,” *IOP Conf Ser Earth Environ Sci*, vol. 588, no. 5, p. 052034, Nov. 2020, doi: 10.1088/1755-1315/588/5/052034.

[138] M. Lumbreras, K. Martin-Escudero, G. Diarce, R. Garay-Martinez, and R. Mulero, “Unsupervised Clustering for Pattern Recognition of Heating Energy Demand in Buildings Connected to District-Heating Network,” *2021 6th International Conference on Smart and Sustainable Technologies (SpliTech)*, pp. 1–5, 2021, doi: 10.23919/SpliTech52315.2021.9566420.

[139] R. Garay-Martinez, B. Arregi, M. Lumbreras, B. Zurro, J. M. Gonzalez, and J. L. Hernandez, “Data driven process for the energy assessment of building envelope retrofits,” *E3S Web of Conferences*, vol. 172, p. 25001, Jun. 2020, doi: 10.1051/e3sconf/202017225001.

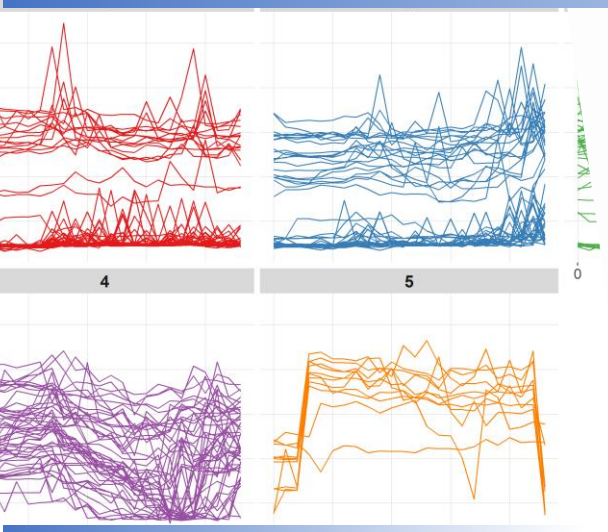
[140] A. G. Marijuan, R. Garay, M. Lumbreras, V. Sánchez, O. Macias, and J. P. S. de Rozas, “RELaTED Project: New Developments on Ultra-Low Temperature District Heating Networks,” in *The 8th Annual International Sustainable Places Conference (SP2020) Proceedings*, Basel Switzerland: MDPI, Dec. 2020, p. 8. doi: 10.3390/proceedings2020065008.

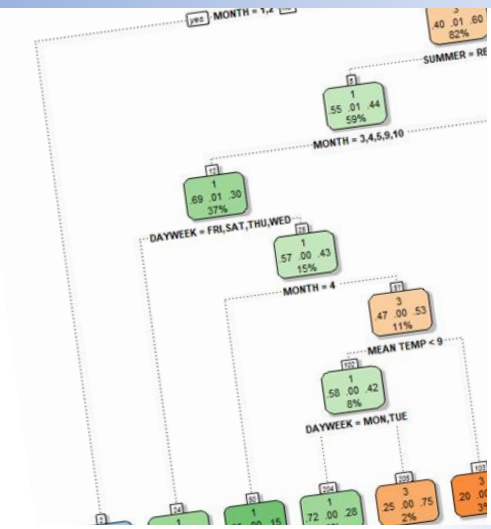
[141] M. Lumbreras, K. Martin-Escudero, G. Diarce, and R. Garay-Martinez, “Data-Driven Analysis of Heating Demand in Buildings Connected to District-Heating:

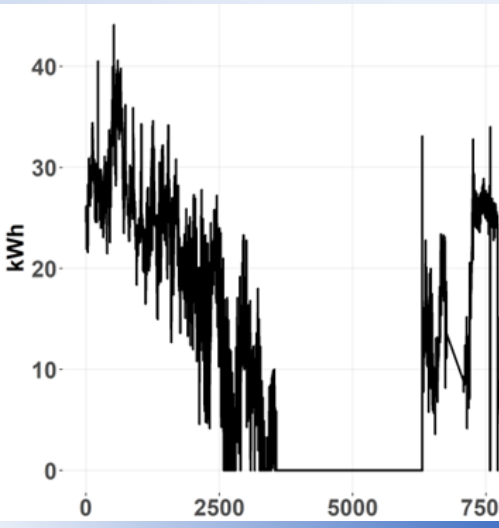
Pattern Recognition and Demand Prediction," in *12th European Conference on Energy Efficiency and Sustainability in Architecture and Planning (EESAP12)*, 2021.

[142] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biol Cybern*, vol. 43, no. 1, pp. 59–69, 1982, doi: 10.1007/BF00337288.

[143] T. Y. Kim and S. B. Cho, "Predicting residential energy demand using CNN-LSTM neural networks," *Energy*, vol. 182, pp. 72–81, Sep. 2019, doi: 10.1016/J.ENERGY.2019.05.230.







enedi

Universidad del País Vasco
Euskal Herriko Unibertsitatea

BILBOKO INGENIARITZA ESKOLA
ESCUELA DE INGENIERÍA DE BILBAO