FACULTY OF ENGINEERING AND ARCHITECTURE

Constructing Realistic Synthetic Load Profiles Based on Privacy Sensitive Residential Smart Meter Data

Robbert Claeys

Doctoral dissertation submitted to obtain the academic degree of
Doctor of Electromechanical Engineering

Supervisors

Prof. Jan Desmet, PhD - Prof. Jos Knockaert, PhD

Department of Electromechanical, Systems and Metal Engineering
Faculty of Engineering and Architecture, Ghent University

October 2024

GHENT UNIVERSITY

ISBN 978-94-6355-898-3
NUR 964
Wettelijk depot: D/2024/10.500/103

Members of the Examination Board

Chair

Prof. Filip De Turck, PhD, Ghent University

Other members entitled to vote

Prof. Geert Deconinck, PhD, KU Leuven
Prof. Chris Develder, PhD, Ghent University
Sam Hamels, PhD, Ghent University
Prof. Dirk Saelens, PhD, KU Leuven

Supervisors

Prof. Jan Desmet, PhD, Ghent University
Prof. Jos Knockaert, PhD, Ghent University

This research has been conducted at **EELab/Lemcko**, the branch of the Electrical Energy Laboratory of Ghent University located on its campus in Kortrijk.

Acknowledgments

I think I did pretty well, considering I started out with nothing but a bunch of blank paper.
Steve Martin

Prospective PhD students are often warned that pursuing a PhD can be a challenging and sometimes lonely journey. However, reflecting on the past years as a teaching assistant and PhD student at Ghent University, I feel both incredibly privileged and grateful for the people that are or have been part of my life. Their contributions, support, and companionship made these years fly by and ensured that my own journey has never felt like a solitary endeavor. While a brief mention in these acknowledgments is insufficient to express my gratitude, I nonetheless wish to express my appreciation to those who have helped shape these last few years.

My deepest gratitude goes to my thesis supervisor, Jan Desmet, for his support and for providing me with the opportunity to pursue this PhD. Our biweekly progress meetings more often than not ended in more philosophical discussions, but the steadfast commitment to the supervision of the students under your care is illustrative of the effort you put into creating a productive learning environment for everyone connected to our research group. I am also deeply appreciative of your flexibility, allowing me the freedom to shape my own research interests, and for your understanding when my other responsibilities occasionally slowed my progress. Additionally, I was fortunate to have a second supervisor, Jos Knockaert, whose external perspective and constructive feedback consistently elevated the quality of my work. Thank you both, I never took your presence and support for granted.

Nothing presented in this thesis was done in isolation, and I am grateful for the amazing colleagues (past and present) at EELab/Lemcko who often served as sparring partners and made it a pleasure to come to the office (the times I actually made it there).

Rémy and Hakim, thank you for the many collaborations, as well as for the input you provided for my own work. Your willingness to offer assistance seemed inexhaustible. Toon, Brecht, and Fien: I will forever have fond memories from our collaboration on energy communities, and wish we had more time to realize all we discussed together. Gianni, we started working at EELab/Lemcko on the same day, and I wish to thank you for teaching me so much about battery systems and their control strategies. Your work has proven invaluable as a stepping stone for some of my own research.

A good set of colleagues is invaluable, so I would be remiss not to acknowledge a second group of colleagues from Ghent University, who without a shadow of a doubt have shaped and enriched my experience over these last few years. I have had the honor of serving on the Executive Board of Ghent University for the final two years of my PhD. Femke, Dries, Jochen, Tim: I cannot sufficiently express my gratitude to all of you for entrusting me with this position. While at times challenging to balance these responsibilities with the work for this thesis, I am incredibly proud of the work we have accomplished together over the past few years.

Fortunately, there is a world beyond Ghent University (although there were weeks when it seemed otherwise). I have been lucky enough to be part of the most inspiring gang of friends, and it is hard to express my gratitude for every break they have provided, big or small, away from the demands of university life. Lunch meetings, movie nights, weekly badminton sessions and shared TV series binges were regular fixtures over the years. Our annual barbecues, getaways, Christmas parties, and New Year's Eve celebrations all became traditions to look forward to, while our trips through the USA and Japan will undoubtedly stand out as highlights for years to come.

To my parents, thank you for your unwavering support and for shaping me to be the person I am today. I would not be who I am today without the example both of you set. I am also deeply grateful to my brother Pieter (who is ten times the researcher I can ever hope to be) for his invaluable guidance at the very start of my PhD journey.

Finally, to Sarah, the single most impressive person I know. You have always been my go-to person these past few years, always there for me in every way imaginable. I could not thank you more for your patience, input, and day-to-day inspiration. Any acknowledgment for your role in this work would inevitably fall short, as you unconditionally supported me, fully aware that you will have to endure hearing "*doctor's orders*" every time we order takeaway in the near future.

Robbert Claeys
Ghent, May 2024

Contents

List of Symbols

Throughout this PhD dissertation, italicized variables indicate scalars, whereas italicized and bold–faced variables indicate vectors.

Alphanumeric symbols	
A_k	k-level DWT approximation function
$\mathcal{B}^i_{\mathcal{I}^x}$	Subset of all $P^i(t)$ that occur in $\mathcal{I}^x$
$\mathcal{D}^i$	Subset of all $P^i(t)$ values labeled as peaks
$\mathcal{D}^i_{\mathcal{I}^x}$	subset of $\mathcal{D}^i$ that occurs in $\mathcal{I}^x$
$H(x)$	Shannon entropy
D_k	k-level DWT detail function
$f^{i,c}_{\mathcal{I}^x}$	Fraction of the consumption of consumer i in $\mathcal{I}^x$
$f^{i,p}_{\mathcal{I}^x}$	Fraction of the peak demands of consumer i in $\mathcal{I}^x$
$\mathcal{I}^d$	Time period at the daily level
$\mathcal{I}^w$	Time period at the weekly level
$\mathcal{I}^x$	Time period at a generic level
$l_1(p, q)$	Wasserstein-1 distance between two distributions p and q
$L^2(\mathbb{R})$	The vector space of square integrable functions
L_F	Load factor
$L^{i,d}_F$	Load factor of consumer i on day d
$\mathcal{O}$	Time complexity of an algorithm
$p(x)$	Probability of a possible outcome x
$\hat{p}(x)$	Learned distribution of $p(x)$
$P^i(t)$	Chronological demand data of consumer i
P^i_{max}	Annual peak demand of consumer i
$\mathcal{P}^i(\tau)$	Load duration curve of consumer i

Greek symbols	
$\Gamma(a, b)$	Gamma function defined by parameters a and b
$\kappa(f(t))$	Curvature function of $f(t)$
μ_i	Mean of the observations in a cluster
$\phi(t)$	Scaling function or father wavelet
$\psi(t)$	Mother wavelet
$\psi^H(t)$	Haar wavelet
σ_κ	Circular shift of κ steps
τ	Normalized time in the LDC expression
τ^*	Threshold value for demand peaks
θ_D	Learnable parameters of the discriminator
θ_G	Learnable parameters of the generator

Mathematical symbols	
$\|F\|$	Absolute value of scalar F
$\|S\|$	Cardinality of a set S
$\langle f, g \rangle$	Inner product of two functions f and g
$\ln$	Natural logarithm
$\hat{f}(\xi)$	Fourier transform of $f(t)$

List of Abbreviations

ACF	Autocorrelation function
AMI	Advanced metering infrastructure
ASHP	Air-source heat pump
BESS	Battery energy storage system
BFS	Backward-forward sweep
CC	Constant current
COP	Coefficient of performance
CV	Constant voltage
DGAN	DoppelGANger
DHW	Domestic hot water
DSO	Distribution system operator
DTA	Data transfer agreement
DWT	Discrete wavelet transform
EMF	Electromagnetic field
EMS	Energy management system
EV	Electric vehicle
GAN	Generative adversarial network
GHG	Greenhouse gas
HC	Hosting capacity
HEMS	Home energy management system
HF	High-frequency
HP	Heat pump
IEA	International Energy Agency
IoT	Internet of Things
IPCC	Intergovernmental Panel on Climate Change
KDE	Kernel density estimation
LDC	Load duration curve
LF	Low-frequency
LR	Learning rate
LSTM	Long short-term memory
LV	Low-voltage
MAE	Mean absolute error

MAPE	Mean absolute percentage error
MPE	Mean percentage error
MRA	Multiresolution analysis
MTLF	Medium-term load forecasting
NDA	Non-disclosure agreement
NIOM	Non-intrusive occupancy monitoring
OLS	Ordinary least squares
pdf	Probability density function
PV	Photovoltaic
RES	Renewable energy systems
RL	Reinforcement learning
RMI	Royal Meteorological Institute
RMSE	Root mean square error
RNN	Recurrent neural network
SCOP	Seasonal coefficient of performance
SCR	Self-consumption ratio
SSR	Self-sufficiency ratio
SH	Space heating
SLP	Synthetic load profile
SoC	State of charge
STLF	Short-term load forecasting
SM	Smart meter
SME	Small and medium-sized enterprise
TSO	Transmission system operator
VPP	Virtual power plant
VSTLF	Very short-term load forecasting
WCSS	Within-cluster sum of squares
wRNG	Weighted random number generator

Samenvatting

Als gevolg van klimaatverandering en de dringende noodzaak om alle sectoren van de samenleving koolstofvrij te maken, staat de residentiële sector voor een kritieke transformatie in een tijdsbestek van minder dan drie decennia. Deze transformatie vereist snelle en substantiële veranderingen, waaronder het wijdverspreid integreren van fotovoltaïsche systemen op daken, een snelle overgang van verwarmingssystemen op gas en olie naar energie-efficiëntere elektrische warmtepompen in goed geïsoleerde huizen, en een versnelde overstap naar elektrische voertuigen in combinatie met een verschuiving naar alternatieve vervoersmodi.

De uitrol van slimme meters in elk huishouden dient als hoeksteen voor deze energietransitie. Ten eerste faciliteren meters met bidirectionele communicatie tussen consumenten, energieleveranciers, en netbeheerders de automatisering van verschillende energiegerelateerde diensten, waardoor de last voor individuele consumenten wordt verlicht en een efficiënter netbeheer mogelijk wordt. Tegelijkertijd bieden slimme meters ongeëvenaarde inzichten in gewoontes van consumenten op granulaire tijdresoluties, waardoor bedrijven de toegevoegde waarde van hun oplossingen nauwkeurig kunnen kwantificeren ten gunste van elke individuele consument. Het is dan ook weinig verrassend dat datagestuurde digitale bedrijfsmodellen in de energiesector snel zijn uitgegroeid tot een miljardenmarkt.

Granulaire gegevens van slimme meters kunnen echter onbedoeld persoonlijke informatie over individuele consumenten vrijgeven aan bedrijven, overheidsinstanties en onderzoeksinstellingen. Daarom hebben regelgevers en wetgevers gereageerd door beperkingen op te leggen aan de publieke toegang tot gegevens van slimme meters, ondanks het aanzienlijke potentieel ervan. In de Europese Unie zijn gegevens van slimme meters bestempeld als persoonsgegevens en het delen en verwerken ervan moet in overeenstemming zijn met de Algemene Verordening Gegevensbescherming van de EU.

Synthetische belastingsprofielen worden ontwikkeld als een oplossing om een balans te vinden tussen het publiek nut van deze gegevens en het behoud van privacy, waarbij privacygevoelige metergegevens worden gebruikt om data te genereren die wel geschikt zijn om onbeperkt te delen. Deze thesis levert een bijdrage in het modelleren van hoogfrequente synthetische datasets op jaarbasis met realistische piekbelasting.

Traditiegetrouw dient **Hoofdstuk 1** van dit proefschrift als inleiding op het onderzoek, waarin de probleemstelling wordt verwoord. De primaire onderzoeksvraag wordt ontleed in enkele secundaire doelstellingen die als tussentijdse mijlpalen zullen fungeren. Elke secundaire onderzoeksvraag wordt behandeld in een apart hoofdstuk.

Om het onderzoek in dit proefschrift in de juiste context te plaatsen en de lezer de nodige achtergrond te geven, worden in **Hoofdstuk 2** verschillende relevante onderwerpen behandeld. Startend van de positionering van de residentiële sector binnen de bredere energiesector worden de inspanningen om deze sector koolstofvrij te maken besproken. Vervolgens worden slimme meters en datagestuurde toepassingen geïntroduceerd, waarbij het huidige marktpotentieel wordt belicht en mogelijke toepassingen van metergegevens op individueel en geaggregeerd niveau worden toegelicht. Tot slot wordt het wetgevend kader voor het delen van gegevens van slimme meters in de Europese Unie besproken, waarbij wordt afgesloten met een overzicht van de soorten persoonsgegevens die kunnen worden afgeleid.

Om tegemoet te komen aan de vraag naar trainingsgegevens voor datagestuurde toepassingen, biedt synthetische modellering van residentiële belastingsprofielen een veelbelovend alternatief voor privacygevoelige metergegevens. Daarom worden in **Hoofdstuk 3** verschillende modelleringstechnieken besproken, van verouderde methoden voor data-anonimisering tot geavanceerde generatieve AI-modellen. De bijdragen van dit onderzoek zijn gericht op top-down modellering, waarbij historische datasets worden gebruikt om gelijkaardige consumenten te groeperen en synthetische belastingsprofielen te genereren. Het meeste onderzoek richt zich echter op energiegegevens voor de korte termijn, voornamelijk dagelijkse of wekelijkse profielen, waardoor er aanzienlijke hiaten zijn voor langetermijntoepassingen.

Het verschil tussen conventionele dagelijkse en wekelijkse belastingsmodellen en de gegevens die in dit onderzoek worden gebruikt, wordt duidelijk in **Hoofdstuk 4**, waar de trainings- en validatiedataset wordt gepresenteerd. De dataset die in dit proefschrift wordt gebruikt, bevat een heel jaar aan verbruiksgegevens die zijn opgenomen met intervallen van 15 minuten, wat neerkomt op 35.040 datapunten per verbruiker, meerdere grootteordes groter dan het aantal observaties in traditionele, korte termijn, belastingsmodellen.

Gezien het complexe samenspel van factoren die samen de output van slimme meters van residentiële consumenten bepalen en de afwezigheid van langetermijnbelastingsmodellen, beargumenteren we dat nauwkeurige synthetische belastingsmodellen op jaarschaal met een resolutie van 15 minuten niet kunnen worden bereikt met één enkele modelleermethode. Deze complexiteit in gedachte stellen we een ontkoppeling van het probleem voor, waarbij voor verschillende tijdschalen geschikte modelleringstechnieken worden gebruikt, elk op maat om de unieke dynamiek vast te leggen die inherent is aan de beschouwde tijdschaal. In **Hoofdstuk 7** en **Hoofdstuk 8** worden de voorgestelde belastingsmodellen op respectievelijk dagelijks en jaarlijks niveau besproken.

Alvorens in te gaan op het modelleren van de belastingsprofielen zelf, wordt in **Hoofdstuk 5** eerst het wiskundige kader voor het definiëren van piekbelastingen geïntroduceerd, gezien het toenemende belang ervan voor moderne toepassingen gericht op vraagrespons en piekscheren. Hoewel de term "*piekvraag*" alom gebruikt wordt door netbeheerders en energiebedrijven, kon er geen rigoureuze definitie gevonden worden die geschikt is voor individuele consumenten. Daarom wordt de lastduurcurve geïntroduceerd om een drempelwaarde te bepalen die uniek is voor elke consument.

Deze definitie wordt gebruikt in een clusteringalgoritme om consumenten met gelijkaardige eigenschappen te groeperen in **Hoofdstuk 6**. Daarin wordt een nieuwe functieset gebouwd op basis van de tijdsgebonden eigenschappen van verbruik en piekvraag. Deze set wordt vervolgens gebruikt in een hiërarchisch clusteringproces om de dataset in 10 clusters te verdelen. Het clusteralgoritme leidt tot compacte clusters met duidelijke verbanden met werkelijke toepassingen waarbij piekvraag in woningen een rol speelt, zoals vraagresponsinitiatieven en de toepasbaarheid van batterijopslagsystemen voor piekscheren. Bovendien laten deze clusters een duidelijk verband zien met de voorspelbaarheid en variabiliteit van consumptie en piekgedrag, waardoor de onevenredige aanwezigheid van pieken in bepaalde tijdsperioden voor elk cluster kan worden gekwantificeerd. Deze geconstrueerde metagegevens op basis van de clusters zullen van grote waarde blijken te zijn tijdens de volgende stappen.

Op dagniveau onderzoekt **Hoofdstuk 7** het gebruik van tijd-frequentie analyse om dagelijkse belastingsprofielen te genereren. De stochastische piekbelastingen worden beschouwd als hoogfrequente componenten die gesuperponeerd worden op een trager variërende, laagfrequente basisbelasting. Een op wavelets gebaseerde multiresolutieanalyse wordt gebruikt om slimme meterdata te ontbinden in hun laagfrequente en hoogfrequente componenten. Vervolgens wordt een stochastische belastingsprofielgenerator op dagniveau geconstrueerd.

Door de hoogfrequente component van een huishouden te herschalen en te verschuiven en deze bovenop de laagfrequente component van een ander huishouden te plaatsen, kan een realistisch synthetisch belastingsprofiel worden verkregen. Er wordt aangetoond dat deze methodologie een nauwkeurige regeling van de verdelingen van en de relatie tussen de dagelijkse piekvraag en het dagelijkse verbruik mogelijk maakt.

Een andere techniek is nodig om jaarprofielen van dagelijkse consumpties te modelleren, aangezien deze meer correlatie vertonen over langere tijdschalen. **Hoofdstuk 8** illustreert dat Generatieve AntagonistenNetwerken (GANs) zeer geschikt zijn voor deze taak. De DoppelGANger-architectuur wordt gebruikt vanwege het vermogen om langetermijncorrelaties te capteren en de mogelijkheid om in batch kortetermijnvariaties op weekniveau te genereren. Op basis van een sensitiviteitsanalyse van de inputs die het generatieproces beïnvloeden, wordt een set parameters geïdentificeerd die de consistente generatie van synthetische data faciliteert die nauw aansluiten bij reële jaarlijkse belastingsprofielen, zowel in het frequentie- als amplitudedomein.

Tot slot bespreekt **Hoofdstuk 9** de validatie van de samengestelde synthetische dataset op een reeks toepassingen van residentiële slimme metergegevens. De dagelijkse belastingsprofielen uit **Hoofdstuk 7** worden geïntegreerd in de jaarprofielen van dagelijkse verbruiken uit **Hoofdstuk 8** om jaarprofielen te verkrijgen met een hoge variabiliteit en een realistische dynamiek op korte en lange termijn. Deze validaties bewijzen de uitstekende overeenkomst tussen de verdeling van echte en synthetische gegevens voor de opgenomen benchmarks, zowel op individueel als op geaggregeerd niveau. Zowel kwalitatieve metadata als een voldoende grote trainingsdataset zijn echter cruciaal om een dataset te verkrijgen met realistische en voldoende diverse belastingsprofielen.

Concluderend kan worden gesteld dat in dit proefschrift een nieuwe tweestapsmethode is voorgesteld om synthetische gegevens te genereren uit privacygevoelige slimme metergegevens. De resultaten bevestigen de bruikbaarheid van deze techniek voor het produceren van hoogfrequente gegevens gedurende een heel jaar, waardoor het een geschikt alternatief is voor slimme metergegevens voor een breed scala aan praktische toepassingen. Omdat de geconstateerde tekortkomingen kunnen worden toegeschreven aan de proof-of-conceptfase die werd uitgevoerd op een beperkte dataset, bieden de voorgestelde methoden potentieel op vlak van zowel schaalbaarheid als toepassing in niet-academische contexten.

Summary

Driven by the urgency of climate change and the pressing need to decarbonize all sectors of society, the residential sector faces a critical transformation in a time frame of less than three decades. This transformation necessitates swift and substantial changes, including the widespread deployment of rooftop photovoltaic systems, a rapid transition from gas and oil-based heating systems to more energy-efficient electric heat pumps in well-insulated homes, and an accelerated adoption of electric vehicles coupled with a shift towards alternative modes of transportation.

The deployment of smart meters at every household serves as a cornerstone for this energy transition. First, meters with bi-directional communication between consumers, energy providers, and grid operators facilitate the automation of various enery-related services, easing the burden on individual consumers while enabling more efficient grid management. At the same time, smart meters provide unprecedented insights into consumer habits at granular time resolutions, allowing organizations to precisely quantify the value proposition of their solutions to benefit each individual consumer. Unsurprisingly, the data-driven digital business models in the energy sector have rapidly evolved into a billion-dollar market.

However, granular smart meter data can unintentionally disclose personal information about individual consumers to companies, public agencies, and research institutions. Consequently, regulators and legislators have responded by imposing restrictions on public access to smart meter data, despite its significant potential for data-driven applications. In the European Union, smart meter data has been labeled as personal data, and its sharing and handling needs to be compliant with the EU's General Data Protection Regulation (GDPR).

Synthetic load modeling has emerged as a solution to balance data utility with privacy preservation, leveraging privacy-sensitive smart meter data to generate synthetic datasets suitable for unrestricted sharing. This PhD dissertation builds upon these techniques, offering advancements in the modeling of annual, high-frequency synthetic datasets that incorporate realistic peak demands.

Following tradition, **Chapter 1** of this dissertation serves as an introduction to our research, articulating our problem statement. Our primary research question is broken down into several secondary objectives that will act as interim milestones. Each secondary research question will be addressed in a separate chapter.

To contextualize the research presented in this dissertation and give the reader the necessary background, **Chapter 2** covers several relevant topics. It begins by situating the residential sector within the larger energy sector and discussing the necessary decarbonization efforts as part of the energy transition. Smart meters and their data-driven applications are subsequently introduced, highlighting the current market potential and exploring possible uses of smart meter data at the individual and aggregated level. Last, the legislative framework governing the sharing of smart meter data in the European Union is reviewed, concluding with an overview of the types of personal data that can be extracted from residential smart meter data.

To meet the demand for accurate training data in data-driven applications, residential load modeling presents a promising alternative to privacy-sensitive smart meter data. As such, **Chapter 3** explores different modeling techniques, starting from legacy data anonymization methods to state-of-the-art generative AI models. This dissertation's contributions are focused on top-down modeling, which employs historical datasets to group similar consumers and generate synthetic load profiles. However, most existing research focuses on short-term energy data, primarily daily or weekly profiles, leaving a significant gap for long-term data applications.

The disparity between conventional daily and weekly load models and the data used in this PhD is made clear in **Chapter 4**, where the training and validation dataset is presented. The dataset used throughout this dissertation features a full year of consumption data recorded at 15-minute intervals, equating to 35,040 data points per consumer, several orders of magnitude larger than the number of data points involved in traditional, short-term, load models.

As such, given the complex interplay of factors that collectively shape the smart meter output of residential consumers and the absence of long-term load models, we argue that accurate synthetic load models at the annual timescale at a 15 minute resolution cannot be achieved with a single modeling method. Recognizing this complexity, a decoupling of the problem is proposed, adopting scale-appropriate modeling techniques for different timescales, each tailored to capture the unique dynamics inherent to the timescale under consideration. **Chapter 7** and **Chapter 8** respectively discuss the proposed load models at the daily and annual level.

However, before tackling the issue of load modeling itself, **Chapter 5** first introduces the mathematical framework to define peak demands, given their increasing importance for modern applications focused on demand response and peak shaving. While the term "*peak demands*" is ubiquitously used by grid operators and utilities, no rigorous definition suitable for individual consumers could be found. As such, the load duration curve is proposed to determine a threshold unique to each consumer.

This definition is used in **Chapter 6** to develop a clustering algorithm to group consumers with similar properties together. A novel feature set is constructed based on the temporal properties of consumption and peak demands, which is subsequently used in an hierarchical clustering process to partition the dataset under consideration into 10 clusters. The clustering algorithm produces compact clusters with clear connections to real-life applications involving residential peak demands, such as demand response initiatives and the applicability of battery storage systems for peak shaving. Furthermore, these clusters demonstrate a clear connection to the predictability and variability of consumption and peak behavior, allowing for the quantification of the disproportionate presence of peak demands in certain time periods for each cluster. This constructed metadata based on the clusters will prove to be invaluable during the following steps.

At the daily level, **Chapter 7** explores the use of time-frequency analysis to construct daily load profiles. The stochastic peak demands are treated as high-frequency components superimposed on a more slowly varying, low-frequency base load. A wavelet-based multi-resolution analysis is used to decompose daily smart meter data in their low-frequency and high-frequency components. Subsequently, a stochastic load profile generator at the daily level is constructed. By rescaling and shifting the high-frequency component of one household and superimposing it on the low-frequent component of a different household, a realistic synthetic load profile can be obtained. This methodology is shown to allow for fine-grained control on the distributions of and relation between the daily peak demand and daily consumption.

A different technique is necessary to model annual profiles of daily consumptions, as they exhibit more correlated behaviors over longer timescales. **Chapter 8** illustrates that Generative Adversarial Networks (GANs) are well-suited for this task. The DoppelGANger architecture is used due to its ability to capture long-term time correlations and its capability to batch-generate short-term intraday variations at the weekly level. Based on a sensitivity analysis of inputs influencing the generation process, a set of parameters is identified that eliminates mode collapse in both the frequency and amplitude domains, facilitating the consistent generation of datasets of load profiles that closely resemble real annual load profiles.

Finally, **Chapter 9** presents the validation of the constructed synthetic dataset on a series of downstream applications of residential smart meter data. The daily load profiles of **Chapter 7** are inserted in the annual profiles of daily consumptions from **Chapter 8** to obtain annual profiles with high-frequency variability and realistic temporal short-term and long-term dynamics. These validations prove the excellent distributional similarity between real and synthetic data for the included benchmarks, at both the individual and aggregated level. However, qualitative metadata and a sufficiently large training dataset are both crucial to obtain a dataset with realistic and sufficiently diverse load profiles.

In conclusion, this PhD dissertation proposed a novel two-step methodology to generate synthetic data from privacy-sensitive smart meter data. The results confirm the viability of this technique for producing high-frequency data over the course of a full year, making it a feasible alternative to smart meter data for a wide range of practical applications. Because the observed shortcomings can be attributed to the proof-of-concept phase conducted on a limited dataset, the presented techniques demonstrate potential for scaling up and applications in non-academic contexts.

Part I

Constructing realistic synthetic load profiles based on privacy-sensitive residential smart meter data

1

Introduction

What is the use of a house if you haven't
got a tolerable planet to put it on?
Henry David Thoreau

At its core, this PhD thesis is about households; about individuals, appliances, and the day-to-day activities inside their own homes. For centuries, this information has remained veiled in privacy, shielded from governments, companies, utilities, as well as grid operators. However, the advent of smart meters and advanced metering infrastructure over this past decade has inadvertently breached this long-standing privacy barrier. Through the widespread recording of energy consumption habits at unprecedented levels of detail, these technological advancements offer unparalleled insights into both the daily routines as well as the habits of households.

Given the heightened awareness surrounding privacy and data security, regulators and legislators have responded by imposing restrictions on access to smart meter data. Nevertheless, the wealth of granular consumption data obtained from smart meters continues to hold immense value in expediting the advancement of new products and services pivotal to facilitating the clean energy transition at the level of individual consumers on the low-voltage grid. This data can not only be used by public research institutions but also by private enterprises, aiding them in refining their digital business models through innovative research and development.

This dissertation focuses on load modeling, a discipline that leverages privacy-sensitive smart meter data to produce synthetic datasets suitable for unrestricted sharing. Grid operators, who collect large volumes of this data but are limited in sharing it, are the primary target audience. Researchers in this field aim to unlock the potential of smart meter data while protecting individuals' privacy. This work contributes to advancing load modeling methodologies that balance data utility with privacy preservation, offering solutions to these challenges.

1.1 Problem statement

Traditional load modeling techniques typically rely on unsupervised machine learning to group consumers with comparable characteristics, followed by the averaging and smoothing of their smart meter data to produce synthetic load profiles. However, contemporary applications of smart meter data emphasize the need for flexibility and accuracy, particularly in capturing both the timing and amplitude of individual peak demands, rendering traditional smoothed curves inadequate. Granular synthetic data, characterized by realistic representations of peak demand behavior, are imperative for addressing these shortcomings and advancing two critical objectives, at the individual and the collective level, in the context of an evolving energy landscape.

First, at the individual level, the transition to a clean energy future necessitates the development of digital business models customized to the diverse needs of consumers. This approach transcends the conventional one-size-fits-all paradigm, ensuring personalized and profitable energy management strategies. To achieve this, accurate load models are crucial for promoting and automating demand response systems tailored to each consumer's unique circumstances. Data with realistic peak behavior serves as the cornerstone to enable private enterprises to accelerate the development, validation, and deployment of energy-efficient systems and services that address the evolving needs of consumers.

Second, at the aggregated level, the remarkable growth of photovoltaic installations, electric vehicles, and heat pumps poses significant challenges to distribution grids, necessitating accurate hosting capacity studies to ensure grid reliability and resilience. Granular smart meter data with accurate representations of peak demand behavior is essential for performing these hosting capacity studies. By leveraging realistic data, stakeholders can identify potential grid constraints, optimize grid operations, and suggest targeted infrastructure upgrades to ensure the reliable and efficient operation of distribution grids in the face of evolving energy trends.

1.2 Objectives and contributions

Building upon the problem statement presented in the preceding section, further explained in Chapter 2, the main objective can be defined as follows:

Primary research objective

➢ Develop and validate a novel methodology for synthetic load modeling capable of reproducing the peak demand behavior in residential energy consumption patterns.

Given the complex interplay of factors that collectively shape the smart meter output of residential consumers, we argue that accurate synthetic load models at this timescale and resolution cannot be achieved with a single modeling method. Recognizing this complexity, we advocate for the decoupling of the problem and the adoption of scale-appropriate modeling techniques for different timescales, each tailored to capture the unique dynamics inherent to the considered timescale.

At the daily level, synthetic data will be generated through a decomposition-recombination procedure rooted in time-frequency analysis and supported by wavelets. At the annual level, recent advances in generative AI are leveraged to generate annual profiles that accurately capture seasonal, weekly, and interday correlations.

To systematically address these methodological challenges and refine the primary research objective, we formulate several additional secondary research questions. These secondary objectives serve as critical milestones, collectively contributing to achieving the primary research question.

Secondary research objectives

➢ Identify the use cases for granular smart meter data, both at the individual and aggregated level. These downstream applications will serve as benchmarks of the synthetic data throughout this dissertation.

➢ Define a rigorous mathematical framework to classify which values can be considered '*peak demands*'.

➢ Construct a feature set capable of incorporating the peak demand behavior, and to use this feature set in a clustering algorithm to investigate and quantify the difference in peak demand behavior for different consumer categories.

➢ Use time-frequency analysis to construct daily load profiles with sufficiently stochastic peak demands.

➢ Train generative adversarial networks to generate annual profiles with realistic multiscale time correlations.

1.3 Thesis outline

With both the research scope as well as the primary and secondary research objectives defined in the previous sections, the methodology and thesis outline form the subject of this section. Figure 1.1 gives a graphical overview of this outline.

Chapter 2 starts by providing the necessary background information and helps explain the context and relevance of this PhD dissertation. Beginning with an examination of the residential sector within the broader energy landscape, we discuss the challenges this sector faces in decarbonizing its emissions. Specifically, we focus on the growing significance of photovoltaic installations, electric vehicles, and heat pumps in shaping the transition towards cleaner energy sources. Subsequently, we explore the role of smart meter data in accelerating this transition, both at the individual and aggregated levels, with particular emphasis on their impact on the distribution grid. This chapter concludes with a summary of the legislative framework around smart meter data, and the personal data that can be extracted from them.

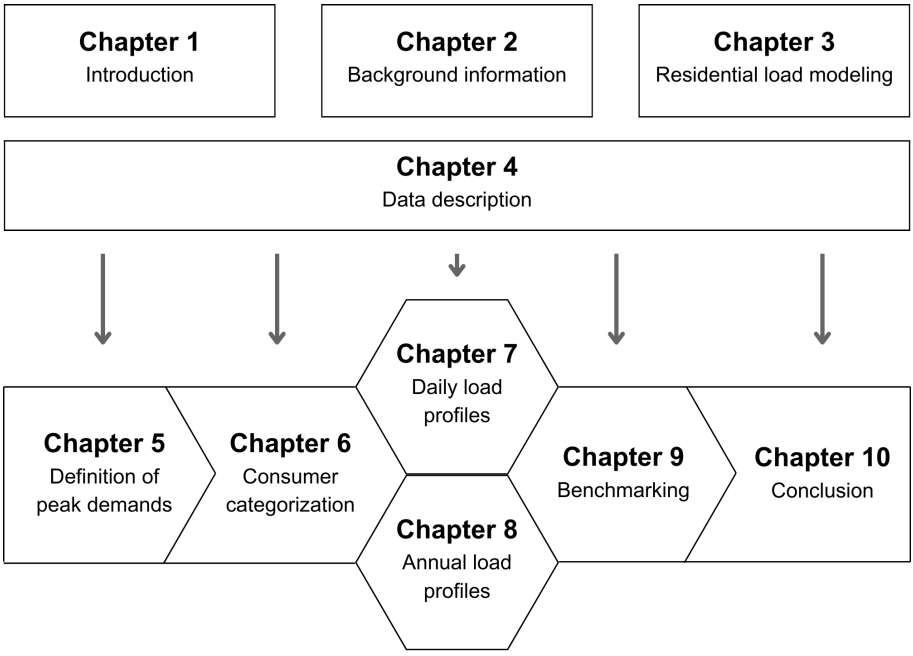

Figure 1.1: Outline of the dissertation, where arrows and connected tiles represent links from one chapter to another.

Chapter 3 subsequently analyzes the state-of-the-art literature on residential load modeling, ranging from bottom-up models to top-down models that use smart meter data to generate synthetic data. This literature review helped refine the secondary objectives of this thesis, while simultaneously resulting in the multifaceted approach towards the modeling of different timescales as discussed in Section 1.2.

The smart meter data used throughout this dissertation are described in Chapter 4. Furthermore, the production data of photovoltaic installations, the electric vehicle charging profiles, as well as the heat pump load profiles are discussed in the remainder of that chapter. Given the data-drive nature of this dissertation, this data will be used throughout all following chapters.

Chapter 5 proceeds by introducing the mathematical framework necessary to unambiguously define peak demands, which will be used for the remainder of this work. Drawing from the concept of load duration curves, frequently employed by electric utility engineers, we establish consumer-specific thresholds to define peak loads.

This framework serves as the cornerstone for Chapter 6, wherein we construct a feature set characterizing both consumption patterns and peak demands across time. This feature set serves as the basis for a clustering algorithm aimed at categorizing consumers and quantifying the stochastic nature of peak demands across a diverse spectrum of residential households.

In Chapters 7 and 8, the synthetic data generation process for daily and annual profiles is established. Chapter 7 starts with a brief exploration of time-frequency analysis and discrete wavelet transformation, laying the groundwork for the subsequent decomposition-recombination process. Smart meter data undergoes decomposition into low-frequency and high-frequency components, where the latter drives stochastic peak demands. By rescaling, shifting and combining these components across a diverse consumer dataset, we construct a robust stochastic daily load profile generator.

Finally, Chapter 8 introduces generative adversarial networks (GANs). As state-of-the-art in synthetic data modeling, generative AI is used to produce annual profiles of daily consumptions. A set of application-specific microbenchmarks is introduced to evaluate the fidelity of the generated data against the input dataset. Additionally, various combinations of neural networks are trained and benchmarked to consistently yield annual profiles exhibiting realistic seasonal, weekly, and interday temporal correlations.

While benchmarks comparing the synthetic data to the original dataset are provided in the preceding chapters, the primary validation of our methodology is presented in Chapter 9. Herein, the performance of the synthetic data is benchmarked against the original dataset across various downstream applications. Building upon the insights obtained in Chapter 2, a comprehensive array of applications at both individual and aggregate levels are analyzed and discussed.

To conclude this dissertation, Chapter 10 provides a summary of the methodology and main conclusions, as well as an outlook on future research directions. The limitations of the proposed methodology are discussed, as well as perspectives for improvements and continuation of this work.

2

Background information

The energy system is collecting more data than ever, but too much of it remains idle, or stuck in siloed storage, with significant untapped potential.

Pauline Henriot

This chapter aims to provide the necessary background to contextualize the research presented in this dissertation. First, Section 2.1 situates the residential sector within the broader energy landscape, emphasizing its significant contribution to total emissions. Section 2.2 subsequently explores the anticipated impact of the energy transition at the level of the individual households, as well as on the low-voltage distribution network.

For the second part of this chapter, our focus shifts towards smart meter data. Section 2.3 highlights the economic opportunities presented by data-driven digital business models, paving the way for subsequent discussions on the diverse applications of smart meter data.

However, the use of smart meter data is constrained by privacy considerations and regulatory frameworks. Section 2.4 covers the privacy-related challenges associated with smart meter data handling. In Section 2.4.1, the legislative framework for EU countries is described, highlighting the classification of smart meter data as personal data under the General Data Protection Regulation (GDPR). Finally, Section 2.4.2 provides insights into the types of information that have previously been extracted from smart meter data, clarifying why such data needs to be classified as personal.

2.1 Residential consumers & their emissions

This thesis primarily centers on the residential sector situated on the low-voltage distribution grid. It is therefore important to contextualize the residential sector within the broader energy landscape, a task undertaken in this section.

In the European Union (EU), households accounted for 27% of the final energy consumption in 2021, according to the most recent data available from Eurostat [1]. Residential energy usage is predominantly allocated to space heating (SH) or cooling, domestic hot water (DHW) production, cooking, lighting, and appliance usage. As depicted in Figure 2.1, natural gas and electricity currently dominate the EU's final energy consumption in the residential sector, while renewables and biofuels contribute to covering 21.2% of the consumption.

Due to both direct emissions from fossil fuel consumption and indirect emissions from electricity generation, the residential sector in the EU contributes to approximately 20% of the total emissions, with fossil-based heating being a significant driver of energy demand in many member states.

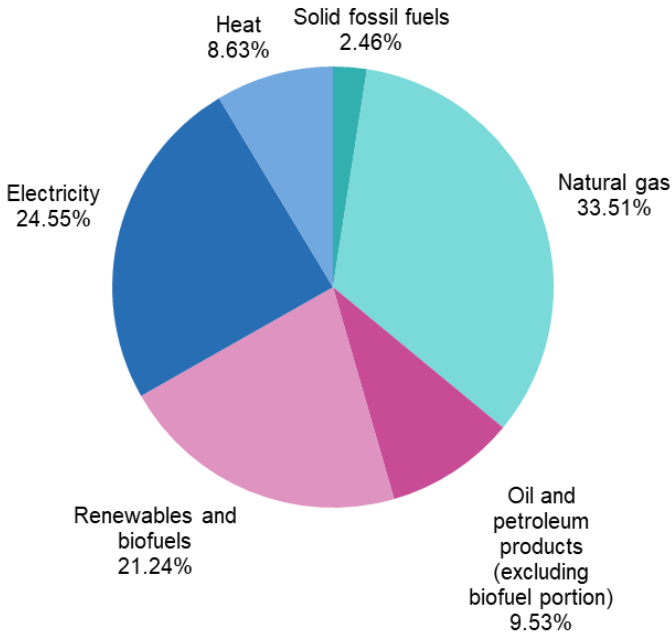

Figure 2.1: Final energy consumption in the residential sector by fuel type in the EU [1].

Examining the situation within our own home country, Belgium, Figure 2.2 illustrates the sector-wise distribution of national greenhouse gas (GHG) emissions, as reported in the most recent National Inventory Report [2]. The residential sector is directly responsible for nearly 15% of emissions, while the transport sector is responsible for more than 20%.

These aggregated numbers obfuscate the size of challenges ahead for residential consumers in Belgium. Over 80% of the residential buildings in Belgium was built before 1981. In 2019, oil and petroleum products still constituted a 29% share of all fuels in the final energy consumption in the residential sector [3]. The annual EU average per capita GHG emissions from residential heating was 696.4 kg. In Belgium, 1210.5 kg was emitted per capita. This highlights the urgent need for a drastic increase in renovation rate of the building stock together with a shift towards more environmentally friendly technologies for residential heating.

Challenges of a similar size lie ahead in the transportation sector. Since 1990, the number of passenger cars has increased by 53%. The evolution of the number of passengers cars is visualized in Figure 2.3 on the following page. With an increase of GHG emissions by 26% between 1990 and 2019, road transport is one of the few sectors that has actually increased its emissions since 1990.

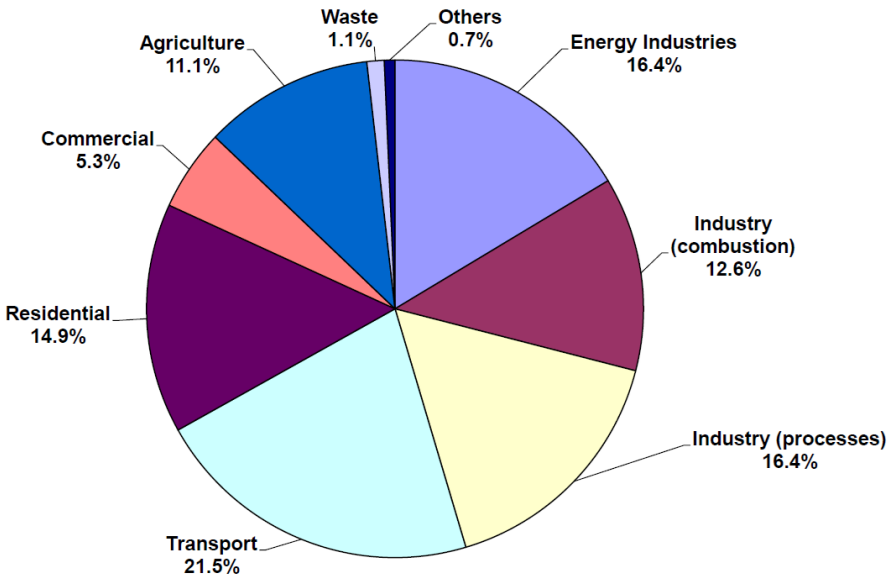

Figure 2.2: Contribution of the main sectors to Belgian greenhouse gas emissions [2].

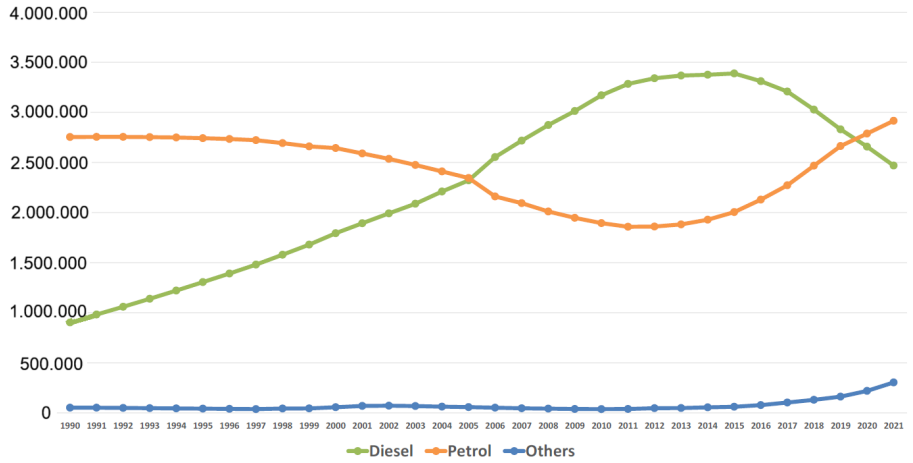

Figure 2.3: Fleet numbers of passenger cars in Belgium [2].

In recent years, a notable shift has begun to manifest, as highlighted by a gradual increase in the '*Others*' category in Figure 2.3, comprising non-conventional engines such as LPG, CNG, petrol conventional hybrid, petrol plug-in-hybrid, as well as battery electric vehicles. This rise signals a growing adoption of electric vehicles (EVs) in Belgium, propelled by a diverse range of fiscal incentives and policy measures implemented by governmental bodies.

While this marks a positive development, the challenges that lie ahead demand urgent attention, necessitating an acceleration of current trends to drive further progress. We look ahead towards the energy transition and its impact on residential consumers and the low-voltage grid in the next section.

2.2 Residential consumers & the energy transition

The term '*energy transition*' refers to the ongoing shift away from traditional, fossil fuel-based energy systems to cleaner, renewable energy sources. This transition involves a transformation of the energy sector, encompassing changes in energy production, transmission, distribution, as well as consumption.

The energy transition is driven by the urgent need to decarbonize the global energy system in response to the looming threat of climate change. Indeed, the consequences of climate change are already being felt around the globe, affecting biological systems and human populations. The Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) is crystal clear in its findings [4], as listed on the following page.

- *Human activities, principally through emissions of greenhouse gases, have unequivocally caused global warming, with global surface temperature reaching 1.1°C above 1850-1900 in 2011-2020.*
- *Continued greenhouse gas emissions will lead to increasing global warming, with the best estimate of reaching 1.5°C in the near term in considered scenarios and modelled pathways. Every increment of global warming will intensify multiple and concurrent hazards.*
- *Continued emissions will further affect all major climate system components, and many changes will be irreversible on centennial to millennial time scales and become larger with increasing global warming. Without urgent, effective, and equitable mitigation and adaptation actions, climate change increasingly threatens ecosystems, biodiversity, and the livelihoods, health and well-being of current and future generations.*
- *Rapid and far-reaching transitions across all sectors and systems are necessary to achieve deep and sustained emissions reductions and secure a liveable and sustainable future for all. These system transitions involve a significant upscaling of a wide portfolio of mitigation and adaptation options. Feasible, effective, and low-cost options for mitigation and adaptation are already available, with differences across systems and regions.*

Given the increasingly urgent need for decarbonization across all segments of society, the residential sector faces a critical transformation within a relatively short time frame of less than three decades. This transformation necessitates swift and substantial changes, including the widespread deployment of rooftop photovoltaic (PV) systems, a rapid transition from gas and oil-based heating systems to more energy-efficient electric heat pumps in well-insulated homes, and an accelerated adoption of EVs coupled with a shift towards alternative modes of transportation.

As responsibility for matters pertaining to rooftop PV installations, household renovations, as well EV charging infrastructure has been decentralized to the regional governments, we narrow our focus to Flanders, the Dutch-speaking region of Belgium. By focusing on Flanders, we aim to provide a scope of the challenges ahead in these key areas for residential consumers.

Consider, for instance, the growth in installed rooftop PV capacity. As of the beginning of 2024, the installed rooftop PV capacity stood at 6,071 MW, marking a significant increase of over 50% compared to the capacity installed in 2020 [5]. Despite the Flemish government's ambition for an annual growth of 450 MW [6], Flemish grid operators are already preparing for a projected installed capacity of 10 GWp by 2030, reflecting the sustained higher growth trajectory [7]. With Flanders boasting a potential rooftop PV capacity of 65 GWp, this accelerated growth appears achievable [8].

EVs are expected to follow a similarly significant growth trajectory. In 2023, Flanders recorded nearly 3.7 million passenger cars, of which 104,000 were fully electric and 366,000 were plug-in hybrid vehicles [9]. By 2030, the Flemish Government anticipates 1.2 million EVs on the road [6], while Flemish grid operators, in collaboration with various sector federations, to proactively prepare their distribution grids for the anticipated surge, aiming to accommodate up to 1.5 million EVs by 2030 [7].

Additionally, beyond passenger cars, both freight transport and public transportation sectors are swiftly electrifying their fleets, accelerating the transition away from fossil-based engines and further reducing emissions in the transport sector.

Long-term policy goals regarding heat pumps are ambitious, yet intermediate targets are comparatively modest. The Flemish Government aims for a climate-neutral residential building stock by 2050 [6]. To achieve this target, the renovation rate of the residential building stock would need to reach 3% annually, implying that 3% of all residential buildings undergo renovation each year. However, the current renovation rate for residential buildings falls below 1%. The Flemish Government has set a goal of installing 42,750 air-to-water and soil-to-water heat pumps each year by 2030 [6]. To contextualize this figure, it is important to note that the Flemish region comprises over 3.3 million residential dwellings [10]. Consequently, grid operators anticipate that only 10-15% of residential dwellings will be heated by heat pumps by 2030 [7], with expectations for the renovation rate to accelerate after 2030.

While all these changes and transitions are driven by the societal need to urgently decarbonize the energy production and consumption, the low-voltage (LV) distribution grids need to be able to accommodate these changes. The inclusion of PV installations at the level of the end user is changing consumers to prosumers, a portmanteau of '*producer*' and '*consumer*', injecting in the grid when production is too high. Similarly, the widespread penetration of EVs and HPs with simultaneous peak consumption can lead to grid congestion. Consequently, the energy transition is fundamentally transforming the operation of the distribution network.

Hence, as the energy transition gains momentum, grid operators are reinforcing their infrastructure to prevent them from becoming bottlenecks. Concurrently, stakeholders are implementing economic incentives to encourage consumers to spread their consumption. These measures include the introduction of peak-based tariffs and the establishment of flexibility markets.

In Flanders, grid operators have assessed the distribution network's status in 2024. Considering the anticipated growth of PV panels and the increasing electrification of transportation and heating, driven by the adoption of EVs and HPs, they have simulated the projected state of the distribution grid infrastructure by 2035. This simulation is depicted in Figure 2.4.

To mitigate the risk of grid congestion, the Flemish grid operators are currently planning to invest 12 billion euros over the course of the next decade [7]. Given that any investments made by the grid operators are financed through grid tariffs paid by all consumers, it is imperative to optimize the efficiency of grid reinforcement efforts in the interest of society at large.

Smart meters and digitalization represent possible avenues towards improving the efficiency of grid infrastructure utilization, through demand side response and flexibility initiatives. The potential of these digital business models and different use cases for smart meter data will be discussed in the next section.

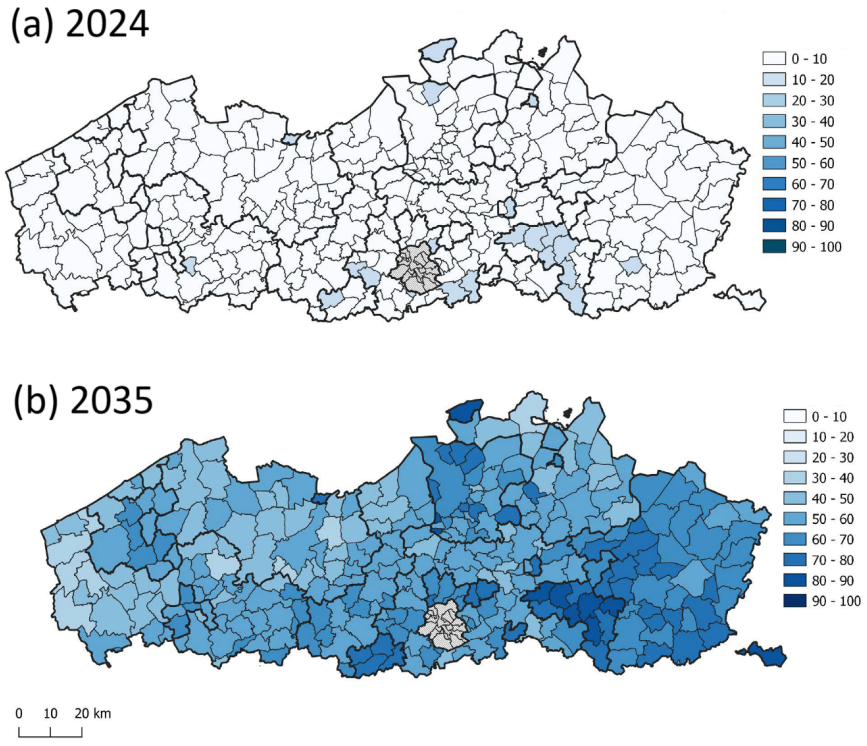

Figure 2.4: Share of current distribution grid potentially in congestion (%/municipality) in (a) 2024 and (b) 2035 [7].

2.3 Uses of smart meter data

The energy sector has undergone rapid digitalization over the past decade, precipitating a shift towards innovative digital business models as alternatives to legacy business models. Unlike their hardware-centric predecessors, these new business models are predominantly software-driven, marking a departure from the dominance of large industrial companies and utilities.

This disruptive market effect is highlighted in Figure 2.5, which displays the trajectory of venture capital inflows into digital energy start-ups since 2015, as analyzed by the International Energy Agency (IEA) [11].

Several noteworthy insights emerge from the data. First, despite the COVID-19 pandemic disrupting global supply chains in 2020, early-stage investments in clean energy start-ups actually increased that year. Secondly, the scale of venture capital financing has experienced a remarkable growth, tripling from its 2016 levels by 2020. Thirdly, the growing importance of "*charging as a service*" underscores the challenges and opportunities accompanying the anticipated exponential growth of EVs.

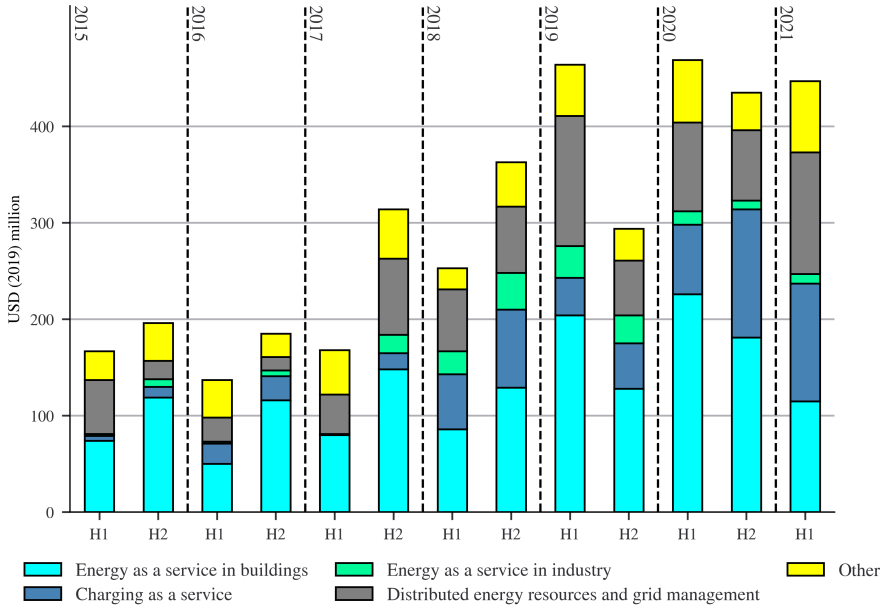

Figure 2.5: Global early-stage venture capital investments in digital energy-efficiency and demand-side flexibility start-ups, by type of new business model, 2015-2021 [11].

At the heart of this digital transformation lie smart meters and the granular data they provide, offering unprecedented insights into consumer habits to empower companies. This metering infrastructure not only enables organizations to precisely quantify the value proposition of their solutions to the benefit of each individual customer, but also expedites the development of new products and services. Furthermore, the establishment of bi-directional communication between consumers, energy providers, and grid operators facilitates the automation of demand response services, unburdening individual consumers.

From the perspective of the individual consumer, smart meters can transmit detailed data related to electricity consumption at intervals typically ranging from 15 minutes to an hour. This granular data collection enables more accurate billing, improved demand forecasting, and effective load management. Additionally, smart meters can transmit information on voltage levels, which allows utilities to detect potential issues such as voltage sags or swells and take timely corrective actions. Furthermore, this improves the utilities' ability to respond to outages to efficiently respond to outages, minimizing downtime and improving overall service reliability.

Beyond data transmission, smart meters can receive information from various market participants. Consumers can be updated on dynamic energy prices, time-of-use rates, and peak pricing schedules, encouraging them to shift their energy usage to off-peak periods. This not only optimizes energy consumption but also reduces costs. Moreover, smart meters can send notifications or alerts regarding demand response events or planned outages, providing consumers with timely and actionable information.

Leveraging digital solutions to their fullest potential thus holds the promise of both streamlining and accelerating the energy transition by facilitating energy efficiency improvements and demand-side flexibility. The integration of smart meters with advanced analytics platforms, machine learning algorithms, and Internet of Things (IoT) devices enables real-time monitoring, analysis, and control of energy systems. This enables grid operators to respond swiftly to changes in energy demand and generation, optimize the use of renewable energy resources, and enhance the resilience and reliability of the grid.

From an investment point of view, the growing interest from venture investors and large corporations in digital energy innovators signals how high the expectations are for growth in this sector. The growth of venture capital directed towards digital energy start-ups therefore not only highlights the democratization occurring within business models, but also underscores the expanding role of digitalization in shaping the modern energy landscape.

Despite ongoing efforts, the IEA has identified several hurdles that must be addressed to scale up current efforts related to digital business models and achieve net-zero emissions by 2050. One such hurdle is the availability of smart meter data for developing and validating customized services within strict regulatory frameworks. Consequently, making datasets of synthetic data publicly available, with properties similar to privacy-sensitive smart meter data, can encourage innovative digital business models.

The synthetic data generated in this dissertation will be benchmarked compared to the original privacy-sensitive dataset under a non-disclosure agreement (NDA). Furthermore, the performance of the synthetic data will be benchmarked based on several downstream applications to evaluate its fidelity. Consequently, we provide an overview of several of these downstream applications where smart meter data is used. Section 2.3.1 examines use cases at the individual consumer level, while Section 2.3.2 considers the aggregated level. For each application, special attention is given to discussing the necessary minimum length and resolution of the required input data.

2.3.1 Uses at the individual level

At the individual household level, smart meter data frequently serves to assess the benefits and economic feasibility of innovative technologies such as PV installations, potentially coupled with battery energy storage systems (BESS). Analyses conducted on larger, representative datasets enable the development of sizing guidelines, facilitating the dissemination of best practices to engineering firms and stakeholders responsible for the sales and installation of these systems. By leveraging smart meter data in this manner, stakeholders can make informed decisions regarding the implementation of sustainable energy solutions, promoting widespread adoption and maximizing the benefits for investing consumers. These applications can be grouped under the broader application category of asset sizing.

For example, [12] examined the benefit of a PV installation for individual consumers based on their smart meter data. The impact on the self-consumption ratio (SCR) and self-sufficiency ratio (SSR) was considered for different consumers under a wide range of PV sizes with various tilt and azimuth angles. Similarly, [13] proposed an optimization of PV sizing and orientation for 13 customer classes, obtained from a large-scale smart meter dataset of hourly consumption data. Sizing schemes for PV-BESS systems based on smart meter data of residential consumers were considered in [14], [15], as well as [16].

As expected, both the temporal resolution and the considered time scale significantly impact the calculated self-consumption and, consequently, the economic viability of these systems. Two recent review papers discussed the effects of the resolution of both the PV production profile and the load profile [17, 18]. They concluded that having a full year of data to capture seasonality is essential. Specifically, (i) accurate results for PV system sizing can be obtained using 15-minute data, although the resolution of the residential load was found to be of greater importance than the PV resolution, (ii) temporal resolutions between 5 minutes and 60 minutes yield reliable results for sizing PV-BESS systems, and (iii) a finer temporal resolution of at least 5 minutes is necessary for the optimal sizing of the battery inverter [19–21]. Coarser resolutions were found to overestimate PV self-consumption, particularly under scattered cloud conditions, thereby misrepresenting the return on investment.

In addition to asset sizing, smart meter data can be used to make informed decisions in load scheduling and load management through data-driven forecasting at the level of the individual household. In the context of residential consumers, historic smart meter data is often used to forecast the load profile and/or PV production. Predicted values subsequently serve as input for the scheduling of BESS or individual appliances to minimize the electricity cost. According to a review study, the most common optimization horizon for these Home Energy Management Systems (HEMS) is 24 hours ahead, using short-term load forecasts (STLF) at a resolution of at most 1 hour [22]. There is an obvious trade-off between the optimal scheduling and an increased computational burden when moving towards finer resolutions.

Illustratively, [23] considered residential consumers with a PV-BESS system. Forecasts of generation and demand were applied, and optimal schedule of the battery was determined. Their results showed an improvement of lifetime value by an average of 160% relative to basic set-point control representative of many systems in operation at the time. Similar challenges were considered in [24–26].

Finally, smart meter data is often used as input to test various Reinforcement Learning (RL) algorithms for the optimal scheduling of residential PV-BESS installations, which allows researchers to quantify the added value of more complex scheduling algorithms compared to more naive benchmarks. This was done in [27], where a relative electricity bill reduction of 14% was achieved when comparing the proposed algorithm with the benchmark approach. An analogous problem was tackled in [28], where the proposed RL algorithm reduced the accumulated electricity cost by 11.38% compared to the benchmark case.

Two other applications at the individual household level can be distinguished: load profiling and consumer characterization, as well as appliance identification. Given sufficiently high resolution of smart meter data, the presence of appliances and renewable energy sources can be detected with high accuracy. However, this capability raises privacy concerns for households, which is one reason access to smart meter data is restricted. Section 2.4.2 provides an in-depth overview of the appliances that can be detected based on smart meter data with resolutions ranging from seconds to hours.

Load profiling and consumer characterization involve grouping individual consumers with similar properties and constructing representative load profiles [29–31]. Given the relevance of this application for this dissertation, Chapter 3 is fully dedicated to the state-of-the-art of residential load profiling. Therefore, the discussion in this section is restricted to the resolution and time scale under consideration for load profiling. Two review papers conclude that smart meter data at the daily or weekly level are typically used for profiling and characterization purposes, with a resolution of at least 15 minutes [18, 32]. Indeed, [33] investigated the performance of three clustering algorithms for smart meter data with resolutions varying from 1 minute to 2 hours, showing that a resolution of at least 30 minutes is sufficiently reliable for load profiling purposes.

2.3.2 Uses at the aggregated level

Analogous to forecasting at the individual level, forecasting based on metering data is a common application at the aggregated level. However, both the end goals of the forecasting and the considered timescales differ. Very short-term load forecasting (VSTLF) focuses on subhourly time horizons, ranging from a few minutes to an hour, with very fine temporal resolutions [34, 35]. VSTLF is necessary for various balancing services, aiding in the real-time scheduling of generation, load frequency control, and resource dispatch [36, 37]. Accurate VSTLF ensures stability in power systems by quickly responding to fluctuations in demand and supply, thereby improving grid reliability and efficiency.

At longer timescales, medium or mid-term load forecasting (MTLF) covers forecasting periods from a few days to several months [34, 38]. This type of forecasting typically uses data at hourly or coarser resolutions, and provides necessary input for the planning and scheduling of preventive maintenance of units [39]. Additionally, MTLF assists in identifying trends and patterns in energy consumption, which can be valuable for strategic planning and decision-making within utility companies.

The aforementioned applications are based on data with very fine or very coarse time resolutions. However, two uses cases for smart meter data at fine resolutions at the aggregated level are highlighted for this dissertation: hosting capacity studies and the case for energy communities and Virtual Power Plants (VPPs).

Hosting Capacity studies

The energy transition fundamentally changes how grid operators need to manage and size the assets of their distribution networks, challenging the traditional perspective on the LV grid. Historically, (LV) distribution systems were designed to distribute power from centralized power plants to end-users connected to the LV grid, often dimensioned to withstand a worst-case concurrent peak demand. However, the expected exponential growth of EVs, HPs, as well as PV installations was introduced in Section 2.2.

Grid operators are aware that the large-scale integration of PV installations, EVs and HPs can lead to LV grids performing unsatisfactory. Research related to the integration of these new appliances are grouped under the term “Hosting Capacity" (HC) studies. A frequently used definition of the HC is the amount of new production or consumption that can be connected to the grid, while not requiring infrastructure modifications, without impairing the reliability or voltage quality for other consumers [40, 41].

A myriad of issues can arise at higher penetration of PV installations, HPs, EVs, or any combination of these appliances. Overvoltage due to PV installations [42, 43], undervoltage due to EV charging [44, 45], thermal overloading due to HP integration and/or EV charging [46, 47], as well as voltage unbalance between phases [48, 49] have all been predicted through various HC studies on different types of LV grids.

Multiple approaches exist to calculate the HC, depending on the available data and the desired accuracy of the analysis. A common denominator in the different approaches is the use of power flow calculations to determine the voltages and currents in the considered network [50, 51].

One of the prerequisites of many HC studies is the data availability of the pre-existing demands on the network [50]. Hosting capacity studies that use time series data were found to be accurate and realistic, although the main drawback is that few scenarios can be tested. However, time series analysis allows to take the intermittent nature of distributed generation into account, while also including all correlations as they occur in reality, such as solar power production and electric heating or cooling [51]. Analogously, HC studies that use a probabilistic approach for the maximum load or generation need to define distributions based on available data, not only for the generation, but also for the pre-existing demand.

In [47] the introduction of EVs and HPs were considered in UK-based networks, and the HC was found to be strongly dependent on the initial demand, highlighting the need for accurate modeling of the a priori loads. Consequently, the availability of residential smart meter data is of paramount importance for accurate HC studies on the LV grid. However, its availability is severely hindered due to privacy concerns, which is discussed in-depth in Section 2.4.

Energy Communities, DC backbones & Virtual Power Plants

The advent of smart meters with bi-directional communication capabilities has opened up new possibilities for aggregating individual end-users and renewable energy systems (RES), thereby facilitating the optimization of consumption or production at a higher aggregation level. These collaborative arrangements, known as energy communities or VPPs, depending on the context, enable more efficient utilization of energy resources and grid infrastructure.

Concurrently, research into alternative grid architectures is gaining traction, with investigations into the advantages of low-voltage DC (LVDC) backbones compared to traditional LVAC grids, particularly in contexts such as the integration of PV systems and EV chargers.

It is evident that research in these areas centers around the concept of aggregation and aims to quantify potential benefits for various stakeholders, including consumers, producers, grid operators, and society at large. By exploring the advantages and challenges associated with these emerging technologies and grid configurations, researchers aim to inform decision-making processes and contribute to the development of more sustainable and resilient energy systems.

Research on the impact of aggregation using smart meter data has shown improved accuracy for forecasting [52, 53], increased PV self-consumption when only considering PV installations [54–56], as well as higher PV self-consumption and a lifetime improvement when a community BESS is integrated in the grid instead of individual BESS together with PV installations [57, 58]. However, saturation effects for these advantages tend to start to occur at aggregation levels higher than 20 individual residential consumers [55, 59].

Similar topics using smart meter data of individual consumers have been investigated for alternative grid configurations such as LVDC systems. In [54], the conversion and cable losses were compared for residential consumers with a PV installation connected to an LVAC and LVDC configuration. Similarly, [60] and [61] compared and quantified the benefits and limitations of a LVDC microgrid versus an LVAC microgrid for EV charging.

Figure 2.6 provides an overview of the findings related to the input data requirements of various data-driven applications that rely on smart meter data. The x-axis represents the required length of the input data of each application, plotted on a logarithmic scale to accommodate the wide range of timescales involved, from seconds to years. The y-axis displays the resolution of the data, indicating the necessary granularity of the data for the considered application.

In the context of this dissertation, particular emphasis will be placed on applications related to asset sizing, specifically for PV systems, PV-BESS systems, as well as HC studies for the sizing of grid assets. As previously discussed and illustrated in Figure 2.6, a time scale of at least a single year is necessary to capture the seasonal nature of the load, while insight in the fine resolution is necessary for the short-term dynamics.

Given the absence of well-established synthetic data models that can generate accurate data over longer timescales with a 15-minute resolution, as will be elaborated in Chapter 3, these applications provide a robust framework for validating the synthetic data generation methodology developed in this work.

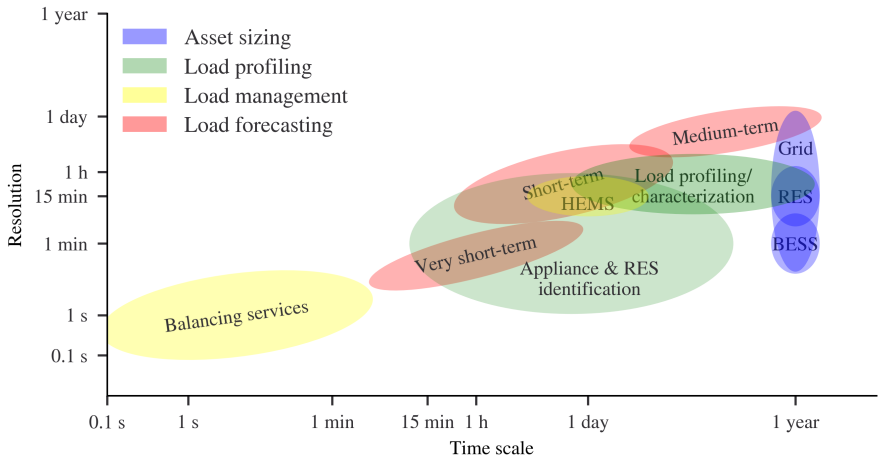

Figure 2.6: Positioning of data-driven applications using smart meter data based on input data requirements, showing the relationship between time scale (x-axis) and time resolution (y-axis) of the input data.

2.4 Smart meter data & privacy

2.4.1 Legislative framework

In the European Union, the right to privacy is attributed the highest possible protection. Both Article 8 of the European Convention on Human Rights by the Council of Europe [62], as well as Article 7 of the Charter of Fundamental Rights of the European Union [63] explicitly mention the right to respect for the private and family life, as well as their home, of every individual.

Given the sensitive nature of smart meter data, legislators have classified them as personal data in Article 23 of the EU Directive concerning common rules for the internal market for electricity [64]:

> **Article 23: Data management**
>
> 1. *When laying down the rules regarding the management and exchange of data, Member States or, where a Member State has so provided, the designated competent authorities shall specify the rules on the access to data of the final customer by eligible parties in accordance with this Article and the applicable Union legal framework. For the purpose of this Directive, data shall be understood to include metering and consumption data as well as data required for customer switching, demand response and other services.*
> 2. *Member States shall organise the management of data in order to ensure efficient and secure data access and exchange, as well as data protection and data security. Independently of the data management model applied in each Member State, the parties responsible for data management shall provide access to the data of the final customer to any eligible party, in accordance with paragraph 1. Eligible parties shall have the requested data at their disposal in a non-discriminatory manner and simultaneously. Access to data shall be easy and the relevant procedures for obtaining access to data shall be made publicly available.*
> 3. *The rules on access to data and data storage for the purpose of this Directive shall comply with the relevant Union law. The processing of personal data within the framework of this Directive shall be carried out in accordance with Regulation (EU) 2016/679.*

Therefore, the data management of smart meter data needs to adhere to the EU's General Data Protection Regulation (GDPR), which applies to any use of personal data or any aggregated data through which it is possible to identify a living person [65].

GDPR imposes strict rules on every step of the data management strategy of smart meter data, limiting its possible uses and sharing of data. Companies working on smart meter data face several legislative hurdles, including but not limited to:

- **Purpose limitation**: GDPR requires that personal data must be collected for specified, explicit, and legitimate purposes. Companies cannot collect or acquire data through a Data Transfer Agreement (DTA) for one purpose and then use it for another without obtaining additional consent, which restricts their ability to repurpose data freely.
- **Data minimization**: Companies are required to collect only the data that is necessary for the intended purpose. This means they cannot collect excessive amounts of personal data beyond what is needed for their services, nor can they require additional data in a DTA. Consequently, DTAs of smart meter data are often tailored to specific purposes to exclude personal data that is deemed non-essential to the purpose of the request.
- **Data security obligations**: Data handlers are required to implement appropriate technical and organizational measures to ensure the security of personal data they process. Data protection methods can include encryption and secure storage systems.
- **Accountability and Compliance**: GDPR requires companies to demonstrate compliance with the regulation by maintaining detailed records of their data processing activities, often through a data management plan of individuals working with the data. Furthermore, companies may be required to appoint a Data Protection Officer responsible for overseeing GDPR compliance within the organization.

It should be evident that GDPR imposes compliance requirements and limitations on companies, which may impact their operations and growth prospects, particularly those engaged in data-driven digital applications. Compliance with GDPR is often complex and may require companies to invest time and resources in understanding and implementing its requirements.

Given the strict requirements on data handling and processing for smart meter data, the following section gives an overview of different types of personal data which have been extracted from smart meter data in academic literature. This summary aims to underscore the importance of data protection measures in safeguarding individual privacy and upholding regulatory compliance within the context of smart metering technologies, given the detailed and often surprising information that can be extracted.

2.4.2 Personal data extracted from smart meter data

Given the different types of personal data that can be extracted from high resolution smart meter data, this section is organized as follows. First, we examine how appliance ownership and usage can be determined for individual consumers. Second, we discuss how information about household occupancy and active hours can be detected and the implications of this automatic detection. To conclude this overview, several works are presented that combine survey information with smart meter data to determine correlations between socioeconomic properties of the consumers and their consumption habits.

Appliance usage

The focus of this paragraph lies on the detection of appliances relevant for the energy transition, i.e., heat pumps, electric vehicles, as well as air conditioning. While smaller household appliances such as TVs, dryers, dish washers, ovens, freezers, laptops and gaming consoles, have consistently been identified based on the characteristics of the aggregated household smart meter data, they are of secondary importance in the context of this dissertation [66–70].

- **Heat pumps**
  In [71], daily consumption information was used to determine the presence of heat pumps, both for space heating and space cooling depending on the season.

  Moving to higher resolution data, the authors of [72] used smart meter data with a resolution of 15 minutes to identify households and automatically differentiate between those with fixed speed and variable speed heat pumps. Analogously, [73] used data with similar resolution to identify both the households with a heat pump, as well as which heat pump reservoir is present.

  Building upon these works, [74] used 15 minute smart meter data to identify households which exhibit atypical cycling behavior of their heat pump. This atypicality was found to be an indicator with respect to energy efficiency and appropriate sizing of their heat pump.

- **Electric vehicles**
  In [75], households with an EV were identified based on their hourly smart meter data, while [76] proposed a detection scheme based on 30 minute data. Analogously, the additional load caused by EV charging at home were determined based on smart meter data with a resolution of 1 minute in [77].

Going one step further, the authors of [78] developed a methodology to determine the model of the EV being charged at the level of the individual household, using estimated values of the battery capacity, charging rates and energy charged per session.

- **Air conditioning**
  In [79], a machine learning-based algorithm was proposed to identify households with air conditioning. Similarly, [80] investigated the accuracy of the detection technique for different data resolutions for Texan households. While the accuracy on 5-minute data was comparable to 1-min data, according to the authors, 15-min data did not yield accurate results. Additionally, the authors of [81] used smart meter data with a resolution of 1 minute to detect when air conditioning switches on and off through the identification of the operation cycle.

Household occupancy

A common application of smart meter data is occupancy detection. While this offers advantages for home automation as well as more efficient energy management of HVAC systems based on occupancy, it comes with several privacy-based dangers. For example, insurance companies could use this data to profile individuals based on their occupancy patterns, charging higher premiums for individuals with perceived higher risk from their occupancy patterns. Similarly, should this data become available to criminals, they could use it to identify periods of low or no occupancy in homes or businesses, making them vulnerable to burglaries. Similarly, stalkers could get access to the routines and habits of their targets.

In [82], the concept of Non-Intrusive Occupancy Monitoring (NIOM) was introduced, where household occupancy is inferred based on electricity smart meter data. In [83], a combination of 1-minute resolution electricity and water consumption smart meters was used to detect occupancy with an accuracy of over 80%.

The authors of [84] investigated whether half-hourly meter data are sufficient to predict the home-occupancy status of households, not only in the present but also in the future. Their study revealed a high predictive power to also establish the future occupancy status of households. Furthermore, they included an analysis of the demographic data, suggesting that households known to be least concerned with privacy are the ones who are more vulnerable to smart meter privacy implications. The household segments that are most vulnerable to the privacy implications of smart meters (i.e., young, educated professional individuals and couples) are those with least privacy sensitivity and protective behavior.

The addition of on-site renewable energy production distorts the aggregated net loads as measured by the smart meter, as renewable energy production can cause false negative detection of occupancy status. However, [85] proposed a method to filter out which fluctuations are caused by the energy production installation. However, the authors concluded that a data resolution on the level of seconds is necessary to reliably detect the occupancy in buildings with prosumer status. In case studies without energy generation, data with averaging time steps up to one hour were sufficient to detect the occupancy.

Socioeconomic properties

The ability to infer socioeconomic properties from smart meter data has been studied in-depth based on a combination of smart meter data and survey information. An illustrative selection of these works is included in this paragraph.

The authors of [86] investigated whether household characteristics could be inferred from 30-minute resolution smart meter data. They were able to predict the employment status of the inhabitant(s), the number of persons in the household, as well as the appliance stock with an accuracy of over 70%.

Similarly, [87] and [88] found significant correlations between the shape of the demand profiles and whether or not someone worked from home, the hours of television watched per week, as well as the highest education level of the occupants. Comparable findings were reported in [89], where education was highly correlated with the household's social class. In [90], in addition to the previously mentioned parameters, the age of the head of household was found to also significantly impact the load behavior. However, these aforementioned studies found inferring information about the household income from the smart meter data particularly challenging.

The influence of household income on its load profile was examined more in-depth in [91] and [92]. The authors of [91] found that, on average, households with large incomes have consumption profiles with a relatively large share of consumption in cheap hours, while households with low incomes have consumption profiles with a relatively large share of consumption in expensive peak hours. Analogously, [92] investigated the impact of both wealth and number of occupants of the household. Unsurprisingly, they found that wealthier and larger households displayed increased peak demands. More surprising was that the experimentally derived distributions of the peak demands for households with more than three inhabitants did not agree with the planning guidelines in UK at the time of publication, meaning there was a real risk of asset undersizing in some neighborhoods.

2.5 Drawbacks of smart meters

While individual consumers are understandably most concerned with the privacy-related drawbacks of smart meters, it is important to acknowledge and touch upon other significant disadvantages associated with these devices and the data they generate.

First, smart meters are technologically more complex than their traditional counterparts, which can lead to issues with data logging and communication. For instance, the United Kingdom's Department for Energy Security and Net Zero reported that by the end of 2023, 3.98 million smart meters were faulty [93]. This resulted in consumers either overpaying for their energy, with difficulties in obtaining refunds from their suppliers, or underpaying, leading to debt accumulation during the peak of the energy crisis in 2022 and 2023.

Second, the vast amounts of data generated by smart meters place considerable strain on data centers, carrying both environmental and economic costs [94]. For example, a network of 10 million smart meters collecting data every 15 minutes generates nearly 300 TB of data annually, assuming each record is 5 kB in size. This immense data volume has driven significant research into data compression techniques within the context of smart grids, highlighting the ongoing challenges of managing electric power big data effectively [95, 96].

Third, smart meters, their communication possibilities, as well as controllers for various smart home appliances are vulnerable to cyber-attacks [97]. Furthermore, smart meters can become the point of entry for larger-scale attacks on the energy grid, potentially leading to wider disruptions [98]. Consequently, the widespread roll-out of smart meters has necessitated the development of advanced cybersecurity systems and encryption methods to mitigate these risks [99]. However, the ongoing need for maintenance and software upgrades can lead to additional costs over time, potentially making the overall system more expensive than initially anticipated.

Finally, concerns have been raised by some consumers and advocacy groups about the potential health effects of the electromagnetic fields (EMFs) emitted by smart meters, particularly given their continuous operation and proximity to living spaces. Although these concerns are understandable, studies have not conclusively demonstrated harmful effects at the levels emitted by smart meters. The scientific consensus generally considers the EMF levels produced by smart meters to be safe [100–102].

2.6 Conclusion

This chapter has provided an overview of the broader context in which this work is situated, addressing the challenges posed by climate change and the efforts to decarbonize the residential and transport sector. Heat pumps, electric vehicles, and photovoltaic installations have a prominent position in the residential sector's energy transition, making the electricity production more sustainable, while electrifying the mobility and heat demands of individual consumers.

However, their widespread deployment may lead to technical challenges, particularly concerning their integration into the LV distribution grid, potentially causing grid congestion. Consequently, as the energy transition gains momentum, grid operators are reinforcing their infrastructure to prevent them from becoming bottlenecks. Simultaneously, stakeholders are implementing economic incentives to encourage consumers to spread their consumption.

Smart meters with bi-directional communication capabilities, coupled with the rise of digital business models, are a cornerstone in the residential energy policies of numerous EU member states, facilitating more effective grid infrastructure utilization. Our discussion has explored the economic opportunities presented by digital business models and the diverse applications of smart meter data, both at the individual and aggregated levels.

However, we have subsequently demonstrated the possibility of extracting personal details from smart meter data, ranging from appliance ownership and occupancy detection to socioeconomic insights such as education level and number of persons living in the dwelling.

It is therefore unsurprising that smart meter data has been labeled as personal data under the General Data Protection Regulation of the European Union. Summarizing the GDPR's key stipulations for data sharing and processing, we found that unrestricted access to smart meter data is unfeasible, given the substantial legislative safeguards in place to uphold consumer privacy.

Considering the benefits offered by smart meter data for digital business models, our aim is to harness the vast potential of this data while respecting the privacy of individual consumers. The next chapter therefore introduces the research field of residential load modeling, with the objective of generating synthetic data mirroring real smart meter data characteristics but devoid of any privacy-sensitive information.

3

Residential load modeling

By democratizing access to data at scale, it will help level the playing field, enabling smaller upstarts to compete with more established players that they otherwise might have had no chance of challenging.

Forbes

This chapter provides an in-depth introduction to the context of this dissertation. We conduct a comprehensive literature review on both the generation methods used to create synthetic energy time series and the various generation approaches with which these methods are applied.

Section 3.1 gives an overview of various generation methods. The four most common generation techniques are identified, and their advantages and disadvantages are discussed.

Section 3.2 is subsequently divided in three subsections, introducing bottom-up, top-down, and hybrid models respectively [30, 103, 104]. Bottom-up load models construct synthetic models by modeling the individual behavior of appliances and aggregating them into a synthetic load profile for a single household. In contrast, top-down models start from an input dataset of smart meter data. The purpose of this technique is to identify different consumer segments, and reconstruct their consumption behavior.

To conclude this chapter, Section 3.3 uses the introduced knowledge of available techniques and approaches to position the methodology followed in this dissertation.

However, before moving the discussion to techniques geared towards the modeling of smart meter data, it is necessary to contextualize why synthetic data is the state-of-the-art of data anonymization techniques, and justify why synthetic data is the only considered avenue in this dissertation.

Table 3.1 summarizes the performance of different legacy anonymization techniques and how each method affects the value of its output dataset [105]. The properties of interest that describe the value of the data are:

- **Protection from re-identification risk**: This risk of re-identification is wider than merely re-identification based on the raw dataset under consideration. Rather, this re-identification risk includes the risk of matching anonymous data with publicly available information, or other auxiliary data, to discover the person the data belongs to.
- **Feature statistics**: This measures describes how well techniques are able to capture and reconstruct the quantitative characteristics of individual data attributes
- **Feature correlations**: Feature correlations pertains to the relationship between pairs of features within the dataset, and they indicate the extent to which changes in one feature are accompanied by changes in another feature in the anonymized data.

In order to be useful for machine learning (ML) tasks, anonymized data needs to preserve both the feature statistics, as well as the feature correlations. As seen in Table 3.1, legacy data anonymization techniques fail to produce datasets that simultaneously protect sensitive information and privacy, while retaining the patterns and relationships of the features necessary for ML algorithms to function effectively.

The five techniques included in the comparison of Table 3.1 are summarized on the following page.

Table 3.1: Properties of data anonymization techniques [105].

Method	Protection from re-identification risk	Feature statistics	Feature correlations
Synthetic data	High	High	High
Randomization	Low	Medium	Low
Permutation	Low	High	Very low
Generalization	Medium	Low	Low
Pseudonymization	Very low	High	High
Data masking	Very low	Very low	Very low

- **Randomization**: The features are modified according to predefined randomized patterns. One example of randomization is perturbation, which works by adding systematic noise to data.
- **Permutation**: This involves changing features, either swapping data points between consumers at a certain time, or permute data of the same consumer at different times. The main disadvantage here is the loss of all correlations.
- **Generalization**: In a generalization anonymization scheme, the granularity of the data is reduced in order to preserve privacy. For example, the age of participants can be binned into broader categories, which lessens the risk of re-identification. However, in the generalization in the context of smart meter data implies losing information about the granular peak demands, the property of interest.
- **Pseudonymization**: Pseudonymization removes all personally identifiable information from a dataset, while replacing those personal identifiers with placeholder values. The link between the placeholder values and the personal identifiers is kept separately in the encryption key. The EU defines data as pseudonymized "*if it cannot be attributed to a specific data subject without the use of separately kept additional information*." Consequently, it is a reversible process that allows the re-identification later on if necessary.
- **Data masking**: This is a method of creating a structurally similar but inauthentic version of the considered data that can be used for purposes such as software testing and user training. Perhaps the best-known instance of data masking is related to credit card numbers. Instead of the full credit card number, a string of X's is returned, except for the final four numbers.

Pseudonymization scores high on the three non-privacy properties, and therefore necessitates a more in-depth discussion on its limitations. As the process is reversible (using the encryption key to translate the placeholder values to the original personal identifiers), a pseudonymized dataset is still considered to be personal data according to the GDPR.

GDPR Article 4, Recital 26 states: "*Personal data which have undergone pseudonymization, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person*." [65]

Thus, pseudonymized data must fulfill all of the same GDPR requirements that personal data has to, and the same regulatory barriers for the sharing of this data exist.

However, a degree of ambiguity exists within the regulatory framework on which data can be considered anonymized versus pseudonymized, sparking debates. The same Recital 26 continues its discussion on pseudonymized data and identifiable natural persons as follows: "*to determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly*."

It is clear that the term "*reasonable*" in the regulation can be interpreted in different ways, particularly for smart meter data where only raw metering data is available without location or consumer information. This ambiguity raises concerns about the potential for re-identification through advanced data analytics or the integration of external data sources. Should it be impossible to identify the natural person corresponding to the data with all the means reasonably likely to be used, the data can be considered anonymized, and GDPR would not apply.

In addition to this regulatory ambiguity arising from the term "*reasonable*", Article 4.1.8/2 of the Flemish Energy Decree obliges the distribution system operator (DSO) to make anonymized energy-related data available for scientific purposes [106]. In order to accommodate requests for smart meter data from various market parties and fulfill their legal requirements, the DSO has made a limited dataset of several hundred anonymized consumption profiles available on its open data platform [107]. However, the challenge remains in balancing transparency and data utility against the risks of re-identification.

The Flemish Regulator overseeing the DSO is aware of this ambiguity in the regulation and the impact of publishing anonymized datasets can have, as seen in its 2023 report on the data management activities of the DSO [108]. The regulator states that "*making a new dataset publicly available requires a trade-off between different aspects: feasibility, data quality, ethical aspects (e.g. privacy, intellectual property, protection against threats), direct added value (for citizens, local governments or other public authorities), societal impact as well as legal requirements.*"

It is our expectation that the public availability of granular smart meter data will be a temporary phenomenon and remain limited in size. Once synthetic data establishes itself as equivalently useful for relevant market parties, the direct added value of privacy-sensitive metering data will be significantly lower, no longer justifying the privacy cost of its public availability. Moreover, the evolution of data protection regulations and public awareness may further restrict the accessibility of such data, emphasizing the need for robust synthetic alternatives.

3.1 Modeling methods

This section is based on [109], a 2024 review study on the generation of synthetic energy time series. The authors identified and characterized 169 articles focusing on the generation of synthetic time series, both consumption and production data. The key findings relevant to our literature overview are presented here.

Figure 3.1 displays the share of each method identified in the aforementioned literature review. Four generation methods taken together are used in more than 60% of the considered articles: Monte Carlo, Generative Adversarial Networks (GANs), Markov models, and weighted Random Number Generators (wRNG). Other techniques such as neural networks (NN), Gaussian mixture models (GMM), variational auto-encoders (VAE), Bayesian networks (BN), regression models, aggregations methods as well as data variation and clustering models cover the remaining share of the identified articles.

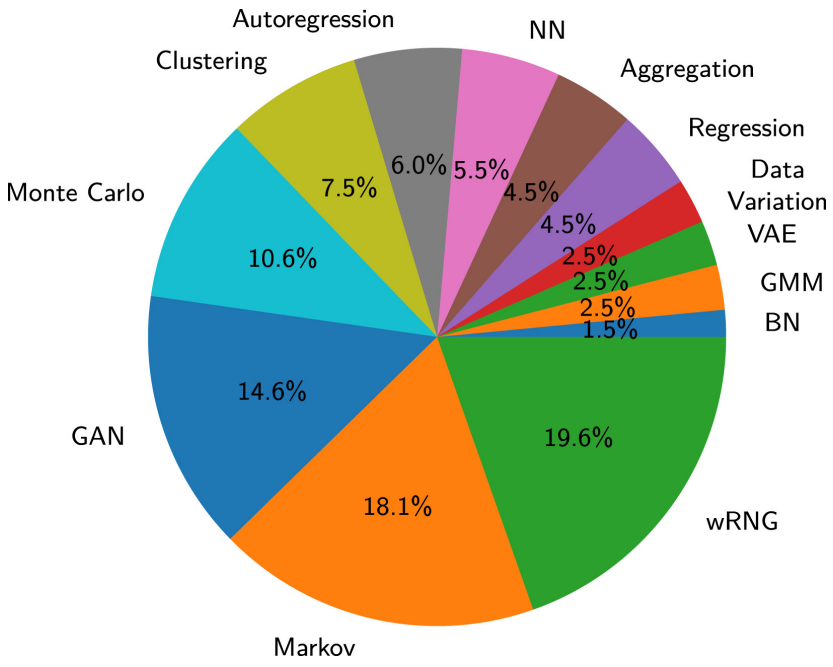

Figure 3.1: Share of the identified methods for generating synthetic energy time series, as included in the review study [109].

The remainder of this section focuses on the principles and strengths of the four most common methods, while individual studies using them for synthetic residential data generation will be discussed in Section 3.2.

Monte Carlo

Monte Carlo simulations leverage random sampling techniques to model the uncertainty and variability inherent in measured time series. In this approach, a model is built based on statistical properties of the original data. By repeatedly running the simulation with random inputs, the method generates numerous possible future scenarios or sequences, each of which represents a plausible realization of the time series. These synthetic series can subsequently be analyzed to assess patterns, trends, and risks under different conditions.

Monte Carlo simulations have several disadvantages [110]. One drawback is that the quality of the synthetic time series heavily depends on the accuracy of the underlying statistical model. Additionally, Monte Carlo methods can be computationally expensive, as they require many iterations to produce a wide range of scenarios, particularly for high-dimensional time series. Moreover, because the approach relies on random sampling, it may struggle to replicate complex temporal dependencies or long-term patterns accurately.

Markov

Markov models model the transition probabilities between different states of a system. The model is built by estimating the transition probabilities from historical data, capturing the likelihood of moving from one state to another. Once the transition matrix is established, synthetic time series can be generated by simulating a sequence of states based on these probabilities. In the context of load modeling, Markov models are particularly useful to capture ON - OFF dynamics of individual appliances in bottom-up approaches, model occupancy of households on a single day, or to model the vacancy of households during holiday seasons.

The disadvantage of Markov models are similar to those of Monte Carlo simulations. Markov models typically require discretizing the system into a finite number of states, which may result in a loss of granularity or over-simplification of continuous processes. Furthermore, the transition matrix is sometimes overly simplistic for time series with long-term dependencies as well as non-stationary behavior. Finally, they struggle with capturing rare events as the transition matrix depends on frequent historical patterns.

wRNG

A weighted random number generator uses random numbers to linearly combine elementary time series. Each elementary time series can represent a specific pattern or characteristic of the data. In bottom-up load models, the time series of an individual appliance can be considered an elementary time series. For top-down models, patterns of the time series such as the trend, seasonal component, or random fluctuation can be considered elementary time series. By increasing its weight, the relative importance of the corresponding time series can be customized.

A first disadvantage is that the quality of the synthetic time series relies heavily on the choice of the elementary series and their corresponding weights. Additionally, this approach assumes linearity in the combination of components, which is not always appropriate for complex, non-linear time series dynamics.

GAN

The rise of generative AI has significantly influenced the field of synthetic data modeling, with GANs quickly becoming a state-of-the-art technique for generating high-fidelity synthetic time series. Their ability to implicitly learn complex, arbitrary data distributions makes them particularly effective in this domain.

GANs consist of two neural networks working in tandem: a generator and a discriminator. The generator aims to produce data that mimics the training set, while the discriminator's role is to distinguish between generated data and real samples. During training, GANs can also be conditioned on additional information, such as weather or calendrical information, to improve the quality and relevance of the generated outputs [111].

Despite the impressive results from well-trained GAN models, they present several challenges. First, GANs function as black-box models, offering limited interpretability, which makes it difficult to understand how specific outputs are generated. As a result, extensive hyperparameter tuning is often required to achieve desirable outcomes, increasing the computational cost of training. Moreover, this challenge is compounded by the inherent instability of GAN training, as the adversarial dynamics between the generator and discriminator can result in mode collapse or failure to converge. Finally, similar to the previous modeling methods, capturing both long and short-term dynamics remains challenging even for GANs, especially for larger time series.

3.2 Modeling approaches

3.2.1 Bottom-up load modeling

While the main focus of this dissertation is on top-down models, we would be remiss not to discuss bottom-up techniques for residential load modeling.

In bottom-up approaches, the total power consumption of an individual consumer is reconstructed by modeling the daily usage patterns of individual devices, encompassing common lighting systems, standard household appliances, and larger equipment like an electric boiler, heat pump, or EV charging station. Time series data representing the power consumption of each electric device are synthesized individually, accounting for factors such as user behavior, weather conditions, and the number of occupants.

State-of-the-art bottom-up load modeling methods typically rely on a combination of extensive measurement campaigns and time-use surveys [104]. These measurements are essential for capturing detailed load profiles of a wide array of household appliances [112–114]. Additionally, time-use surveys serve to connect these load profiles with metadata from diverse consumer demographics, as well as external factors such as weather conditions and calendar information. In instances where measurement data is lacking, methods such as Markov chains and probability density functions are employed to construct density functions for power demands at different times, based on the metadata [115–118].

Two examples of bottom-up load models are the CREST model for describing individual households [113] and the IDEAS and StROBe model for both individual and district-level simulations [119, 120]. Both models utilize time-use surveys as a basis for household occupancy and appliance usage.

Bottom-up modeling approaches are particularly suitable for the so-called "*behind the meter*" applications, e.g., to evaluate the effectiveness of demand response initiatives, where the demand of individual appliances can be curtailed or shifted based on external signals [121, 122]. In contrast, synthetic data obtained through top-down modeling, do not yield any information about the appliances of that individual consumer.

However, bottom-up models do not leverage the abundant data currently collected through the widespread installation of smart meters, opting instead for time-use surveys that are only conducted by a few instances. These are only launched once every 5-10 years in the EU, and annually in the US [123, 124]. Moreover, they encounter scalability challenges due to the reliance on survey data and the complexity involved in integrating the various degrees of freedom related to household appliances, dwelling properties, and occupancy constraints based on demographic factors [125].

3.2.2 Top-down modeling

In contrast to bottom-up models, top-down models treat the residential sector as an energy sink and are not concerned with specific end-uses at the level of the individual household [103]. Hence, studies employing datasets from smart meters fall under the category of top-down models, as they lack information regarding the load behavior of individual appliances behind the meter.

For top-down models, historical consumption datasets supplemented with information on the residents of different dwellings are often used as data source if the corresponding metadata is available and can be used. Such metadata can include (i) the number of residents in the dwelling, (ii) their age, (iii) which major appliances are present, (iv) the size of the dwelling and its isolation level, (v) income, etc. It is therefore unsurprising that such metadata is not often available with the raw consumption data, which can directly be attributed to privacy concerns, as discussed in Chapter 2 [30].

Load profiling for top-down models that use historical datasets in the absence of such metadata generally includes three stages [31]:

1. **Dataset segmentation**: Initially, an unsupervised machine learning clustering method is used to group consumers with similar consumption habits. This can be done on the chronological load profiles themselves, or new application-specific features can be constructed from the measured data.
2. **SLP construction**: Typical or synthetic load profiles are generated for each cluster, often through statistical measures such as the mean or median.
3. **Inferring characteristics**: Consumers' characteristics are inferred from the synthetic load data for each individual cluster, using available or constructed metadata.

Building upon the discussion in Section 3.2.1 and this definition of top-down models, it becomes evident that the terminology of the modeling technique reflects the hierarchical positioning of data inputs in relation to the residential sector. Top-down models use estimates of the total residential sector energy consumption and other relevant variables to attribute energy usage to various characteristics of subsets within the residential sector. In contrast, bottom-up models calculate energy consumption for individual houses (or groups of houses) and extrapolate these findings to represent larger geographic regions.

The literature review on different techniques for dataset segmentation through unsupervised machine learning is part of Chapter 6, while the methods for daily and annual SLP construction are discussed in Chapters 7 and 8 respectively. Instead, the review of this section will only focus on the timescales and resolution of the data used in various modeling studies.

Relatively few evidence-based studies have been conducted that aim to produce synthetic data over a full year, despite its many downstream applications. Rather, the focus of traditional studies has been on short-term generation at the daily or weekly level.

For example, [126] examined the relation between dwelling characteristics and daily consumption behavior. Seasonality was included by considering the daily profiles for different seasons. Similarly, [127] considered the difference between urban and rural dwellings, as well as the impact of the number of bedrooms and inhabitants, by quantifying differences in their averaged daily load profiles.

In the same vein, [128] analyzed and clustered the daily load profiles for a full year to identify the potential for demand response programs for different consumer categories, a similar problem as was tackled in [129, 130].

In [131], three years of smart meter data was used to construct synthetic data. However, daily load profiles were constructed for each of the considered days. The seasonality in the occurrence of each type of daily profile was analyzed, and variations due to both temperature and the COVID-19 pandemic were quantified.

An open-access tool based on Generative Adversarial Networks (GANs) was presented for the random generation of synthetic daily load profiles in [132]. In order to generate synthetic data for successive days, the authors proposed using a Markov chain mechanism to mimic the interday behavior of the real data. Analogously, [133–135] all used GANs to generate synthetic data at the daily level.

However, as a study from 2024 concludes [136]: "*Most existing generative research focuses on short-term energy data, primarily daily profiles, leaving a significant gap in long-term data applications. Privacy concerns and data scarcity further exacerbate these challenges, often limiting the availability of energy data from individual buildings and community power grids.*"

However, while the authors of [136] proceeded to propose a methodology able to handle long-term annual consumption profiles, their research focus was not on the residential sector.

3.2.3 Hybrid modeling

In recent years, models that share characteristics with both the bottom-up and the top-down subcategories have been built to support the bottom-up advantage of incorporating demand side management while using the available data for top-down models. As these models cannot be placed in either category, these have since been called hybrid models [30].

For instance, [137] started from a set of five Irish reference buildings, representative for 82% of the Irish building stock, which is a contribution based on a top-down approach. However, they introduced occupancy-based models and an appliance stock based on number of residents per households for these representative buildings, which is a bottom-up approach. Similarly, [138] combined bottom-up occupancy-behavior modeling with top-down building stock characteristics such as insulation values, square footage, as well as building dimensions.

Furthermore, many hosting capacity studies, including this dissertation for a benchmark validation included in Chapter 9, use a hybrid methodology. The load behavior of EVs and/or HPs is often modeled separately from the smart meter data of the individual households under consideration as this allows for the evaluation of different EV charging strategies or flexible heating behavior. For instance, EV charging profiles are often generated based on measured datasets of arrival and departure times, energy charged per session, as well as assumptions on charging speeds [61, 139]. These additional loads are subsequently overlaid onto the synthetic data of residential consumers to model the impact on the LV distribution grid of higher penetration levels of these technologies.

3.3 Proposed approach

The primary objective of this research is the construction of synthetic load profiles at the annual levels, with a time resolution of 15 minutes. As mentioned in Section 1.2, we advocate for the decoupling of the problem and the adoption of scale-appropriate modeling techniques for different timescales, each tailored to capture the dynamics shaping the load profile in the timescale under consideration.

The high-level methodology is visualized in Figure 3.2 on the following page, decoupling the problem into generating (i) an annual load profile of daily consumptions through GANs, and (ii) daily profiles with a given daily consumption through a wRNG method, using low-frequency trends and high-frequency fluctuations as elementary time series.

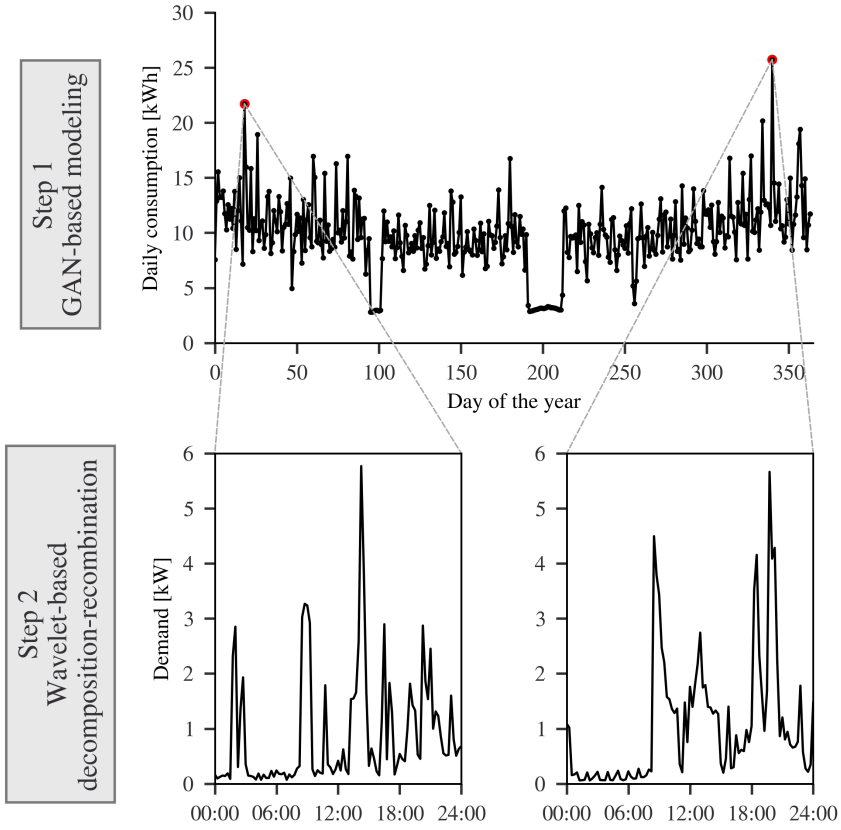

Figure 3.2: Two-step methodology followed in this dissertation for generating an annual load profile at high resolution.

At the daily level, the decomposition-recombination scheme involves breaking down daily load profiles into low-frequency (LF) and high-frequency (HF) components, which are then used as the elementary time series in the wRNG method. This step is detailed in Chapter 7. By confining the wRNG generation process to the daily level, we avoid the method's disadvantages when modeling longer-term dynamics, while still leveraging its flexibility to capture realistic daily peak demands through customizable elementary series.

The HF component captures the stochastic behavior of the individual consumer, while the LF component is a less granular form of the original data. Intelligently recombining the LF component of consumer i with a rescaled and time-shifted version of consumer j subsequently yields synthetic load profiles with realistic variability, as well as consumption and peak behavior.

The properties of several legacy data anonymization techniques were already given in Table 3.1. It is clear that the proposed decomposition-recombination scheme at the daily level finds its origin in a combination of the generalization and superposition scheme, preserving their strengths while circumventing their disadvantages.

- **Low-frequency signal**: A LF approximation of the smart meter reading of consumer i can be considered a generalization of its original data by reducing its granularity.
- **High-frequency signal**: The superposition of a rescaled and time-shifted HF signal of consumer j on the LF component of consumer i can be considered a permutation. Instead of permuting individual measurements, we permute the rescaled and shifted stochastic HF components, again minimizing the risk of re-identification.

At the annual level, GANs are used to capture the complex dynamics and time correlations at the seasonal, monthly, biweekly, weekly and interday level. As discussed in Section 3.1, GANs have demonstrated strong performance in modeling intricate long-term patterns in time series, where traditional methods often fall short. Privacy is conserved as no memorization occurs during the training process. Chapter 8 will explore the GAN generation process in more detail. As Section 3.1 highlighted the instability of GAN training as a major disadvantage, particular emphasis will be placed on careful hyperparameter tuning to mitigate this challenge.

Our methodology therefore follows the top-down modeling approach, starting from a dataset of historic smart meter data. Clusters of consumers with similar properties will need be determined in Chapter 6, and we will show it is necessary to take the constructed metadata into account in the process of linking the generated daily load profiles together in order to obtain synthetic data with sufficiently high fidelity.

However, before moving the discussion towards the construction of peak-based properties and the subsequent clustering approach in Chapters 5 and 6, the datasets used throughout this dissertation still need to be introduced. This is done in the next chapter.

4

Data description

If you torture the data long enough, it will confess.
Ronald H. Coase

In this chapter, an exploration of the various data sources used throughout this PhD dissertation is provided. We aim to provide a transparent overview of the sources accessed and the rationale behind their selection.

First, Section 4.1 is devoted to a discussion on the smart meter data that will be used for the feature construction, clustering, as well as synthetic data generation. As the data itself is subject to a non-disclosure agreement (NDA), this section aims to give an overview of the kinds of consumers and households within the dataset. Special attention is directed towards the extraction of metadata to facilitate the discussion in later chapters.

Section 4.2 subsequently introduces the production data of PV installations, while the generation of EV charging profiles forms the subject of Section 4.3. In each instance, a succinct overview is provided, with references to primary literature for further elaboration. Finally, Section 4.4 wraps up this chapter with a discussion on how the additional loads due to the integration of heat pumps are modeled.

The constructed PV, EV, and HP profiles will be used in Chapter 9 for the validation of the synthetic data in a series of downstream applications linked to the energy transition.

4.1 Smart meter data

The used dataset used in this work comprises 1,422 consumers on the low-voltage distribution grid in two Flemish towns in a suburban area, measured at a 15-minute resolution during one year, leading to 35,040 time points per consumer. The data were provided by Fluvius cvba, the Flemish distribution network operator. The metering infrastructure was installed during a proof-of-concept study on digital meters in Flanders during the period 2010–2014. As more than 3,000 consumers spanning different generations and building types participated in this study, the dataset can be considered sufficiently diverse for consumers on the low-voltage distribution grid.

Several preprocessing steps were undertaken to obtain the final dataset, leading to a reduction from over 3,152 load profiles to 1,422 data entries. These preprocessing steps are as follows:

- A first preprocessing step involving possible missing data was performed by the distribution network operator before providing the dataset for this research;
- Only meters that had measurements for the full year 2013 were included, given the purpose of this dissertation. This excluded 198 entries that ended before 01/01/2014, and 47 entries that started after 01/01/2013.
- We excluded 1,105 meters that logged injected energy in 2013, as it is known that the presence of a PV installation can induce behavioral changes to increase PV self-consumption [140]. Furthermore, the metering data for households with PV installations merely included information on the net consumption and injection, not the gross consumption which is necessary for the proposed methodology.
- Following the Eurostat classification [141], 98 meters exhibiting an annual consumption lower than 1,000 kWh and 125 consumers with annual consumption higher than 15,000 kWh were excluded, as these were assumed to not be representative for typical household habits, or could include small and medium-sized enterprises (SMEs), meaning commercial meters, on the low-voltage distribution grid.
- Finally, data from 157 meters were excluded for either being faulty, or for corresponding to non-residential consumers, but rather to, e.g., communal areas in apartment buildings.

As the raw smart meter data contains no metadata to evaluate and interpret models in the later chapters of this dissertation, we manually construct metadata based on the consumption habits.

In order to construct our own metadata, the 1,422 individual profiles were manually and visually categorized based on the thermal images obtained via heatmaps of their demand profile. This heatmap is the visualization of the matrix obtained by reshaping the 35040×1 vector of the chronological data to a 96×365 matrix. The entries belonging to the days of the start and end of daylight saving time are removed before reshaping the matrix, resulting in a 96×363 matrix. These days contain 92 and 100 data points, and would therefore distort the heatmap.

Based on the obtained heatmaps, five categories are introduced that are able to describe the typical low-voltage consumers in Flanders: four habit-specific categories and one so-called regular residential consumer for all consumers that do not fit one of the four special categories.

The introduced categories are inspired by three available Synthetic Load Profiles (SLPs) for Flanders: one for households with electric heating, one for households without electric heating, and one for non-residential consumers with majority of their consumption during the day [142]. Based on the observations in the heatmaps, two types of electric heating were determined, and a category for households equipped with air-conditioning was introduced.

- **SME profile**: consistent load profile with 9–18h habits on weekdays and absent on weekends, as shown on Figure 4.1a.
- **Electric heating**: consumption late in the evening and at night, superimposed upon a regular consumption profile. Two substructures are observed:
  - **Ripple control heating**: These profiles exhibit the same moment during weekdays when the heating is turned on, and different behavior is observed for weekdays and weekends, as shown on Fig 4.1b;
  - **Continuous heating**: Unlike the ripple control heating, the moment of switching on the heating is stochastic and no difference in heating behavior between weekdays and weekends can be observed, as shown on Fig 4.1c.
- **Air conditioning**: profiles with a significant electric load during summer months, superimposed upon a regular consumption profile. This heatmap is not shown for brevity.
- **Regular residential consumer**: the remaining load profiles not belonging to one of the above categories. There are typically (but not necessarily) characterized by a morning and evening peak, with demands concentrated during the evening as shown on Figure 4.1d.

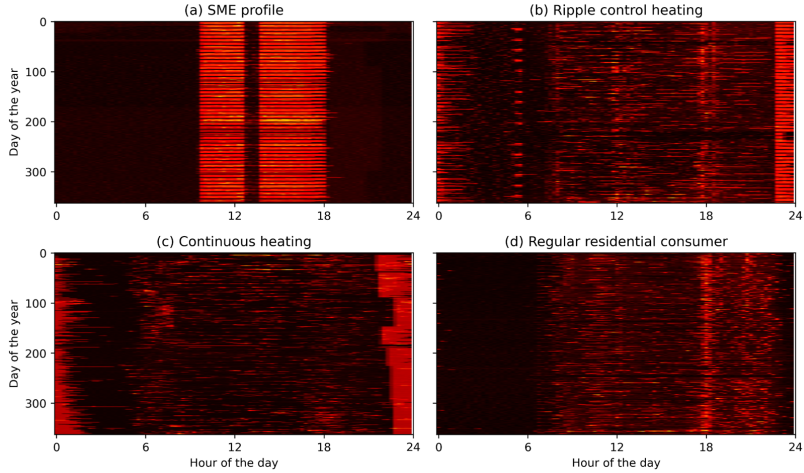

Figure 4.1: Heatmaps of four different types of consumers: (a) an SME, (b) a consumer with ripple control heating, (c) a consumer with continuous electric heating, and (d) a regular residential consumer.

As Table 4.1 shows, the majority of the considered consumers does not fall within a category with specific features such as the SME or the electric heating profiles, but can be considered a regular household. Table 4.1 gives an overview of the number of profiles in each category, split for different consumption ranges. The density of the regular demand profiles is highest in the range of 2–4 MWh per year, consistent with typical consumption ranges for Flemish consumers without a form of electric heating [143]. Similarly, the other categories are mostly situated at higher average yearly consumption.

As the focus of this dissertation is on residential consumers, the 20 identified SMEs are removed, leading to a final dataset of 1,402 consumers.

Table 4.1: Dataset composition by the considered profile categories.

	Regular	Ripple e-heating	Cont. e-heating	SME	Airco
1 – 2 MWh	143	3	1	0	0
2 – 3 MWh	260	19	6	2	0
3 – 4 MWh	254	25	4	2	0
4 – 5 MWh	223	17	15	0	0
5 – 6 MWh	126	14	13	5	0
6 – 7 MWh	86	10	14	1	3
> 7 MWh	114	19	18	10	15
Total	1,206	107	71	20	18

4.2 Solar yield data

The PV yield profile for an installation of a given size, orientation and azimuth is obtained through the methodology presented in [12], starting from the diffuse horizontal irradiance and the direct normal irradiance. The data required for this methodology were provided by the Belgian Royal Meteorological Institute (RMI). The measurements were performed during one complete year, and originate from a weather station situated in West Flanders (latitude 50.90° N - 3.12° E, 25 m above sea level).

In order to obtain the AC power output of the PV system under consideration, a two-step approach is followed:

1. First, the solar irradiance on a tilted plane is calculated. The Hay & Davies transposition model is used to estimate this irradiance, starting from the horizontal irradiance components provided by the RMI [144].
2. Second, the AC power output corresponding with this irradiance is calculated by using PVLib, an open source Python library containing models for simulating the performance of PV energy systems [145].

The properties of the PV model Sunpower SPR-230-WHT-U, Mono-Si and 230 W peak power are used for the simulation. Furthermore, the normalized efficiency curve of the considered inverter is shown in Figure 4.2. It is worth mentioning that the inverter sizing ratio is assumed to be 1.

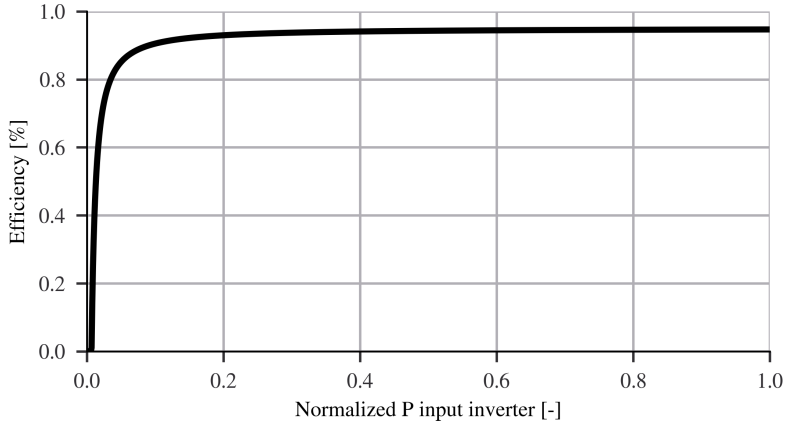

Figure 4.2: Inverter efficiency curve.

4.3 Electric vehicle data

The profiles for EV charging are stochastically generated based on real-world arrival and departure times obtained from ELaadNL [146]. ELaadNL is a knowledge and innovation center for smart and sustainable charging of electric vehicles, an initiative of the joint Dutch grid operators.

We refer to [61] for an elaboration on the EV charging profile generator used for this dissertation, and limit this section to a concise summary of several probabilities involved in the stochastic “dumb" charging profile generator, as smart charging will not be considered for the discussed used cases.

The probability densities used to determine arrival times and the initial SoC of the EVs are depicted in Figure 4.3(a) and (b) respectively [61]. Furthermore, a dynamic charging process is applied to avoid overcharging degradation. The charging process consists of the typical constant current (CC) stage followed by a constant voltage (CV) stage based on the approach presented in [147]. The obtained dynamic voltage and current charging curves, as well as the SoC curve, are presented in Figure 4.3(c) and (d) respectively. Once the voltage reaches a predefined maximum voltage level, the charging stage switches from CC to CV.

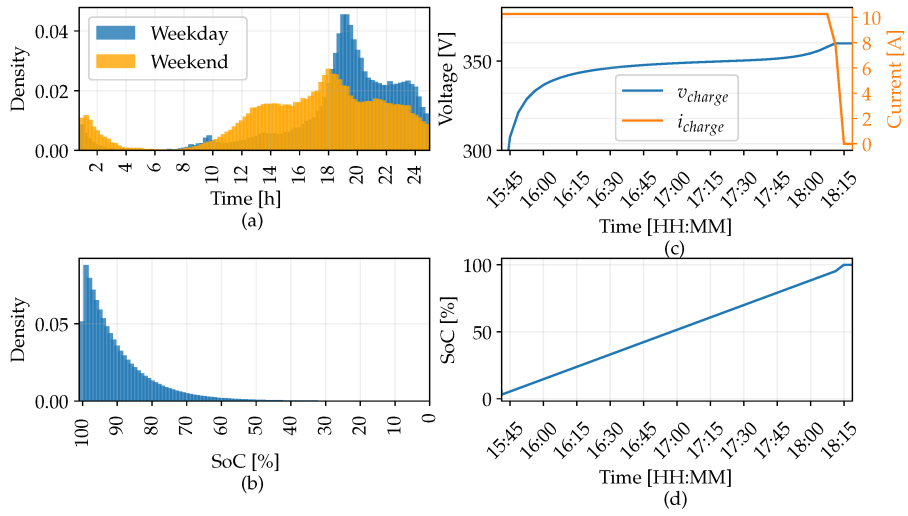

Figure 4.3: Probability densities and dynamic charging curves, with: (a) the distribution of arrival times for weekdays and weekends; (b) the SoC upon arrival at destination. Both right panels are representative of a randomly selected day with: (c) the charging voltage and current; (d) the SoC curve, reproduced from [61].

4.4 Heat pump data

Two separate data sources are used in this dissertation to construct the load profiles corresponding to the integration of heat pumps in residential dwellings. The first data source will provide the normalized load data, while the second source gives insight in the scaling this normalized load.

In the interest of reproducibility, we opt to use the open source '*When2Heat*' dataset for the normalized heating profiles, as published in Scientific Data [148]. This dataset comprises synthetic national time series of both the heat demand and the coefficient of performance (COP) of heat pumps with an hourly resolution. It covers 16 different countries (including Belgium), and spans a total of 10 years, ranging from 2008 to 2018. Furthermore, the variable COP time series are included for different heat sources (air, water, ground), as well as different heat sinks (floor heating, radiators, and water heating).

These estimated COP curves from [148] are given in Figure 4.4. These estimations are obtained based on manufacturer data, as well as adjusted for real-world inefficiencies.

The normalized time series data from *When2Heat* are subsequently rescaled based on the heating demand of individual dwellings under consideration. Hence, information about the building stock of the region is needed. For this, we make use of two studies conducted by the Building Physics research group of Ghent University for the Flemish Energy and Climate Agency.

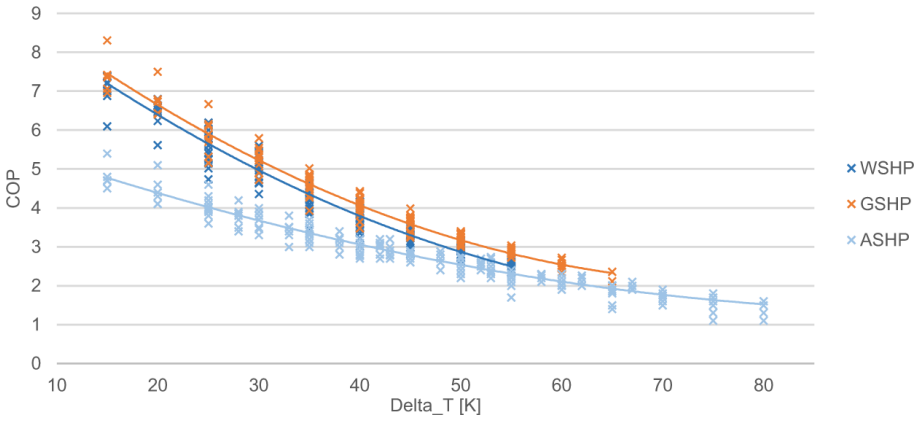

Figure 4.4: Estimation of COP curves, distinguishing between air-source heat pumps (ASHP), ground-source heat pumps (GSHP), and groundwater-source heat pumps (WSHP) [148].

To estimate the heat demand of homes, we use the E-level as used by the Flemish administration. This is a score that indicates how energy efficiency a dwelling is: the lower the E-level, the more energy efficient the building. The E-level depends, among other things, on thermal insulation, air-tightness, compactness, orientation, conscious ventilation losses, as well as the fixed installations (for heating, domestic hot water, ventilation, and cooling). Since 2006, building permits for newly built houses require these houses to reach a certain E-level, with the mandatory level becoming increasingly strict.

In the first relevant study [149], the authors analyzed the total measured energy consumption of over 250.000 individual dwellings and compared it to the E-level. Table 4.2 displays the mean value of the gas usage for households grouped in binned E-levels, depending on whether space heating (SH) with and without domestic hot water (DHW) production is provided through gas. As heat pumps can provide both SH and DHW, this information is of particular interest for case studies investigating the impact of HP integration.

These normalized energy demands can be combined with the results of [150], which allows us to determine the heating demand for different types of dwellings. For this study, the raw data was supplemented with survey data, such as the building topology. We combine the average size of each type of house for different E-levels with the average normalized real annual gas usage to obtain Table 4.3. This table lists the annual gas consumption for a typical detached and terraced house when gas is used to provide space heating and produce domestic hot water.

These annual gas usages can now be transformed to annual electricity demands, in case a heat pump were to provide similar demands. For this, we use the seasonal coefficients of performance (SCOP) of each individual installation. This SCOP averages out the instantaneous COP over a full year.

Table 4.2: Considered average normalized real annual gas usage for space heating (SH) and/or domestic hot water (DHW) production.

	Normalized annual gas usage	
E-level	SH & DHW	Only SH
0 – 20	42 kWh/m²	31 kWh/m²
21 – 40	47 kWh/m²	42 kWh/m²
41 – 60	54 kWh/m²	48 kWh/m²
61 – 80	62 kWh/m²	62 kWh/m²
81 – 100	74 kWh/m²	74 kWh/m²

Table 4.3: Assumed average annual gas consumption (kWh/year) for SH and DHW production for a detached house and a terraced house.

	Annual gas usage	
E-level	Detached house	Terraced house
0 – 20	10,080	9,240
21 – 40	11,280	10,340
41 – 60	12,960	11,880
61 – 80	14,880	13,640
81 – 100	17,760	16,280

We assume an SCOP of 2.60 for the air-to-air heat pump, while air-to-water heat pumps with floor heating reach an SCOP of 3.70. This illustratively leads to the additional annual electricity demands as given in Table 4.4, for a detached dwelling of 240 m². These total annual demands can now be combined with the normalized load profiles from the *When2Heat* dataset to obtain electric load profiles, depending on both the dwelling type and heat pump installation under consideration.

It is important to recognize that using the SCOP is a coarse approximation to obtain average values. First, heat pump efficiency decreases in colder weather. Converting gas consumption to electricity consumption using a constant SCOP may not account for performance variations due to temperature fluctuations, potentially underestimating electricity use during peak cold periods. Second, some heat pump systems rely on auxiliary electric heaters during extreme cold, when the heat pump alone cannot meet demand. This extra electricity consumption is not captured by the SCOP, leading to further underestimation if based solely on gas-to-heat conversions. While using the instantaneous COP for different years would provide more accurate estimates, this level of detail is not necessary for the heat pump use cases presented in Chapter 9.

Table 4.4: Average increase in electrical consumption for detached dwellings of 240 m² if SH and DHW are provided by a heat pump.

E-level	Air-to-air heat pump	Air-to-water heat pump
0 – 20	3,848 kWh/year	2,926 kWh/year
21 – 40	4,306 kWh/year	3,274 kWh/year
41 – 60	4,948 kWh/year	3,762 kWh/year
61 – 80	5,681 kWh/year	4,320 kWh/year
81 – 100	6,780 kWh/year	5,156 kWh/year

5

Definition of peak demands

The definition of a good mathematical problem is the mathematics it generates rather than the problem itself.
Andrew Wiles

As established in Chapter 2, contemporary use of smart meter data often involves knowledge of the timing and amplitude of peak demands at the level of the individual household. Due to this interest, we wish to include features related to them in the clustering scheme.

The question now arises: how can we rigorously define what exactly a '*peak demand*' entails? Despite being ubiquitous in academic literature and studies by grid operators, a rigorous definition is rarely given. The response to this question as proposed over the course of this chapter is derived from following works:

- R. Claeys, T. Delerue, and J. Desmet, "Assessing the influence of the aggregation level of residential consumers through load duration curves," 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe), Bucharest, Romania, pp. 1-5, 2019. [151]
- R. Claeys, H. Azaioud, R. Cleenwerck, J. Knockaert, and J. Desmet, "A novel feature set for low-voltage consumers, based on the temporal dependence of consumption and peak demands." *Energies*, vol. 14.1, p. 139, 2020. [152]

5.1 Properties of peak demands

Intuitively, it is clear that peak demand occurrence at the level of the individual household is the result of a stochastic process at the sub-meter level which depends both on appliance ownership as well as consumer behavior. A local maximum in the load profile arises due to the stochastic simultaneous usage of individual appliances when the household is occupied.

However, this intuitive explanation lacks a mathematical background and leaves a lot of room for interpretation, and as such is not suitable for further analysis. Therefore, we first characterize peaks from five intuitive properties that one would expect from such a definition, before moving on to introduce a more rigorous mathematical framework:

1. Amplitude-wise, all demands above a certain predetermined threshold can be considered peak demands.
2. This threshold has to be established at the level of the individual household. A one-size-fits-all approach is not appropriate for determining which values can be considered peaks.
3. Compared to the total time period under consideration, peak demands only arise a very small fraction of the time.
4. Longer periods of elevated consumption should not count as peak demands at the level of the individual consumers. For example, several uninterrupted hours of elevated consumption due to the operation of an electric heater should not be labeled as peak demand behavior.
5. The timing of a demand peak at the level of the individual consumer is irrelevant for this first definition, as many applications at the individual level are time-agnostic. For example, capacity-based tariffs often do not take the timing of a demand peak into account, only its amplitude.

Based on these properties, we propose to use the Load Duration Curve (LDC) of each individual consumer to define which of its measurements constitute a peak demand on an annual basis. Throughout the analysis in the remainder of this section, we will benchmark the results with these five proposed properties.

The LDC of an individual consumer is obtained by ordering the smart meter measurements in descending instead of chronological order.

5.2 Load duration curve (LDC)

At the macrogrid level, the LDC has traditionally been used by electric utility engineers for network planning purposes, to analyze the utilization of power plants, as well as characterizing the cyclic behavior of electricity demand over a longer time period [153–155].

While the LDC has not traditionally been used to model individual consumers, it was successfully used by Poulin et al. to investigate the value of peak shaving for commercial, institutional, and industrial consumers [156].

Encouraged by these findings for non-residential consumers, the analytical form of the load duration curve is used in this work to construct the peak-based features. Based on the shape of the LDC, a threshold unique to each consumer can be proposed, and every demand higher than this threshold can be considered as a peak.

The analytic expression introduced by Poulin et al. is taken as the starting point for our analysis [156]. Traditionally, $P^i(t)$ denotes the chronological demand data of a specific consumer i. The LDC $\mathcal{P}^i(\tau)$ corresponding with this demand profile can subsequently be described by following equation:

$$\mathcal{P}^i(\tau) = 1 - a\tau - b\tau^c + \frac{d}{1 + e^{f(\tau - g)}} - \frac{d}{1 + e^{-fg}} \tag{5.1}$$

The variables $\mathcal{P}^i$ and τ in the expression of the LDC respectively denote the normalized demand and “time", or ordered rank of the consumption value, i.e., both scaled such that their range spans the interval $[0, 1]$. This allows for a scale-independent comparison of consumers, merely comparing the shape of their LDCs, as visualized in Figure 5.1.

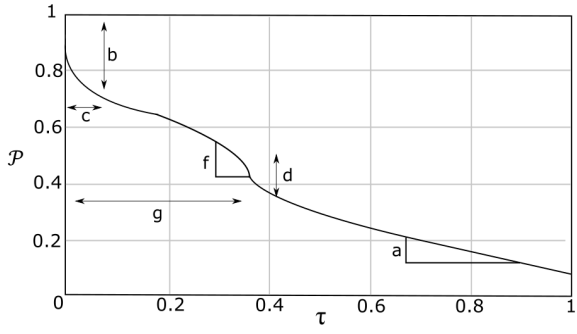

Figure 5.1: Shape of the 6-parameter LDC for commercial, institutional and industrial consumers as proposed by Poulin et al. [156]

The six parameters included in Equation (5.1) show a clear connection to customer operations, and thus are relevant for consumer clustering purposes. The peak height and duration are correlated with b and c respectively, while parameters d, f and g are linked to respectively the height, slope and location of the step. Finally, a yields information about the slope of the linear segment of the LDC. These six parameters and their qualitative relation to the shape of the LDC are displayed in Figure 5.1.

A constrained least-squares fit is used to determine these six parameters for each individual consumer in the considered dataset. To the best of our knowledge, prior to our initial reporting, the proposed relation was not yet validated for residential consumers, as [156] considered a dataset comprised of commercial, institutional and industrial consumers.

Therefore, an initial validation of the model for individual residential consumers has been done via the coefficient of determination R^2 of the performed fit. To perform the curve fitting, the Python package `lmfit` was used. The constraints used during the curve fitting algorithm for each individual parameter are given in Table 5.1. These boundaries were chosen nearly identical to those used for the LDC fitting procedure in [156].

The only deviation from the constraints is the lower bound of the g parameter. Therein, a lower bound of 0.1 for g was assumed. However, residential consumers are more peak-intensive and their peaks are more stochastic. It is expected that this behavior is reflected in the shape of the LDC with a shorter duration of the peak and step. Therefore, the lower bound for g, the parameter linked to the location of the step, can be taken smaller than the aforementioned 0.1. A value of 0.025 was chosen for this lower bound.

The median coefficient of determination R^2 of the curve fittings is 0.987, highlighting an excellent fit between the proposed mathematical expression and the empirical LDC. Similar results for the coefficient of determination were reported in [156], allowing us to extend the application range of the proposed relation to include residential consumers in addition to the previously validated commercial, institutional and industrial consumers.

Table 5.1: Constraints considered for the LDC curve fitting procedure.

Parameter	Minimum	Maximum
a, b, c	0	1
d	0.005	1
f	25	$+\infty$
g	0.025	1

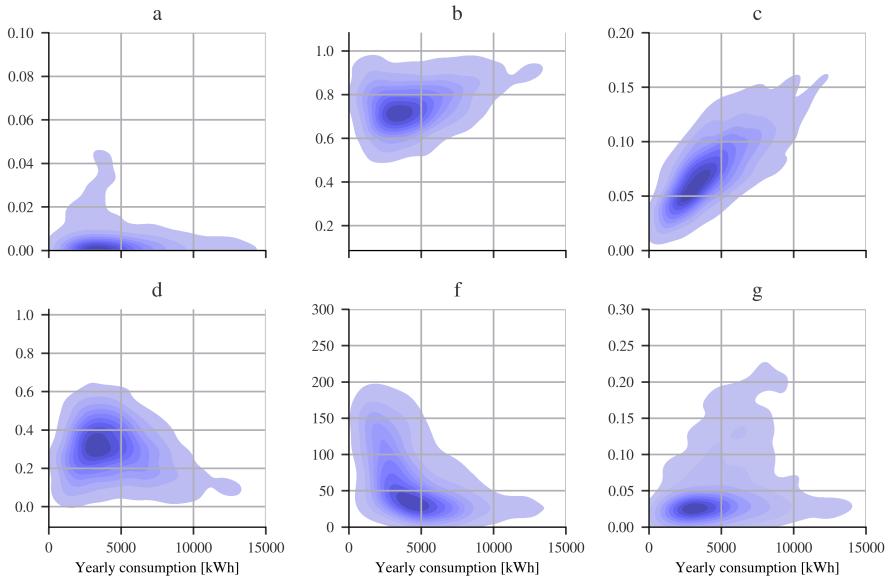

Figure 5.2: 2D kernel density estimation (KDE) of the parameters included in Equation (5.1), plotted versus the yearly consumption.

The distributions of the six parameters obtained from the curve fitting procedure are given in Figure 5.2.

It is noteworthy that negligibly low values of the parameter a are obtained, corresponding to a nearly non-existent slope for the residential LDC. Intuitively, one would indeed expect residential consumers to spend the majority of their year on a certain base load, i.e., the aggregated standby demand of the appliances in the household. As such, this would correspond to a saturation effect towards this standby demand being present in the household LDC for $\lim_{\tau \to 1} \mathcal{P}^i(\tau)$, in contrast to the decreasing slope in Equation (5.1).

Consequently we can simplify this six-parameter expression, and propose a five-parameter LDC model for residential consumers in Equation (5.2), which is visualized in Figure 5.3.

$$\mathcal{P}^i(\tau) = 1 - b\tau^c + \frac{d}{1 + e^{f(\tau - g)}} - \frac{d}{1 + e^{-fg}} \tag{5.2}$$

A second possible improvement entails incorporating possible correlations between the values of the parameters in Equation (5.1) and properties of the consumer, such as the annual consumption.

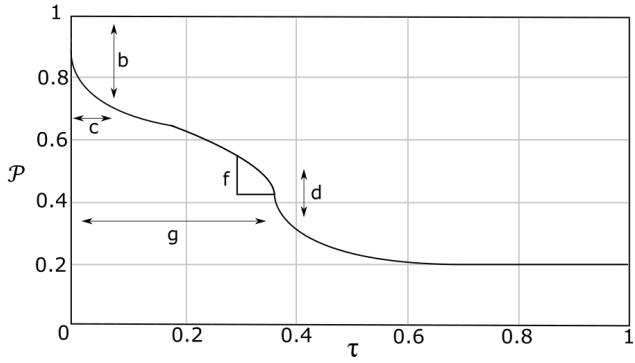

Figure 5.3: Shape of the proposed 5-parameter LDC.

Despite a large spread being present in the scatterplot, the parameter c describing the power law in Equation (5.1) is noticeably correlated with the yearly consumption. Consequently, Equation (5.3) fixes the parameter c at the value $c_0 + c_1 Y$, and reduces the number of parameters to be fitted to four. The values of c_0 and c_1 are determined by an ordinary least-squares fitting procedure on the relation between Y, the yearly consumption in kWh, and c, as shown in Figure 5.2.

$$\mathcal{P}^i(\tau) = 1 - b\tau^{c_0+c_1 Y} + \frac{d}{1+e^{f(\tau-g)}} - \frac{d}{1+e^{-fg}} \tag{5.3}$$

The evaluation of all three LDC models, based on the coefficient of determination R^2, is listed in Table 5.2. As expected given the observed values of a, the 6-parameter and 5-parameter models exhibit identical performance. Furthermore, while a decrease in median R^2 value can be observed for the 4-parameter model, this value is still acceptable. However, the observed mean value is significantly lower and exhibits an increasing difference with the median value, highlighting that the 4-parameter model of Equation (5.3) leads to a worse fit for a non-negligible amount of consumers.

Table 5.2: Comparison of the fitting result of the three considered LDC models.

Model	Median R^2 value	Mean R^2 value
6-parameter model, Equation (5.1)	0.987	0.977
5-parameter model, Equation (5.2)	0.987	0.977
4-parameter model, Equation (5.3)	0.968	0.937

As expected, given the high spread in the linear relation between c and the yearly consumption, the reduction in accuracy of modeling the LDC is a trade-off that has to be made in order to incorporate the dependency on the consumer's yearly consumption and reduce the complexity of the considered model.

It is worth noticing that an evaluation solely based on R^2 leads to an incomplete picture. Indeed, a very high R^2 value can be an indication of overfitting. Furthermore, more complex models will always be able to capture more variability in the data. Hence, researchers typically strive for a trade-off between the goodness of fit and simplicity of the model. To this end, both the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are considered. Both of these criteria measure the relative quality of statistical models, balancing the performance of the model with the number of parameters it contains. The criteria are defined as follows:

$$AIC = 2k - 2\ln(\hat{L}) \tag{5.4}$$

$$BIC = k\ln(n) - 2\ln(\hat{L}) \tag{5.5}$$

Where k denotes the number of parameters, $\ln(\hat{L})$ is the log-likelihood of the model on the data to incorporate the goodness of fit, and n is the number of data points. It is clear that these criteria reward the goodness of fit by the second term, the log-likelihood function, while penalizing increasing complexity of a model by the first term, the number of parameters.

Therefore, we use the AIC and BIC to select the most appropriate model among the set of three, balancing goodness of fit with model complexity. When comparing multiple models, the model with minimal value of each criterion is considered the best model when considering the trade-off. For 91% of the consumers in the considered dataset, both criteria are minimal for the 5-parameter model. The reduction in complexity for the 4-parameter model does not offset the lower performance in the fitting procedure, leading to the 5-parameter being preferred.

Given the importance of the fitted parameters of the LDC for the remainder of this work, the complexity of the model is of lesser importance. However, given that the 5-parameter model is preferred for the vast majority of the dataset whilst exhibiting superior performance in the evaluation based on the coefficient of determination, further analyses are performed on Equation (5.2), the model that exhibited superior performance in the fitting procedure.

5.3 From LDC to peak demands

The validated analytic expression of the load duration curve can now be used to introduce a binary classification for peak demands, i.e., all values $\mathcal{P}^i(\tau)$ for τ smaller than a certain threshold τ^* can be considered peaks for the individual consumer while all other values cannot. The challenge now lies in determining τ^*, the value of this threshold.

The only condition a proposed expression or value for τ^* has to fulfill for the purposes intended in this work is that it has to be sufficiently small in order to yield usable results. Although the term '*usable*' implies a certain level of arbitrariness, it should be clear that a threshold value that labels 50% of all demands on yearly basis as peaks is not practical for e.g. peak shaving algorithms. Therefore, given the continuous nature of the load duration curve, it is inevitable that any proposed threshold value will have its own advantages and disadvantages.

We propose using the point of maximum curvature as this threshold, for τ sufficiently small. This guarantees the presence of the point of maximum curvature in the exponential decay present in the LDC. Intuitively, the curvature of a function is the amount by which this function deviates from a straight line in a certain point. Therefore, the maximum of this curvature function denotes the point where the curve has the sharpest bend.

For τ sufficiently small, the LDC as defined in Equation (5.2) can be approximated by Equation (5.6), which is dominated by the power law responsible for the peak demand features and the steep decay of the LDC:

$$\mathcal{P}^i(\tau) \approx 1 - b\tau^c. \tag{5.6}$$

Using the point of maximum curvature of Equation (5.6) as the threshold value to define the area of peak demands has two major advantages. First, this threshold is different for each individual as it depends on the shape of the individual load duration curve, allowing for a differentiation among low-voltage consumers. Second, the point of maximum curvature for an analytic function can be unambiguously described analytically. The curvature function $\kappa(\tau)$ of Equation (5.6) is given by [157]:

$$\kappa(\tau) = \frac{\left|(\mathcal{P}^i)''(\tau)\right|}{\left[1 + \left[(\mathcal{P}^i)'(\tau)\right]^2\right]^{\frac{3}{2}}} \tag{5.7}$$

Maximizing $\kappa(\tau)$ with respect to τ yields following value for the point of maximum curvature:

$$\tau^* = \left(\frac{c-2}{b^2 c^2 (2c-1)} \right)^{\frac{1}{2(c-1)}} \tag{5.8}$$

This expression can now be used to determine the point of maximum curvature for every LDC in the consumer dataset. Figure 5.4 illustratively displays the exponential decay of an LDC together with its point of maximum curvature. Before τ^*, the LDC is decaying rapidly, but the rate of decay is slowing down as τ increases. After the point of maximum curvature, the LDC continues to decay, but the rate of decay becomes progressively slower.

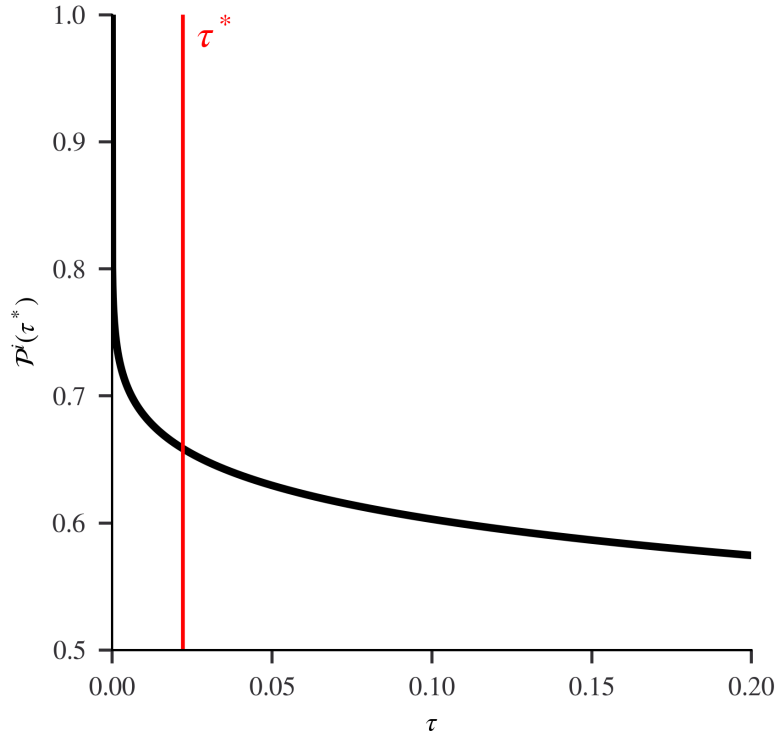

Figure 5.4: Example of the point of maximum curvature for the LDC described by $\mathcal{P}^i(\tau) \approx 1 - b\tau^c$ at small values of τ, with $b = 0.5$ and $c = 0.1$.

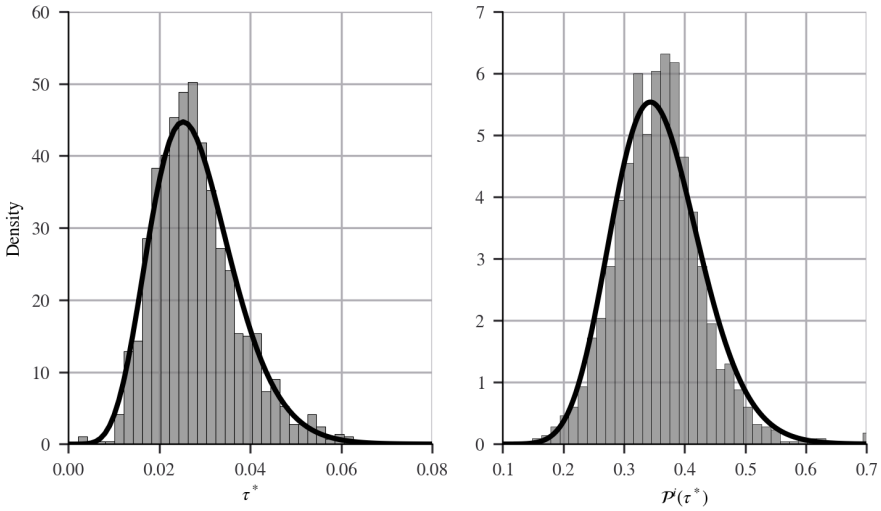

Figure 5.5: Density histogram of the calculated values of τ^* and $\mathcal{P}^i(\tau^*)$, as defined by Equation (5.8).

The histogram of the calculated values of τ^* and the corresponding value $\mathcal{P}^i(\tau^*)$ for the considered dataset are given in Figure 5.5. A beta probability density function is successfully fitted and shown to be able to describe the density functions, as shown overlaid in Figure 5.5. The distribution of τ^* has a $10-90$ percentile range of [0.017, 0.041], with a mean value of 0.028.

Translating this mean value of the normalized time τ to a yearly basis means that, on average across the distribution, 2.8% of the values on a yearly basis can be labeled as peaks, corresponding with 981 values of the 35.040 data points. Furthermore, the distribution of $\mathcal{P}^i(\tau^*)$ shows the large potential of peak shaving initiatives for residential consumers: the mean value of $\mathcal{P}^i(\tau^*)$ is 0.35, i.e. 35% of the original maximum demand.

5.4 Conclusion

In this chapter, we have proposed a methodology to unambiguously define what exactly a '*peak demand*' of an individual consumer encompasses. Despite being a common term in academic literature and studies by regulators and grid operators, no rigorous definition that suits the purpose of this dissertation could be found in literature.

Inspired by its use by electric utility engineers for network planning purposes, we introduced the load duration curve (LDC) to describe the demand behavior of individual consumers. After validating several analytic expressions for the LDC, a 5-parameter model was retained, which is used for the remainder of this work.

Determining the point of maximum curvature in the exponentially decaying part of the LDC allows us to define a unique threshold for each individual consumers. Values above this threshold will be considered peak demands. Moving forward, the findings from this chapter will serve as the foundation for the feature construction and clustering scheme used in Chapter 6 to group similar consumers together, and to shed light on the stochastic nature of these peak demands.

6

Consumer categorization

The world is now awash in data, and we can see consumers in a lot clearer ways.
Max Levchin

In this chapter, we shift our focus from the theoretical approach of defining the peak demands to its practical application in consumer categorization. Specifically, our objectives in this chapter are twofold: first, to cluster similar consumers together in an unsupervised machine learning algorithm taking into account several peak-based properties, and second, to investigate the stochastic nature of peak demands.

This chapter is organized as follows: Section 6.1 provides the necessary background on the unsupervised machine learning algorithms used in this chapter. Section 6.2 introduces the features that will be used for the clustering. The clustering results are presented and discussed in Section 6.4, while the distributional analysis on the constructed features in Section 6.5 sheds light on the predictability of peak demands and its implications for the applicability of demand response and peak shaving initiatives.

Sections 6.2 to 6.5 of this chapter are derived from:

- R. Claeys, H. Azaioud, R. Cleenwerck, J. Knockaert, and J. Desmet, "A novel feature set for low-voltage consumers, based on the temporal dependence of consumption and peak demands." *Energies*, vol. 14.1, p. 139, 2020. [152]

6.1 Unsupervised machine learning

Machine learning (ML) is a discipline of computer science that gives machines the ability to implicitly learn without being explicitly programmed in a rule-based manner. In unsupervised learning, the data does not contain any labels or metadata, meaning the input data is not paired with the desired outcome data. Instead, unsupervised learning is most often used for tasks such as clustering, dimensionality reduction, or anomaly detection.

For this chapter, we will mainly focus on clustering. During a clustering process, the machine organizes unsorted data according to parallels and patterns, classifying similar inputs together in categories. Consequently, we can use unsupervised ML to partition a dataset of smart meter data belonging to different consumers, in order to group those consumers with comparable properties together to get insight in their consumption behavior.

We first discuss two traditional techniques used to cluster smart meter data in literature. Section 6.1.1 first explores the popular k-means algorithm, while Section 6.1.2 reviews the hierarchical clustering scheme. In subsequent sections, we examine the applications for smart meter data, starting with the construction of custom features in Section 6.2.

6.1.1 K-means clustering

The primary problem that k-means clustering tries to solve is the task of partitioning a given dataset into k distinct, non-overlapping clusters. Hence, it is important to note that the number of clusters is the input of the k-means algorithm.

Mathematically, given a set of vectors $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$, where each vector is an observation in a d-dimensional space, k-means partitions these n observations into $k(\leq n)$ sets $\mathbf{S} = \{S_1, S_2, ..., S_k\}$, to minimize the within-cluster sum of squares (WCSS). This is expressed in Equation (6.1).

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in S_i} \|\boldsymbol{x} - \mu_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i \tag{6.1}$$

With μ_i the mean, or centroid, of points in the set, or cluster, S_i:

$$\mu_i = \frac{1}{|S_i|} \sum_{\boldsymbol{x} \in S_i} \boldsymbol{x} \tag{6.2}$$

Equation (6.1) is typically solved iteratively. First, k points are randomly initialized. These are called the means, or cluster centroids. Second, each observation $\boldsymbol{x}_n$ is categorized to its closest mean. Afterwards, the mean's coordinates are updated, their new values being equal to the averages of the observations that have been categorized in that cluster. Finally, this process is repeated for a given number of iterations.

Despite being widely used, the k-means algorithm has some major drawbacks. First, its output is heavily influenced by the initial selection of cluster means. Different initializations can lead to different final clustering results, and in some cases, it may converge to suboptimal solutions. From a mathematical point of view, the k-means algorithm can get stuck in a local optimum instead of searching for the global optimum.

Second, k-means tends to produce clusters of roughly equal size. That makes it unsuitable for datasets with clusters of varying densities. Furthermore, this makes it highly sensitive to outliers. As the sum of squared distances between data points and the cluster centroids is minimized, these outliers disproportionally influence the position of the centroid.

The final drawback is that the dependence on the number of clusters k as input value. As the optimal k is often not known in advance, this has to be determined. However, determining this is often somewhat subjective, with many evaluation metrics being used to determine the optimal number of clusters, given the results for a sweep of k.

Some common evaluation methods are the silhouette score [158], the Calinski-Harabasz index [159], the Dunn index [160], the Davies-Bouldin index [161], as well as the gap statistic [162]. These indices are commonly used together with the k-means algorithm in academic literature for smart meter data clustering.

For example, [163] used a k-means clustering algorithm to construct 19 clusters for Spanish electricity consumers based on their smart meter data. The optimal number of clusters was determined based on the Davies-Bouldin index, as well as the average within cluster variance. Analogously, [164] used both the silhouette score and Davies-Bouldin index to k-means cluster London-based consumers into three distinct categories. In [165], k-means was used together with the Davies-Bouldin index to obtain between 8 and 10 clusters for a dataset of smart meter data of Irish consumers.

The authors of [166] used a combination of the Dunn, Davies-Bouldin and silhouette index to cluster daily load profiles of Korean consumers into 21, 27, or 16 different clusters, depending on which residential site was considered. The same three indices were used in [167] to cluster Irish consumers into four distinct consumer groups.

6.1.2 Hierarchical clustering

Hierarchical clustering is the common alternative to k-means clustering. While k-means aims to partition a dataset in a fixed number of clusters, hierarchical clustering (as the name suggests) seeks to build a hierarchy of clusters. As such, in contrast to k-means, it does not require the number of clusters to be determined in advance, which is a major advantage.

A second advantage is that hierarchical clustering allows for a graphical representation of the results in a so-called dendrogram. A dendrogram is a tree-like structure that displays the relationship between clusters, and it visually depicts the sequence of cluster fusions or splits that occur during the hierarchical clustering process.

The vertical y-axis of a dendrogram represents the distance between clusters or data points. The height of each fusion or split in the dendrogram corresponds to the distance at which the clusters were merged or split. Consequently, a dendrogram is a visual aid that helps to interpret what occurs when moving from k to $k+1$ clusters. Furthermore, cutting the dendrogram at a certain distance divides the dendrogram into distinct clusters. The choice of y-value determines the granularity of the clustering, with lower y-values resulting in more clusters and higher y-values leading to fewer clusters.

There are two main categories of hierarchical clustering:

1. **Agglomerative clustering**: This is a so-called “bottom-up" approach. It starts with each observation as a separate cluster, and subsequently iteratively merges the closest pairs of clusters until only one cluster remains.

2. **Divisive clustering**: In this “top-down" approach, all data points start in a single cluster, and the algorithm recursively divides the dataset into smaller clusters until each data point is in its own cluster.

Divisive hierarchical clustering is less common and computationally more intensive than agglomerative clustering. As such, we will only consider agglomerative clustering.

While agglomerative clustering offers its advantages over k-means, it comes with its own set of challenges. It is computationally intensive and often requires storing distance or linkage matrices, which can consume a significant amount of memory for large datasets. This can limit the scalability of hierarchical clustering approaches compared to k-means clustering.

Agglomerative clustering has a large degree of freedom on which measure for (dis)similarity is used to determine which clusters should be combined. This measure for dissimilarity is a combination of (i) a distance d between single observations of the considered dataset, such as the traditional Euclidian distance, and (ii) a linkage criterion, which specifies the dissimilarity of sets as a function of the pairwise distances of one or multiple elements in those sets.

To illustrate the impact of different linkage criteria, consider several common criteria:

- **Complete-linkage clustering**, also called furthest neighbor clustering, is based on maximum distance. The similarity D of any two clusters A and B is the similarity of their most dissimilar pair:

$$D(A, B) = \max_{a \in A, b \in B} d(a, b) \tag{6.3}$$

- **Single-linkage clustering**, also called nearest neighbor clustering, is based on a minimum distance. The similarity of any two clusters is the similarity of their most similar pair.

$$D(A, B) = \min_{a \in A, b \in B} d(a, b) \tag{6.4}$$

  It is clear that the complete-linkage criterion will tend to lead towards spherical clusters, whereas the single-linkage criterion will lead to elongated clusters. Indeed, the nearest neighbor clustering will produce clusters where nearby elements in the cluster have similar properties, but there is no guarantee that the observations at opposite ends of the cluster will exhibit similar behavior.

  Consequently, we will not use either of these linkage criteria. Instead, we will consider the more popular Ward's linkage for the hierarchical clustering algorithm in Section 6.4.

- **Ward's linkage** aims to minimize the variance when merging two clusters, effectively seeking to minimize the increase in total within-cluster variance after merging [168]. It selects the merge that will lead to the smallest increase in variance, and therefore assumes that this merge will result in the most compact and homogeneous clusters.

  Ward's linkage scheme is of interest when expecting clusters of varying sizes, as this criterion tends to produce more balanced and cohesive clusters compared to other linkage criteria [168].

Similar to k-means clustering, hierarchical clustering with different linkage criteria has successfully been used to group smart meter data. In [169], Japanese consumers were hierarchically clustered in 5 distinct groups based on their smart meter data, using both the complete and Ward's linkage criteria. Greek consumers were categorized into ten clusters in [170] using Ward's linkage, as well as other clustering schemes. Smart meter data of Portuguese consumers were clustered in 10 groups using Ward's linkage criterion in [171].

6.2 Feature construction

As discussed in the previous section, clustering algorithms are performed on a set of vectors $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$, where each vector is an observation in a d-dimensional space. Intuitively for smart meter data, this clustering occurs on the raw smart meter data over a certain period, often one day or a single week.

An alternative method to clustering via chronological measurements is grouping consumers based on similar properties, also called features. This dissertation follows the feature construction and evaluation approach. The advantages of using a limited set of features during a clustering process are multifold. First, artificial overfitting due to high dimensional data can be avoided [172]. Furthermore, computational time is saved and allows easier interpretation if the features are chosen to be application-dependent [173].

Features can be constructed by performing operations on the default chronological features, e.g., combining all daytime consumption in one single feature. However, more advanced features can also be constructed, ranging from features generated in the frequency domain [174, 175], to features related to the shape of the distribution of the load, such as the load factor [176].

Features constructed in literature are often application-dependent, i.e., depend on the goal of the work. One goal of this chapter is to investigate the temporal connection between consumption and peak demand behavior, to gain insight in the stochasticity of residential peak demands. Therefore, the features in this work are linked to either the consumption or the occurrence of peak demands.

Previous studies incorporating temporal properties of these peaks in the clustering process either take the timing and the amplitude of the daily peak demand into account [177, 178], or use statistical measures of the distribution of the measurement data [179].

Two different temporal levels relevant for consumers on the low-voltage distribution grid are considered for these time-dependent features: the daily and the weekly level. At the daily level, the time periods are defined based on the time of day. Similarly, at the weekly level the distinction between weekday and weekend is maintained. In order to distinguish between intervals $\mathcal{J}$ defined on either the daily level and the weekly level, two notations are introduced: $\mathcal{J}^d$ and $\mathcal{J}^w$. The superscript d stands for daily, w for weekly.

While the definition of the intervals $\mathcal{J}^w$ is unambiguous, i.e., weekdays versus weekends, distinct intervals at the daily level $\mathcal{J}^d$ for residential consumers are not universally agreed upon. In [177], Haben et al. identified four key time periods for residential consumers: overnight, breakfast, daytime and evening period.

Inspired by their findings, the distinction as listed in Table 6.1 is introduced for the daily level. In this work, the daytime period is further subdivided in a morning and afternoon range. Furthermore, while the daytime period in [177] ended at 15:30, it is extended to 18:00 for this work.

Other temporal levels can easily be incorporated in the feature set, e.g., the seasonal influence by including four time periods at the annual level corresponding with the seasons. However, this seasonal variation is omitted, as these features were found to not significantly impact the clustering result and rather obfuscated the results, limiting the ease of interpretation.

Based on these time periods at two different temporal levels, a two-pronged approach is introduced. The first class of features considers the relation between the temporal property and the consumption: which fraction of the demand occurs during a certain predefined time interval? In contrast, the second class of features considers the temporal properties of the peak demands: when do these peak demands occur? In order to unambiguously define which values constitute a peak, we analyze the analytic fitted form of the load duration curve as defined in Chapter 5.

Table 6.1: Definition of the considered time periods $\mathcal{J}^d$ at the daily level, based on the hour of the day, h.

$\mathcal{J}^d$	Definition
Early morning	$h \in$ [06:00 – 08:30 [
Morning	$h \in$ [08:30 – 12:00 [
Afternoon	$h \in$ [12:00 – 18:00 [
Evening	$h \in$ [18:00 – 22:30 [
Night	$h \in$ [22:30 – 06:00 [

The individual features are suitable to characterize consumers, e.g., for assessing household compatibility with renewable energy sources (households with high daytime consumption are more favorable for rooftop-integrated PV installations without a battery), or for the timing of the individual peak demands, which is beneficial information for distribution network operators.

However, it is the knowledge on the fraction of the demand combined with the simultaneous occurrence or absence of peak demands in that time period that can elucidate the stochastic nature of these peak demands. Consumers that consistently exhibit a disproportionate amount of peak demands in a certain time period can be targeted for peak shaving initiatives, either via demand response programs or by utilizing an energy storage system.

Temporal dependence of consumption

Let $\mathcal{B}^i_{\mathcal{J}^x}$ be the subset of all measured values $P^i(t)$ of consumer i that occur in one of the previously defined time periods $\mathcal{J}^x$, with the superscript x denoting the considered temporal level. This yields following definition of this subset:

$$\mathcal{B}^i_{\mathcal{J}^x} = \left\{ P^i(t) \mid t \in \mathcal{J}^x \right\}, \quad x \in \{d, w\} \tag{6.5}$$

The fraction $f^{i,c}_{\mathcal{J}^x}$ of the demand of consumer i in time period $\mathcal{J}^x$ is given by:

$$f^{i,c}_{\mathcal{J}^x} = \frac{\sum_{y \in \mathcal{B}^i_{\mathcal{J}^x}} y}{\sum_t P^i(t)} \tag{6.6}$$

This definition yields a total of seven features: five for the daily level, two for the weekly level. However, as the subsets $\mathcal{B}^i_{\mathcal{J}^x}$ for a given temporal level x are disjoint by construction, the sum of $f^{i,c}_{\mathcal{J}^x}$ over all $\mathcal{J}^x$ for a fixed x is equal to 1. Therefore, this reduces down to five linearly independent features: four for the daily level, one for the weekly level.

Temporal dependence of peak demands

The features related to the peak demands are treated in a different way than those linked to the consumption. While the amplitude of the demand $P^i(t)$ at a certain point in time is important to determine the fraction of consumption that happens in a time interval, only the presence of peak demands is of importance for the second set of features, not the size of the peaks.

Let $\mathcal{D}^i$ be the subset of all measured demand values $P^i(t)$ of consumer i that are classified as peak demands, as defined in Chapter 5. As the LDC is normalized with respect to the annual peak demand P^i_{max}, the value $\mathcal{P}^i(\tau^*)$ has to be rescaled:

$$\mathcal{D}^i = \left\{ P^i(t) |\ P^i(t) \geq \mathcal{P}^i(\tau^*) \cdot P^i_{\text{max}} \right\} \tag{6.7}$$

Analogous to the previous section, let $\mathcal{D}^i_{\mathcal{J}^x}$ now be the subset of $\mathcal{D}^i$ that occurs in time period $\mathcal{J}^x$:

$$\mathcal{D}^i_{\mathcal{J}^x} = \left\{ P^i(t) \mid P^i(t) \geq \mathcal{P}^i(\tau^*) \cdot P^i_{\text{max}} \wedge t \in \mathcal{J}^x \right\}, \quad x \in \{d, w\} \tag{6.8}$$

The number of peak demands per time interval is given by via the cardinality of the set $\mathcal{D}^i_{\mathcal{J}^x}$, i.e. $\left|\mathcal{D}^i_{\mathcal{J}^x}\right|$. Hence, the fraction of peak demands for consumer i in a certain time period, $f^{i,p}_{\mathcal{J}^x}$, is given by:

$$f^{i,p}_{\mathcal{J}^x} = \frac{\left|\mathcal{D}^i_{\mathcal{J}^x}\right|}{\left|\mathcal{D}^i\right|} \tag{6.9}$$

Analogous to the features related to the temporal aspect of the consumption behavior, this leads to another five linearly independent features. Consequently, this brings the number of considered linearly independent features for the clustering algorithm up to ten parameters.

The constructed features can now describe the temporal behavior and distribution of the consumption and peak demands, as illustrated for one randomly chosen regular household, household 802, in Figure 6.1 on the following page.

Both the fraction of the consumption and the fraction of peaks are shown for each time period at the daily and the weekly level. Major differences between the distribution describing the consumption and peaks can be observed. At the weekly level, 65% of the household's peaks are observed in the weekend, while only 35% of the consumption occurs during weekends. Similarly, more than 25% of consumption for this consumer happens at night, as defined by Table 6.1, while 10% of the peak demands lie in this time period.

It is this difference between distributions of consumption and peak behavior at the same temporal level that forms the subject of the following sections, as the presence or absence of differences can clarify whether or not peak demands tend to be more stochastic or more deterministic.

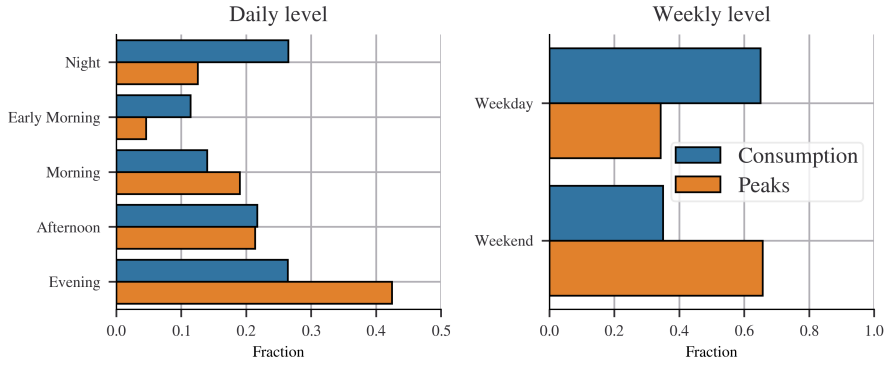

Figure 6.1: Example of the 14 features describing the temporal behavior of the consumption and peak demands at the daily and weekly level of household 802, $f_{\mathrm{J}x}^{802,c}$ and $f_{\mathrm{J}x}^{802,p}$ respectively.

6.3 Feature set transformation

One additional step is performed on the proposed set of features before proceeding to either the clustering algorithm or the distributional analysis: the feature transformation. Depending on the proposed methodology for each analysis, a different feature transformation is more appropriate. Therefore, this section discusses both the proposed methodology for each performed analysis, as well as the corresponding most suitable feature transformation.

6.3.1 Clustering algorithm

No additional information or metadata is included in the dataset of load profiles. As such, the true underlying structure or the optimal amount of clusters to segment the dataset into is unknown. Therefore, unsupervised machine learning is used to cluster those profiles that exhibit similar behavior.

The majority of the rich body of literature available on the unsupervised clustering of load profiles, whether chronologically ordered profiles or based on a constructed feature set, is based on one of two techniques: either a variant of the k-means clustering algorithm or via agglomerative clustering. In the structured literature review on the classification of consumption profiles performed by Tureczek and Nielsen, 65% of the considered papers included a k-means-based method, while another 29% performed analyses based on agglomerative clustering [180].

There are several differences between k-means and agglomerative clustering, both from a conceptual viewpoint, as well as the computational aspect. Agglomerative clustering offers a visualization in a so-called dendrogram of the clustering results, intuitively showing how substructures in the dataset emerge when dividing or merging clusters. Furthermore, when a feature set is used as input for the agglomerative clustering, further analysis on the merging of clusters offers the possibility of tracking which features are the driving force that distinguish clusters. However, agglomerative clustering is a so-called greedy algorithm: at each step, the two closest clusters as defined by a linkage method are merged together. Therefore, agglomerative clustering techniques are prone to yield a sub-optimal solution instead of a global optimum.

In contrast, given an input k, the number of desired clusters, a k-means algorithm partitions the dataset into k clusters. However, k-means tends to get stuck in a local minimum instead of the global minimum. The main challenge for a k-means approach lies in finding the optimal amount of clusters. From a computational point of view, k-means is preferable for larger datasets as the time complexity for k-means algorithms typically is linear in the input data size, $\mathcal{O}(n)$, while the time complexity for agglomerative clustering is quadratic, $\mathcal{O}(n^2)$.

In this work, an agglomerative clustering algorithm with Ward's linkage method is used, as implemented in Python's `SciPy` package [181, 182]. The main contribution of this work is introducing and validating a novel feature set. Therefore, the visualization and emergence of substructures in the dataset in the clustering process is of major importance, justifying the choice for an agglomerative clustering algorithm. The proposed linkage method minimizes the total within-cluster variance for each merging step.

Following the arguments presented by Kang and Lee in [183], it is a necessary condition for clusters to have a usable, non-trivial size in order to be useful in real life applications according to expert opinions. Therefore, Ward's linkage method can be deemed appropriate, as the tendency of clustering algorithms to propose singular clusters that contain outliers is avoided.

For the proposed feature set, Ward's linkage method for agglomerative clustering relies on the Euclidean distance between the 10 linearly independent features in the 10-dimensional feature space. Therefore, obtained results will depend on the scale of the input features. However, when looking at both Table 6.1 and Figure 6.1 it is clear that the proposed features are not yet at the same scale.

By construction, the proposed time periods are not of the same scale, e.g., the weekend period is not the same length as the weekday period, nor is the early morning of similar length as the night interval. Therefore, even a uniform distribution would not lead to similarly scaled features, leading to a distortion of importance of several features.

Therefore, an initial transformation is performed that rescales the features based on the length of their time period such that in the case of a uniform distribution, the value of all features $f_{\mathrm{J}x}^{i}$ would be equal to 1. Any deviation of a uniform distribution will then lead to a deviation of this unity value for each parameter, while avoiding an artificial inflation of the importance of an individual feature or one set of features.

However, of the ten proposed linearly independent features, eight are defined on daily basis, while only two are defined on weekly basis. While this initially proposed transformation aims to give each individual feature the same weight, the two sets of features defined on different temporal levels are not a priori equally represented in the feature set. Consequently, instead of transforming the features on weekly basis to be equal to 1 in the case of a uniform distribution, they are assigned an additional weighting factor equal to 2 to partially offset the numerical advantage of daily features.

In summary, the two sets of features proposed in Section 6.2 are transformed in a two-step transformation before being used as input for the hierarchical clustering, using Ward's linkage method. First, the features are rescaled based on the length of the time interval in which they are defined, which leads to individual features of the same scale.

In the second step, an additional weighting factor is assigned based on the amount of features for each temporal level. A weighting factor of 2 is proposed for the weekly-level features, which partially offsets the numerical advantage daily-level features have in the proposed feature set. Further increasing this weighting factor would put a higher emphasis on the difference between weekdays and weekends in the clustering algorithm.

6.3.2 Distributional analysis

The distribution of features $f_{\mathrm{J}x}^{i}$ at the daily or weekly level x can yield interesting information. As mentioned before, households with high daytime consumption are ideal candidates for PV installations, whereas households that exhibit a large amount of peak demands in a certain time interval, could be targeted via demand response initiatives.

However, it is the difference between the distributions describing the consumption and peak behavior at the daily or weekly level that yields information about the disproportionate presence of peak demands at a certain time interval, and thus about how stochastic the presence of peak demands are for an individual household. Therefore, two measures are proposed to investigate these distributions.

At the level of the individual distributions, we propose using the concept of entropy at the daily or weekly level to characterize the variability of household behavior. Similar to the goal of this work, [184] introduced entropy to study the variability of households, not with respect to features based on consumption of peak behavior, but based on the variability of consumption behavior described by the frequency of different representative daily load shapes during the year. Shannon entropy as introduced in information theory is defined in Equation (6.10), with x_i being a possible outcome and $p(x_i)$ the probability of this outcome [185].

$$H(x) = -\sum_{i=1}^{k} p(x_i) \ln p(x_i) \tag{6.10}$$

This definition of entropy has several interesting properties for this dissertation. First, in the case of a uniform distribution, the entropy reaches its maximum and thus yields maximum uncertainty. Second, any deviation from a uniform distribution results in a decreasing entropy, and thus less uncertainty. If there is no uncertainty, then the entropy becomes 0. In order for these properties to hold in the analysis of the introduced features, it is important that these features are consistent with the assumptions in the Shannon entropy. First, in the case of a uniform distribution, the entropy becomes maximal. A uniform distribution for the consumption behavior would entail having the same consumption at each time period.

However, as mentioned in Section 6.3.1, the periods as defined in Section 6.2 are not of equal length, which leads to inequal features in the case of a uniform consumption distribution. Therefore, the features are rescaled based on the length of the interval for which they are defined such that a uniform consumption distribution leads to identical consumption-related features. Furthermore, Equation (6.10) is defined for probabilities $p(x)$. As such, the features defined on the five periods at the daily level are rescaled to 0.2, while those defined at the two periods at the weekly level are rescaled to 0.5.

The second part of the distributional analysis entails a comparison between the distributions of the consumption behavior and the occurrence of peak demands at the daily and weekly level. As such, a measure for the distance between these two distributions has to be introduced.

The Wasserstein-1 distance is used to characterize the distance metric between two probability distributions [186]. The definition of this distance as integrated in Python's `SciPy` package is used [182]:

$$l_1(p, q) = \inf_{\pi \in \Gamma(p,q)} \int_{\mathbb{R}\times\mathbb{R}} |x - y| \, d\pi(x, y) \tag{6.11}$$

Here, p and q are two distribution functions, and $\Gamma(p, q)$ is the set of probability functions on $\mathbb{R} \times \mathbb{R}$ whose marginals are p and q on the first and second marginals respectively. This Wasserstein distance is also commonly called the earth-mover's distance. Intuitively, it can be seen as the minimum amount of "work" that has to be done to transform one distribution into the other, if each distribution could be considered as a pile of earth. The "work" takes into account both the distance it has to move, as well as the amount of earth it has to move. As such, distributions P and Q that are different over "long" (horizontal) regions will be far away from each other in the Wasserstein distance sense [187].

It is this property that makes it suitable for this work. As the time periods at the daily level were arbitrarily introduced, a distance metric that takes the horizontal difference into account instead of performing a point-wise comparison partly compensates the arbitrary nature of the definition. For a given consumption behavior, this allows us to identify distributions of the peak demands that are closer in time.

This property is illustrated in Figure 6.2, where an artifical distribution of the consumption is compared with two peak demand distributions. As the first distribution of the peak demands has a maximum chronologically closer to that of the normalized consumption, its Wasserstein distance is lower.

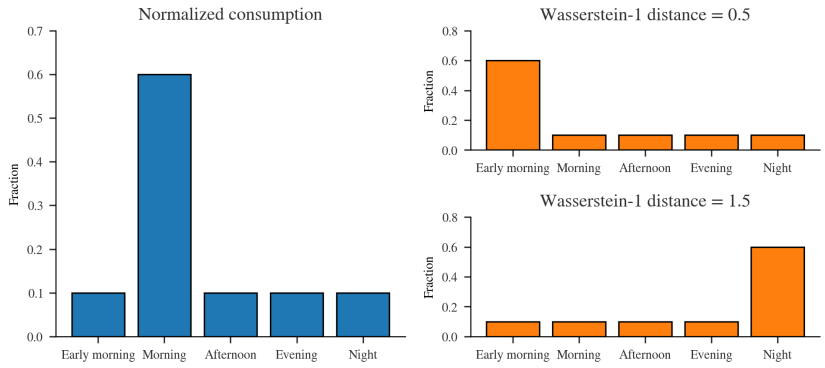

Figure 6.2: Illustration of how the Wasserstein distance is able to capture chronological differences between the feature distributions.

6.4 Clustering results

First, the results for the agglomerative clustering algorithm are illustrated based on the calculated dendrogram. The results for a low number of features are benchmarked to the available synthetic load profiles in Flanders, subsequently highlight how differences in feature behavior lead to the emergence of distinct and compact clusters, and argue how this knowledge can be leveraged from the viewpoint of demand response programs or peak shaving initiatives.

The dendrogram visualizing the hierarchical clustering process using Ward's linkage method on the proposed feature set is shown in Figure 6.3. Two horizontal cuts are included in the figure. The black line at $y = 30$ denotes the height where three clusters are obtained. This can serve as an initial benchmark, as there are three synthetic load profiles available for low-voltage consumers in Flanders: residential with and without electric heating, and non-residential.

The red line was chosen such that 10 disjoint clusters emerge, leading to the color threshold of the highlighted clusters in the dendrogram. This threshold of 10 clusters was chosen based on two independent studies stating that for practical considerations, the total number of clusters should not exceed 10 [173, 183]. This argument is based on the opinions of industrial experts, as these clusters are often used for tariffing or marketing purposes.

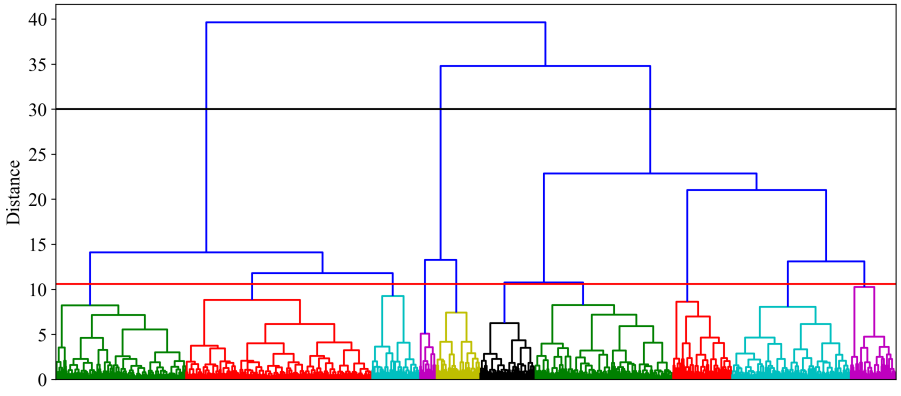

Figure 6.3: Dendrogram obtained via hierarchical clustering, with a color threshold highlighting 10 clusters. Individual profiles are given on the x-axis, while the y-axis denotes the distance.

First, it is necessary to benchmark the clustering result to the available residential SLPs in Flanders. As the color threshold and further discussion in this section is based on 10 clusters, the analysis is performed based on the highlighted 10 clusters. By tracking the merging clusters into the three branches of the dendrogram at the cut $y = 30$, a benchmark can be performed.

Figures 6.4 and 6.5, displaying respectively the distributions of the 14 untransformed features for the individual clusters, and the distributions of the yearly consumption of the consumers assigned to each clusters, allow for an interpretation of the obtained clusters based on consumer properties.

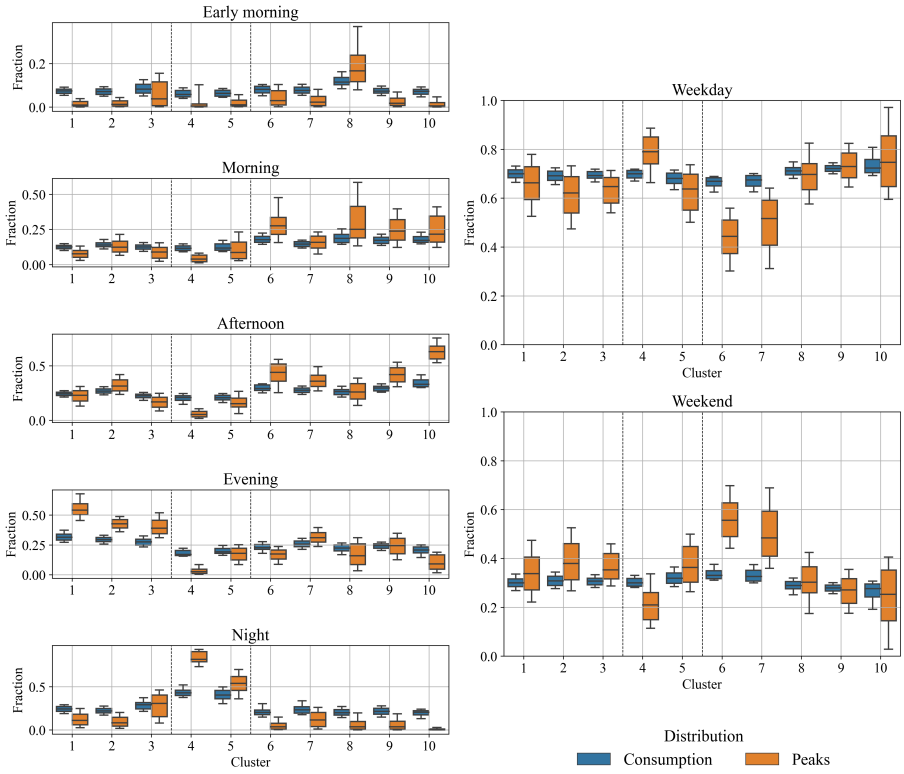

Figure 6.4: Boxplots visualizing the distribution of the 14 untransformed features for 10 clusters, with the 10 features at the daily level displayed on the left and the 4 features at the weekly level on the right. The whiskers of the boxplots describe the [10, 90] percentiles.

The first branch separates into clusters 1 – 3, the second into clusters 4 – 5, while the final branch leads to clusters 6 – 10. The clusters originating from these three branches are partitioned by dashed lines in Figure 6.4 for an easier comparison.

The following discussion on the benchmarking of the results is based on the observed feature distributions in Figure 6.4. The first branch groups consumers with a high fraction of consumption and peaks in the evening, which is typical for regular households.

The second branch, containing clusters 4–5, groups consumers with a high fraction of the consumption and peaks at night. This is encouraging, as this could indicate the presence of electric heating, either space heating or domestic hot water production, one of two major categories of residential consumers.

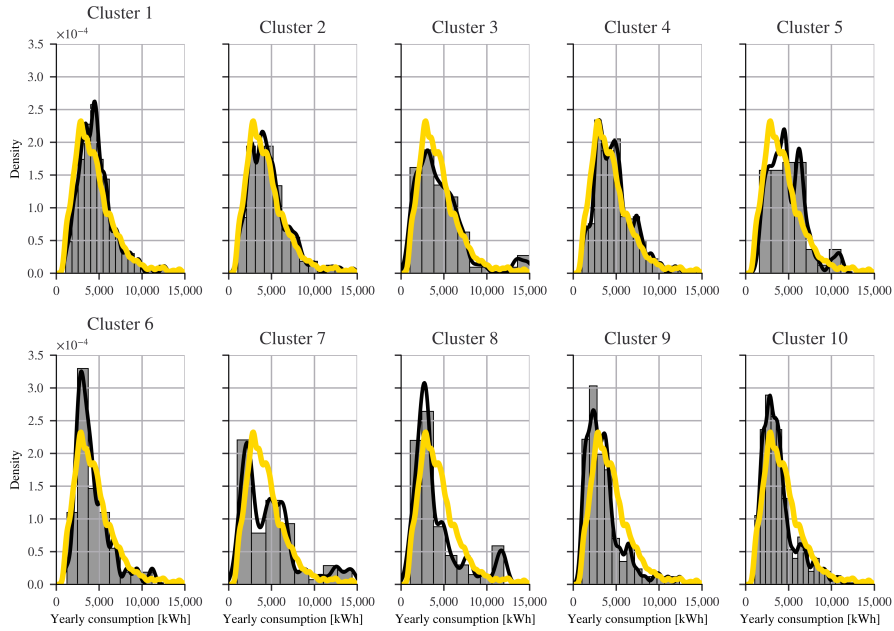

Figure 6.5: Density histograms with shared y-axes, displaying the distribution of the yearly consumption of the individual consumers assigned to each of the 10 considered clusters. The black plot denotes the density for the individual cluster, while the plot in yellow indicates the density of the full dataset.

The interpretation of the third branch is less straightforward, as the properties of the clusters composing this branch are more diffuse: (i) clusters 6–7 group consumers with a disproportionate amount of peaks during the weekend, (ii) cluster 8 collects the consumers with a significant amount of peaks during the early morning, whereas (iii) clusters 9–10 exhibit a large number of peaks during the morning and afternoon.

It is clear that for each time period and in the same branch of the dendrogram, the differences between the fractions of total consumption for that period are limited. Rather, the peak-based features are the driving force to merge similar clusters together before the consumption-based features emerge as driving force to form three clusters reminiscent of the Flemish SLPs. Furthermore, the clustering process yields compact clusters with comprehensive results.

This illustrates the usefulness of a feature set that includes the temporal properties of peak demands, especially with the advent of capacity-based tariff schemes for low-voltage consumers. With the introduction of capacity-based tariffs, it is no longer sufficient to know when consumption occurs. Additional knowledge about when peak demands tend to happen is vital to offer consumers the most suitable techno-economic solution.

As a post-hoc validation of the performance of the proposed feature set in determining customer categories, the clusters of the different consumer types in the dataset as introduced in Section 4 are determined and given in Table 6.2. Furthermore, the mean values of τ^* and $\mathcal{P}(\tau^*)$ within each cluster are given in Table 6.3.

Clusters 4 and 5 are predominantly populated by households with electric heating, while cluster 10 groups households with high daytime consumption. However, not all profiles with electric heating are categorized inside clusters 4–5. This is further investigated in Figure 6.5, which displays the density plots of the yearly consumption for each individual cluster compared to the density plot of the full dataset. Matched against the density plot of the complete dataset, clusters 4, 6, 8 and 10 are skewed towards households with lower to average yearly consumption in the Eurostat classification.

Table 6.2: Relative frequency of consumer categories over the 10 clusters.

	1	2	3	4	5	6	7	8	9	10
Ripple e-heat	2%	6%	6%	18%	36%	9%	20%	2%	2%	0%
Cont. e-heat	14%	8%	15%	11%	31%	1%	10%	4%	4%	0%
Airco	17%	17%	6%	0%	6%	0%	33%	6%	17%	0%
Regular	17%	25%	5%	0.1%	1%	7%	17%	8%	16%	5%

Table 6.3: Mean value of τ^* and $\mathcal{P}(\tau^*)$ in each of the 10 clusters.

	1	2	3	4	5	6	7	8	9	10
τ^*	2.8%	2.8%	2.8%	2.9%	3.0%	2.5%	3.1%	2.4%	2.7%	2.6%
$\mathcal{P}(\tau^*)$	36%	36%	34%	35%	38%	48%	37%	35%	34%	34%

This distribution for cluster 4 is expected and can clarify the diffusion of households with electric heating over different clusters. As the demand profiles of these households can be considered an aggregation of the profile of a regular household with a load profile of an electric heating appliance, the features connected to the peak demands are intrinsically linked to the behavior of that load profile and the timing of the peak demands without the electric heating. The heating load profile for households with otherwise relatively low yearly consumption dominates the aggregated load profile, and consequently encounter the majority of their consumption peaks during the night, consistent with the behavior of cluster 4. For households with electric heating in e.g. cluster 3, the consumption and peak demands during the evening outweigh those occurring during the night.

Table 6.3 shows that there is little difference between clusters in the average behavior of τ^* and $\mathcal{P}(\tau^*)$, except for two notable clusters. Clusters 6 and 8 exhibit a significantly lower mean value of τ^*. For cluster 6, this reduction in τ^* is accompanied by a significant increase of $\mathcal{P}(\tau^*)$. Hence, the exponential decay in the LDC is not as steep or long as in other clusters, and the LDC moves from its exponential decay into the step function faster than average. In contrast, cluster 8 exhibits a smaller mean τ^* with a similar $\mathcal{P}(\tau^*)$ as other clusters, meaning the exponential decay of the LDC is steeper than average.

It can be concluded that the proposed feature set is able to capture the known consumer categories from existing SLPs, and thus passes our self-imposed benchmark test. Three clusters can be attributed to known differences in behavior for low-voltage consumers: the presence of electric heating is captured in clusters 4–5, while the high daytime consumption is present in cluster 10. Deviations from these two clusters for electric heating can be traced back to differing contributions of the electric heating load to the total yearly consumption of the households.

6.5 Distributional analysis

In this section, we investigate the stochasticity of residential peak demands. The common scientific consensus is that peak demands at the residential level are stochastic and hard to forecast. However, we aim to quantify the stochastic nature arising from the intuitive relation between consumption and peak demands in a certain time period.

To investigate this relation, the distributions of the features at the same time levels are analyzed. On the one hand, the Shannon entropy is used to characterize the variability of each type of feature. On the other hand, the Wasserstein-1 distance is used for an in-depth analysis of the stochastic nature of the peak demands, by comparing the distributions of the household consumption and peak demand behavior respectively.

The variability of the daily and weekly consumption and peak patterns are described by the entropy of their probability distribution, where the individual fractions are normalized with respect to the length of the considered time period. A uniform distribution with maximum uncertainty leads to a maximal value of the entropy, while the absence of uncertainty leads to an entropy value of 0. For example, a situation where all peak demands occur during the night due to an electric heating would lead to 0 entropy at the daily level for the consumption probability distribution.

The obtained distributions for the entropy at the daily and weekly level for the consumption and peak probability distributions of the full dataset are given in Figure 6.6. At the daily level, the peak demands exhibit a much larger variability than the consumption. This is unsurprising, given the continuous nature of the consumption. At the weekly level, this difference is less pronounced.

A beta distribution was successfully fitted to each individual density histogram. The 2-parameter beta probability distribution, defined on the interval [0,1], is defined as follows, with $a > 0$ and $b > 0$:

$$f(x, a, b) = \frac{\Gamma(a, b)x^{a-1}(1-x)^{b-1}}{\Gamma(a)\Gamma(b)} \tag{6.12}$$

The beta function offers several properties that make it suitable to describe the obtained distributions. First, it has a finite support: the regular 2-parameter beta function in Equation (6.12) has a [0,1] support. As the entropy can vary from 0 to a maximum of $-\ln(0.2)$ for the daily level and $-\ln(0.5)$ for the weekly level, the finite support of a rescaled and shifted beta function is appropriate.

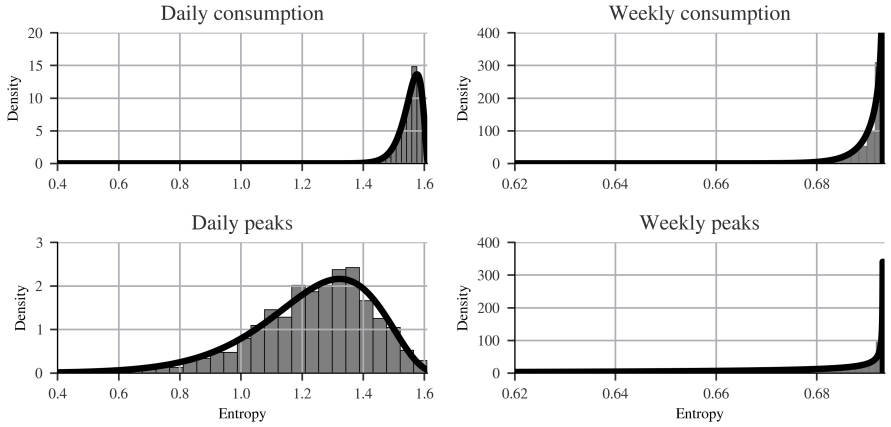

Figure 6.6: Density histograms of the entropy of the consumption and peak demands at the daily and weekly level, with a fitted beta probability density function overlaid in black.

Second, as can be observed in Figure 6.6, the shapes of the daily and weekly behaviors differ significantly. The two shape parameters a and b in the definition of the beta probability function allow us to describe the four distributions with the same formula. For the distributions shown in Figure 6.6, it merely means that $b > 1$ for the distributions at the daily level, while $b < 1$ for those at the weekly level.

The relation between the entropy and the clusters obtained in Section 6.4 is investigated in Figure 6.7, which displays the mean values of the entropy for each individual cluster.

The significantly lower entropy of the probability distribution describing the peak demands can be traced back to the clustering results. The overwhelming presence of peak demands during the night period results in low entropy for cluster 4, while cluster 10 exhibited a majority of its peaks during daytime. Similarly, half of the peak demands for cluster 1 occurred during the evening. On a weekly basis, clusters 6 – 7 showed a significant amount of peak demands during the weekend, leading to a lower entropy for this period.

A low entropy of the probability function describing the peak demands can be taken as an indicator for the presence of a large amount of peaks in a certain time period, which can be leveraged to target demand response programs or peak shaving initiatives. Furthermore, a clear relation can be observed between the obtained clusters on the introduced feature set and the entropy values of the peak demands. The lower values in certain clusters can be traced back to differing intercluster consumer operations at the daily or weekly level.

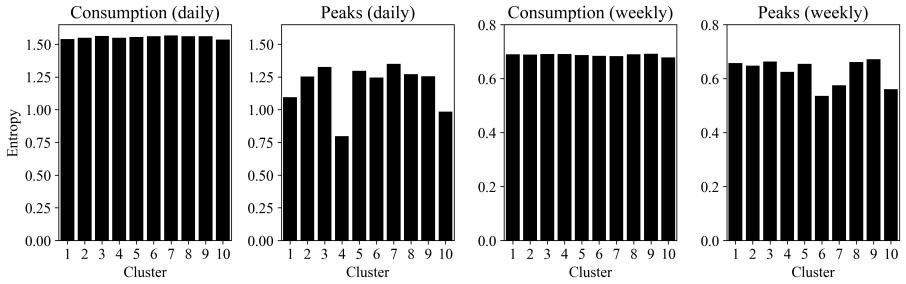

Figure 6.7: Mean values of the entropy of the distributions of the normalized consumption and peak demands for each cluster.

However, the stochastic nature of these peak demands remains an open question. The probability distributions of the peak demands tend to be significantly more variable than those of the consumption behavior, according to the entropy. Even so, this entropy as a single variable does not reveal anything about whether or not the amount of peaks in a certain time period is disproportional relative to the consumption in that time period.

Therefore, the Wasserstein-1 distance is used to quantify the difference between the probability distributions of the consumption and peak demands at the daily and weekly level for each individual consumer. A larger distance corresponds to a stronger deviation of the peak distribution from the distribution of the consumption, and thus peaks are more deterministic.

Figure 6.8 and Figure 6.9 display the distributions for the Wasserstein-1 distances at the daily and weekly level respectively, for each cluster. Analogous to Figure 6.5, the distribution of the Wasserstein-1 distance calculated for each profile in the full dataset is included for comparison to cluster-specific behavior. The distributions of the Wasserstein-1 distances further confirm the findings concerning the behavior of consumers constituting each cluster.

At the weekly level, clusters 6 and 7 show a major deviation from the dataset behavior, due to the presence of a disproportionate amount of peak demands in the weekend. Similarly for the daily level, cluster 4 displays a large Wasserstein-1 distance, pointing to the electric heating which pushes nearly all peak demands to nighttime.

Clusters 1 and 2 exhibited similar behavior for their consumption at the daily level in Figure 6.4. However, households in cluster 1 are characterized by an even higher amount of peak demands in the evening than those in cluster 2, translating to a higher than average Wasserstein-1 distance for cluster 1 at the daily level.

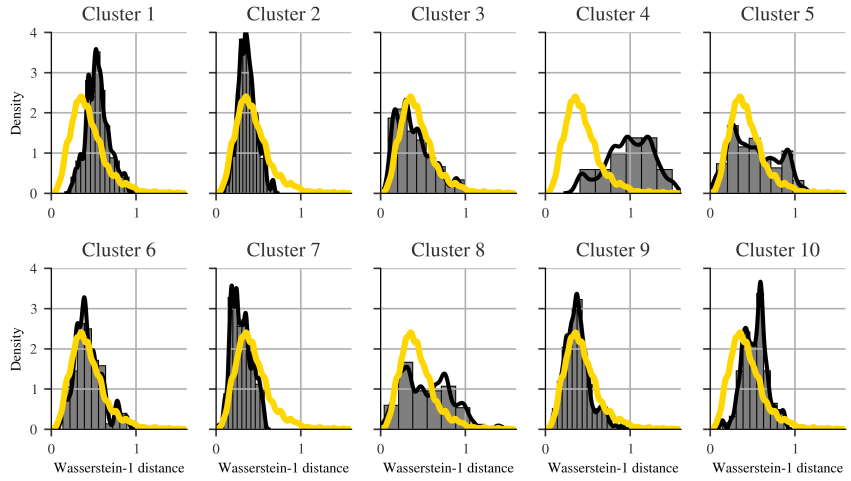

Figure 6.8: Histograms of the Wasserstein-1 distance between the distributions of the consumption and peak demands probability functions at the daily level. The black plot denotes the density for the individual cluster, while the plot in yellow indicates the density of the full dataset.

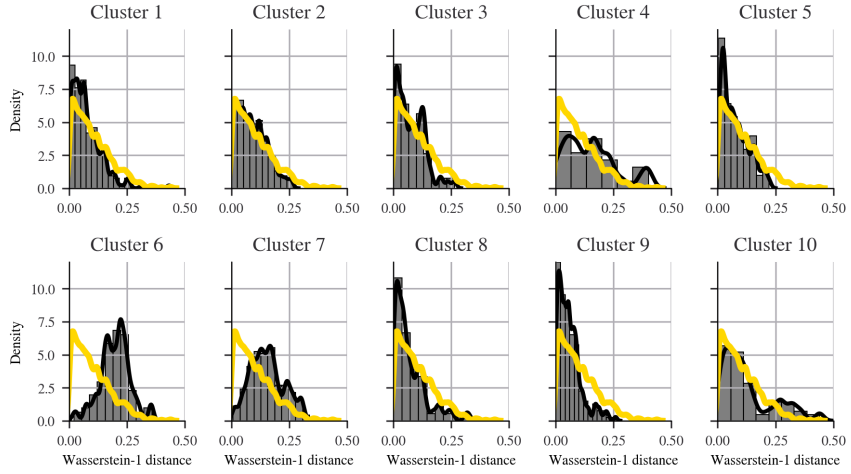

Figure 6.9: Histograms of the Wasserstein-1 distance between the distributions of the consumption and peak demands probability functions at the weekly level. The black plot denotes the density for the individual cluster, while the plot in yellow indicates the density of the full dataset.

This variability and disproportionate amount of peaks in a certain time interval offers insight in possibilities for targeted demand response initiatives or peak shaving via a residential energy storage system.

While cluster 6–7 and 8–9 have a similar consumption pattern, the time of occurrence of peak demands is significantly different, which leads to distinct solutions. As peak demands are typically generated by the simultaneous use of individual appliances, targeted demand response initiatives can be effective for cluster 6 and 7, where the majority of peaks occurs in the weekend.

Spreading the use of individual appliances over different days or being mindful of the simultaneous use in the weekend by inducing behavioral changes can reduce the number of peak demands. However, this requires a trigger for the behavioral changes and for these appliances to be available in different time periods. If this is not an option, investing in an energy storage system applying a peak shaving algorithm during weekends, while e.g,. maximizing the PV self-consumption during weekdays could offer an alternative, although the economic viability depends on the local tariff structure and the investment cost.

In contrast, cluster 8 is characterized by peak demands in the early morning and during the daytime, while households in cluster 9 exhibit peaks during the whole day. Consequently, for these households, a PV installation combined with a storage system can already offer a solution to reduce the demand from the grid, while maintaining a high self-consumption.

As a final check on the stochastic nature of peak demands, the relationship between the consumption in a time period and the presence of peak demands is investigated. Figure 6.10 displays the relations between the (untransformed) fractions of the consumption and peak demands at the daily level, with an ordinary least-squares (OLS) regression fit overlaid given the observed linear relation.

The coefficients obtained in the OLS regression for $f^p = a \times f^c + b$, with f^p and f^c the fraction of respectively the peak demands and consumption in that time period, are given in Table 6.4. As the presence of electric heating heavily skewed previous results for the consumption and peak demands at night, consumers with and without electric heating are treated separately for this analysis.

A correlation between the fraction of the consumption and that of the peak demands is present in Figure 6.10 and 6.11. As the presence of consumption in a certain time period is a prerequisite for a peak demand, some relation between the two types of parameters was expected.

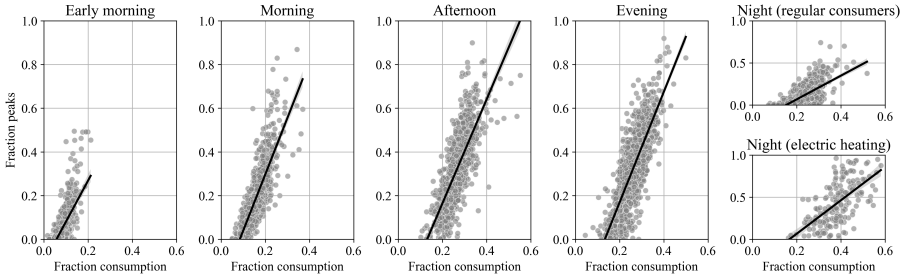

Figure 6.10: Relation between the fraction of consumption and peak demands in the time periods at the daily level, with an OLS regression estimate overlaid.

At first sight, the linear relation could be interpreted as an indication of predictability of peak demands in a certain time period. However, it is the spread on this relation that is the indicator of the stochasticity of the peak demands. For example, if 30% of a household's total consumption is observed occurring during the evenings, the results shown in Figure 6.10 suggests that 30–60% of the peak demands can occur in this same time period. This large uncertainty, which is present for each of the considered time periods, severely limits the usability of this linear relation, observed for the full dataset.

However, the knowledge of the introduced clusters can partly alleviate this uncertainty. This is illustrated in Figure 6.12 for clusters 1–3, which group households with a large fraction of their consumption during the evening, with a high number of peak demands simultaneously occurring in this time period. While we should be cautious drawing conclusions based on clusters that only include a limited amount of households, it appears that the spread on the fraction of peak demands for the individual clusters is smaller than those in Figure 6.10 for the full dataset, while the linear correlation that was observed before is nearly non-existent in some relations.

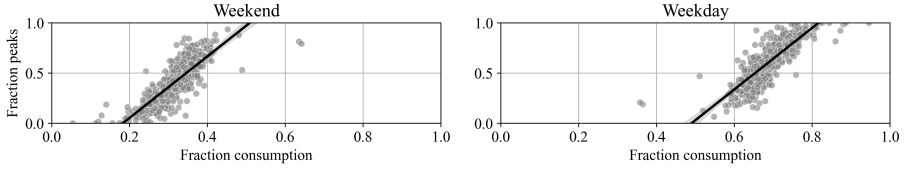

Figure 6.11: Relation between the fraction of consumption and peak demands in the time periods at the weekly level, with an OLS regression estimate overlaid.

Table 6.4: Coefficients of the OLS regressions shown in Figure 6.10 and 6.11 at the daily and weekly level.

Time period	a	b
Early morning	1.88	-0.11
Morning	2.58	-0.22
Afternoon	2.34	-0.31
Evening	2.53	-0.33
Night (regular consumer)	1.41	-0.21
Night (electric heating)	1.99	-0.33
Weekday	3.06	-1.50
Weekend	3.06	-0.56

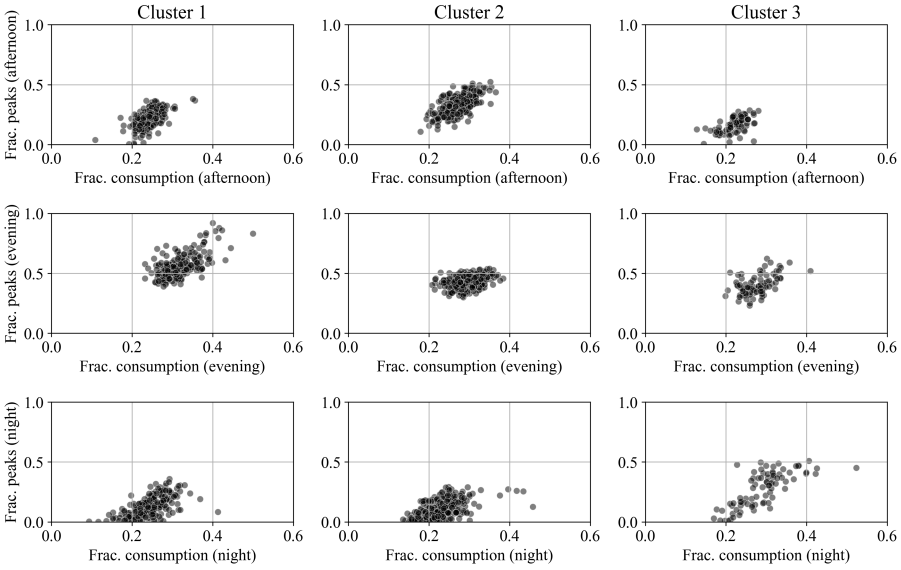

Figure 6.12: Relation between the fraction of consumption and peak demands in three time periods at the daily level for clusters 1–3.

6.6 Conclusion

In this chapter, we set out to cluster households with similar properties together. However, in contrast to traditional studies, a feature set was constructed that would be able to capture the behavior of both consumption and peak demands of individual households.

The primary methodological contribution of this chapter to consumer categorization through clustering is the introduction of a novel feature set designed to better capture temporal dependencies in energy consumption and peak demand patterns using annual data with 15-minute resolution. The analytic expression of the LDC, validated in Chapter 5, was used to identify peak demand values for each consumer. Two types of features were then constructed: (i) the fraction of consumption occurring during specific time periods at daily or weekly intervals, and (ii) the fraction of peak demands within those same periods. This feature set and clusters resulting from a clustering algorithm demonstrate clear relevance to demand response initiatives tailored to residential consumers. Additionally, to our knowledge, this is the first use of the Wasserstein-1 distance as a metric for temporal-based features in this context.

The proposed feature set was used in a hierarchical clustering process to build 10 clusters from the considered dataset of 1,402 consumers from a suburban region in Flanders. The clustering algorithm yielded compact clusters that showed a clear connection to real-life applications concerning the peak demands such as demand response initiatives, or the applicability of, e.g., battery storage systems for peak shaving purposes. Furthermore, differences in the behavior of the peak demands were found to be the main drivers of the clustering procedure. The presence of electric heating could be identified for several clusters, while others exhibited high daytime consumption during weekdays.

In the final analysis of this chapter, the stochastic nature of the peak demands was investigated by considering the relation between the consumption and the presence of peak demands in the same time period. The disproportionate presence of peak demands in a certain time period was quantified, and a linear relation was observed between the fraction of the consumption and peak demands in each time period. The spread on the results quantified the stochasticity of the peak demands, which limited the general applicability of the found relations. However, the obtained clusters showed a clear relation to the predictability and variability of the consumption and peak behavior, reducing the stochasticity of these peak demands and when they tend to occur.

The next two chapters focus on the generation of synthetic profiles. Given the limited number of consumers in some clusters, the 10 clusters themselves will not be used for the generation process from the very start in the following chapters. Rather, the metadata extracted in this chapter and the constructed clusters will be used throughout for additional illustrative validations, as well as to ease computational burdens.

7

Construction of daily load profiles via decomposition-recombination

If you want to find the secrets of the universe,
think in terms of energy, frequency and vibration.
Nikola Tesla

In this chapter, we present a methodology for the generation of daily load profiles rooted in time-frequency analysis. Time-frequency analysis allows for the decomposition of smart meter data into a generalized low-frequency component as well as a high-frequency stochastic component.

This chapter is structured as follows. First, a generic description of these times series is given in Section 7.1. We subsequently give some background information on time-frequency analysis, wavelets and the discrete wavelet transform in Sections 7.2 and 7.3.

Finally, we introduce a wavelet-based decomposition and recombination step for the synthetic data generation starting from Section 7.4 to Section 7.7. To wrap up this chapter, the limitations of the proposed methodology are discussed in Section 7.8.

Sections 7.4 to 7.8 of this chapter are derived from:

- R. Claeys, R. Cleenwerck, J. Knockaert, and J. Desmet, "Stochastic generation of residential load profiles with realistic variability based on wavelet-decomposed smart meter data", *Applied Energy*, vol. 350, p. 121750, 2023. [188]

7.1 Descriptions of daily load profiles

The daily load profiles under consideration are time series consisting of 96 measurements, one every 15 minutes. As mentioned in Section 3.3, we aim to deconstruct the privacy-sensitive daily load profiles into a generalized low-frequency (LF) approximation, and a high-frequency (HF) component that contains the stochastic consumption behavior.

Considering smart meter data on a daily level as a slowly changing base load, often linked to the occupancy of the dwelling, combined with a random perturbation is certainly not novel. For example, outside of the frequency-based decomposition studies that serve as the basis for this work, [189] described the daily load profile of a single household as a sum of two components:

$$x_n(t) = p_n(t) + e_n(t) \tag{7.1}$$

The authors considered $p_n(t)$ a daily routine specific to each energy consumer, while $e_n(t)$ was suggested to be a deviation from the underlying energy pattern. Drawing parallels to our approach, $p_n(t)$ is the LF approximation, while $e_n(t)$ contains the HF details.

Several techniques exist to decompose a signal into frequency components, of which the Fourier transform is the best known. The orthonormal basis of this Fourier decomposition is comprised of periodic sine and cosine functions with infinite support. Fourier established that any 2π-periodic function can be written as:

$$f(t) = a_0 + \sum_{k=1}^{+\infty} (a_k \cos(kt) + b_k \sin(kt)) \tag{7.2}$$

These Fourier coefficients can be used to characterize households. For example, [190] used the coefficients belonging to the 24h, 12h, and 8h cycles to group households based on their base load behavior. Similarly, [169] investigated whether the 24-hour or 12-hour frequency component was strongest to group households.

However, as residential daily load profiles are highly non-stationary due to the occurrence of HF stochastic peaks, the Fourier transform is inefficient to use. It would require a large number of harmonics to accurately describe the volatile peaks [191]. In contrast, wavelet-based decompositions are suitable to describe the considered load profiles. Wavelets are oscillatory functions with zero mean that have good time localization properties as they decay in a limited time window.

Heisenberg's uncertainty principle dictates that one cannot measure with arbitrarily high resolution in both time and frequency. Chronological data in a standard basis is completely localized in the time domain, but not at all in the frequency domain. In the case of Fourier bases we get exact localization in frequency, but none in time as the basis elements have infinite support. The localization properties of several decompositions are illustrated in Figure 7.1.

Using a wavelet basis, higher frequencies can be well localized in time, but the uncertainty in frequency localization increases as the frequency increases, which is reflected as taller, thinner cells with increase in frequency. Consequently, a wavelet basis and the discrete wavelet transformation (DWT) can be used to efficiently decompose a signal into a limited number of components, outperforming the Fourier transformation [191].

The following sections give some background about wavelets and how they can help make a multi-resolution analysis, before we delve into the decomposition-recombination scheme. We recommend readers familiar with the theory behind wavelets and multi-resolution analysis to skip the two following sections.

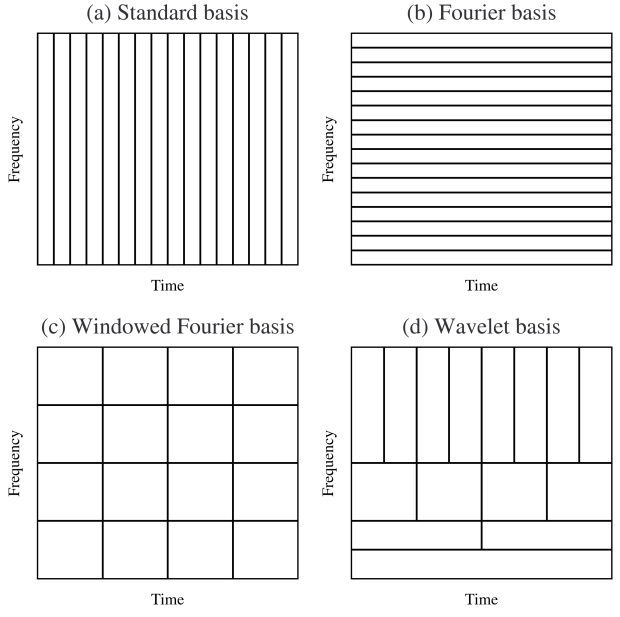

Figure 7.1: Schematic of time-frequency plane decomposition using different bases: (a) standard basis, (b) Fourier basis, (c) windowed Fourier basis, and (d) wavelet basis.

7.2 Introduction to wavelets

Some mathematics is needed to rigorously define what a wavelet is. As the name suggests, a wavelet is a 'small wave', which grows and decays in a limited time period. If we consider the traditional wave functions, the sine and cosine functions, it should be clear they are not wavelets. Indeed, these traditional wave functions oscillate up and down for all values of t if we were to plot $\sin(t)$ or $\cos(t)$ for $t \in]-\infty, +\infty[$.

In order to define a wavelet in the time domain, a real-valued function $\psi(t)$ has to satisfy two essential properties:

1. The integral of $\psi(t)$ is zero:

$$\int_{-\infty}^{+\infty} \psi(t)dt = 0 \tag{7.3}$$

2. The square of $\psi(t)$ integrates to unity:

$$\int_{-\infty}^{+\infty} \psi^2(t)dt = 1 \tag{7.4}$$

The first equation shows that any positive deviation from zero by $\psi(t)$ must be canceled out by negative deviations. Therefore, $\psi(t)$ must resemble a wave function. Furthermore, Equation 7.4 guarantees that the non-zero activity of our wave function $\psi(t)$ is limited to an interval of finite length.

In addition to the duo of equations mentioned above, which define the shape for a wavelet in the time domain, some research areas impose the so-called 'admissibility condition' to define wavelets. This condition guarantees the existence of the inversion formula for the continuous wavelet transform, and denotes any function $\psi(t) \in L^2(\mathbb{R})$ a wavelet if it satisfies the following condition

$$\int_{-\infty}^{+\infty} \frac{|\hat{\psi}(\xi)|^2}{|\xi|} d\xi < \infty, \tag{7.5}$$

with $\hat{\psi}(\xi)$ the Fourier transform of $\psi(t)$.

Figure 7.2 on the next page displays three wavelets that satisfy the aforementioned conditions as illustration. It is clear that the wavelets are (i) localized in time, and (ii) show oscillating behavior.

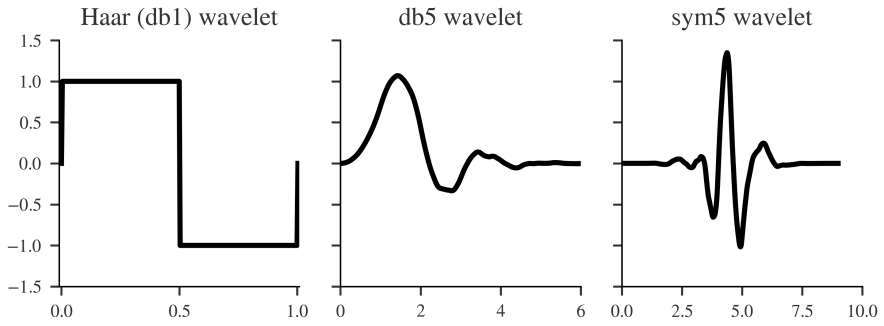

Figure 7.2: Three functions which are classified as wavelets.

The Haar wavelet is the first known wavelet, reported in 1910 by Hungarian mathematician Alfred Haar [192]. It is an odd discontinuous rectangular pulse, and is defined as follows:

$$\psi^H(t) = \begin{cases} 1, & 0 \leq t < 1/2 \\ -1, & 1/2 \leq t < 1 \\ 0, & \text{otherwise} \end{cases} \tag{7.6}$$

Due to its simplicity, the Haar wavelet is often used for educational and testing purposes. For this dissertation, we will also initially only consider the Haar wavelet, while extensions to other wavelets will be concisely discussed.

Individual wavelet functions have the intriguing property that, starting from a single wavelet, a whole family of functions can be constructed which all satisfy the wavelet requirements. Consider a generic wavelet function $\psi(t) \in L^2(\mathbb{R})$, which can subsequently be translated and dilated. This is illustrated in Equation (7.7), with a as scaling parameter and b the translation parameter.

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \psi\left(\frac{t-b}{a}\right) \tag{7.7}$$

In the context of wavelet families, the construction of wavelets from translations and dilations of a single function $\psi(t)$, this initial wavelet is called the *mother wavelet*.

The scaling parameter a is a measure for the degree of compression, whereas the translation parameter b specifies the time location of the wavelet. If $|a| < 1$, the resulting wavelet is a compressed version of the mother wavelet as it has a smaller support in the time domain. Consequently, these wavelets will correspond more with high frequency components of the signal.

Varying the scaling parameter therefore allows wavelets to describe both high and low frequencies localized around the center $t = b$. Analogous to the Fourier transform, the question now arises whether an orthonormal basis of wavelets derived from a mother wavelet can be found to decompose an arbitrary time signal into wavelet-based components with different frequency ranges. This so-called multi-resolution analysis (MRA) is examined in-depth in the following section.

7.3 Multi-resolution analysis through wavelets

The previous section showed that wavelets are suitable to describe functions at different resolutions. By compressing or stretching a single mother wavelet, high or low frequencies can respectively be captured. The matter that now presents itself is how wavelets can be used to decompose an arbitrary time signal into low-frequency and high-frequency components.

Intuitively and operationally, the decomposition of an arbitrary function $f(t)$ into high and low frequency components is illustrated in Figure 7.3. The considered signal is passed through a high-pass and a low-pass filter respectively.

Hence, the original signal is decomposed in a LF approximation A_1 and a HF detail D_1. Each decomposition step downsamples the signal by a factor of 2. A k-level decomposition of a signal $f(t)$ therefore decomposes the signal as follows:

$$f(t) = D_1 + ... + D_k + A_k = \sum_{i=1}^{k} D_i + A_k \tag{7.8}$$

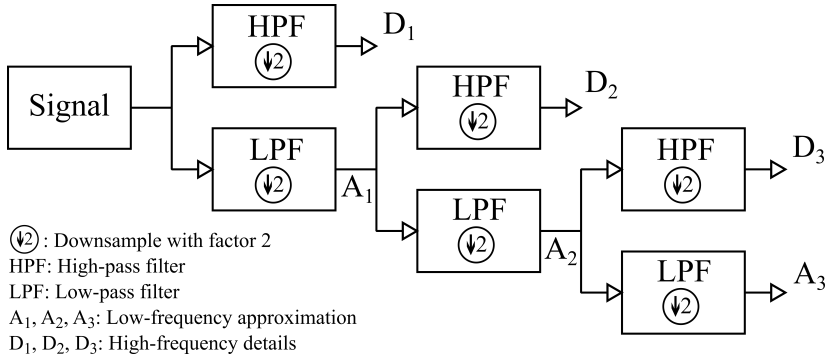

Figure 7.3: Illustration of a three-level multi-resolution analysis using the discrete wavelet transformation.

The link between both the low-pass and high-pass filters and the wavelets introduced in the previous section is not immediately clear. However, the mathematical framework presented by Mallat and Meyer sheds light on this relationship [193]. We summarize their arguments in the following paragraphs, while referring to the original work for a robust proof.

Firstly, an MRA is nothing more than a hierarchical framework which approximates the original function $f(t)$ at multiple resolutions by orthogonal projections onto a family of spaces $\{\mathbf{V}_j\}_{j\in\mathbb{Z}}$. This is illustrated in Figure 7.4 where projections of a function f are shown for different subspaces $\{\mathbf{V}_j\}$. As seen, the resolution of subspace $\{\mathbf{V}_j\}_{j\in\mathbb{Z}}$ is taken to be 2^{-j}. Therefore, a common notation of the subspaces is $\{\mathbf{V}_{2^j}\}_{j\in\mathbb{Z}}$ or $\{\mathbf{V}_{2^{-j}}\}_{j\in\mathbb{Z}}$, depending on convention.

Let $\{\mathbf{V}_{2^j}\}_{j\in\mathbb{Z}}$ now be a multi-resolution analysis of $L^2(\mathbb{R})$. It can be proven that a function $\phi(t) \in L^2(\mathbb{R})$ exists such that $\phi(t)$ satisfies the dilation equation $\phi_{2^j}(t) = 2^j\phi(2^j t)$. The set of dilated and translated functions in Equation (7.9) then forms an orthonormal basis of $\mathbf{V}_{2^j}$:

$$\left(\sqrt{2^{-j}}\ \phi_{2^j}(t - 2^{-j}n)\right)_{n,j\in\mathbb{Z}^2} \tag{7.9}$$

This function $\phi(t)$ is called the scaling function, also known as the father wavelet in the context of MRA.

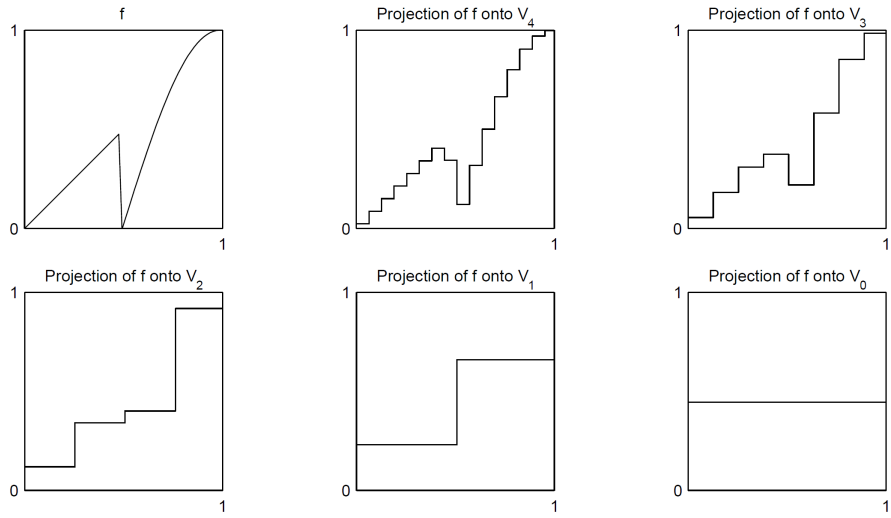

Figure 7.4: Projections of a function f on different subspaces $\{\mathbf{V}_j\}$, formed by the Haar multiresolution analysis. [194]

The approximation of the function $f(t)$ at the resolution 2^j, also denoted as $A_{2^j} f$ can now be computed by decomposing the signal on this orthonormal basis, which can be shown to be equivalent to the presence of a low-pass filter:

$$A_{2^j} f = \left(\langle f(t), \phi_{2^j}(t - 2^{-j} n) \rangle \right)_{n \in \mathbb{N}} \tag{7.10}$$

However, if a signal is approximated at the resolution 2^{j+1} and 2^j, there is a difference of information between these two scales, which restricts the ability to reconstruct the original signal. This difference is called the detail function. Consequently, it is necessary to also determine the detail function to unambiguously reconstruct the original signal. This gives rise to a description equivalent to the high-pass filter included in Figure 7.3.

The approximations at resolution 2^j and 2^{j+1} are equal to the orthogonal projections of the original signal on $\mathbf{V}_{2^j}$ and $\mathbf{V}_{2^{j+1}}$. The detail function at resolution 2^j is then given by the orthogonal projection of the original signal on the orthogonal complement of $\mathbf{V}_{2^j}$ in $\mathbf{V}_{2^{j+1}}$. This orthogonal complement is denoted as $\mathbf{O}_{2^j}$.

It is possible to now construct an orthonormal basis for $\mathbf{O}_{2^j}$ consisting of dilated and translated wavelets as introduced in Equation (7.7). Similar to the set defined in Equation (7.9), a binary scaling and dyadic translation is chosen to obtain such an orthonormal basis:

$$\left(\sqrt{2^{-j}}\, \psi_{2^j}(t - 2^{-j} n) \right)_{n,j \in \mathbb{Z}^2} \tag{7.11}$$

The detail $D_{2^j} f$ of the function $f(t)$ at resolution 2^j is similarly determined by the following set of inner products with the wavelet basis:

$$D_{2^j} f = \left(\langle f(t), \psi_{2^j}(t - 2^{-j} n) \rangle \right)_{n \in \mathbb{N}} \tag{7.12}$$

Consequently, it is clear that applying a discrete wavelet transform (DWT) is equivalent to passing the signal through a band-pass filter bank, either a high-pass or low-pass filter. Passing the signal through a high-pass filter yields the oscillating HF component, while the low-pass filter yields an approximated signal at lower resolution.

In this work, the Haar wavelet as given in Equation (7.6) is initially used as the mother wavelet $\psi(t)$. It is crucial to bear in mind that for this case, the sum of the high-frequency detail signal is equal to zero. Consequently, the daily energy of the LF approximation is equal to that of the original smart meter data. This property will be vital for the remainder of this chapter and the proposed methodology.

Following the reasoning introduced in [191], the discontinuous nature of Haar's wavelet is suitable to describe the load profile of individual consumers, capturing the behavior of turning appliances on and off. Furthermore, the considered metering infrastructure logs averaged power demand with a resolution of 15 minutes, likewise leading to a discontinuous profile. Additionally, the scaling function for the LF approximation is given in (7.13).

$$\phi(t) = \begin{cases} 1, & 0 \leq t \leq 1 \\ 0, & \text{Otherwise} \end{cases} \tag{7.13}$$

7.4 Wavelet-based decomposition

Using the framework from the previous section, we can now decompose daily load profiles into a LF and HF component. Figure 7.5 illustrates how a three-level MRA on a residential load profile is performed using the Haar wavelet. The original load profile is measured at a 15-minute resolution. The MRA introduces subspaces with a 30-minute, 1-hour, and 2-hour resolution to approximate the signal A_1, A_2 and A_3 respectively. The HF details are subsequently captured in the signals D_1, D_2 and D_3. The nested subspaces of the MRA which translate to the downsampling by a factor 2 for each decomposition step can be observed.

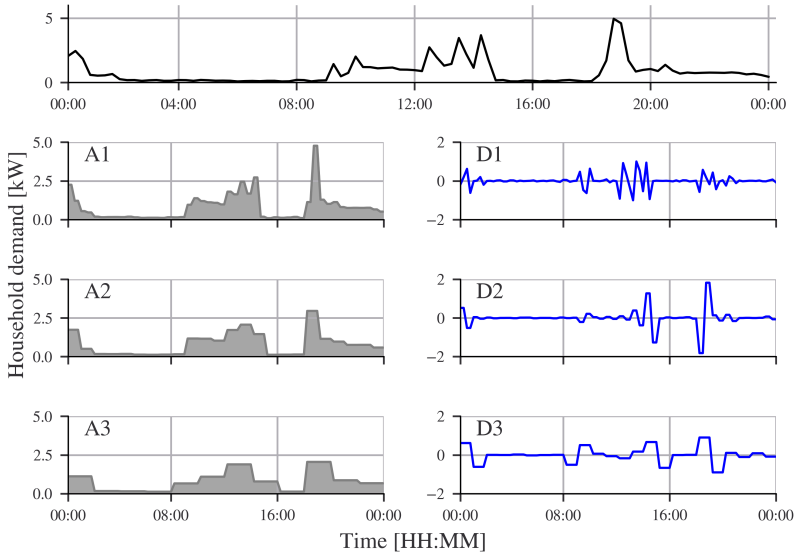

Figure 7.5: Application of a three-level DWT MRA decomposition.

Figure 7.5 additionally illustrates the main difficulty of constructing synthetic load profiles for residential consumers. The low-frequency approximations fail to adequately capture the peak demands at increasing levels of decompositions. The low-frequency approximation A_2 only captures 65% of the original daily peak demand.

This observation for a single day for a single consumer can be extended to the considered dataset of residential load profiles. Figure 7.6 visualizes the distributions of the annual peak demand as well as the mean monthly peak demand, compared to the original load profiles. Similar results can be observed: a three-level approximation tends to underestimate the annual and mean monthly peak demand by a factor 2.

This significant contribution of the HF detail to the peak demand of residential consumers limits the possibility of using previously reported methodologies for superimposing extracted HF variabilities onto a LF approximation or modeled load without any modification. Previous works using DWT MRA for load modeling either studied non-residential consumers [195, 196], or aggregated load profiles at feeder-level [197].

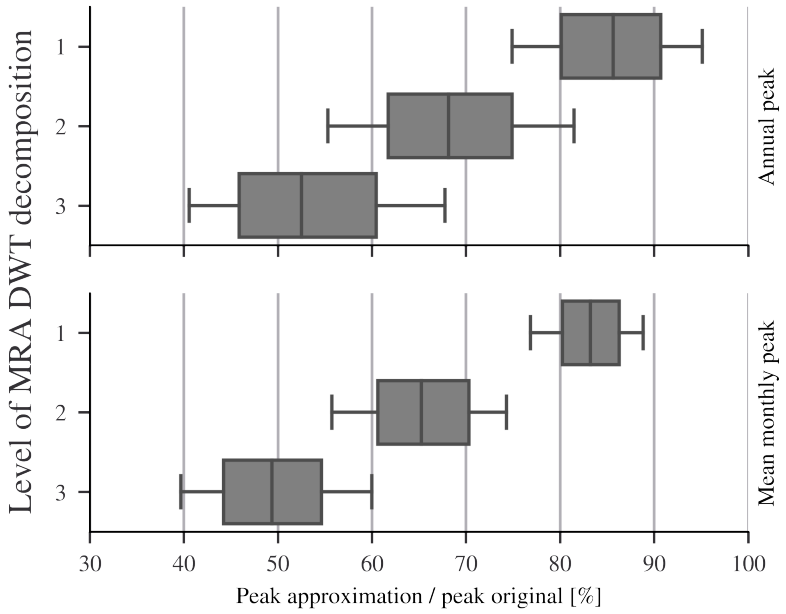

Figure 7.6: Boxplot visualization of the distributions of the annual (top) and mean monthly (bottom) peak demands of the k-level low-frequency approximation compared to the original load profiles. The whiskers of the boxplot denote the [10, 90] percentiles.

7.5 Recombining with superimposed variability

Our aim is to construct synthetic load profiles with a given daily consumption, corresponding peak demand and accompanying variability in the load profile. The question now arises: given a daily consumption value and a maximum value, how can a synthetic profile with realistic variability be constructed? Inspired by [195] and [196], a component based on the HF details will be added to a LF approximation to construct a synthetic profile.

However, in both [195] and [196] the extracted HF detail of the daily profile was each time normalized with respect to the peak daily value of the LF approximation. This normalized profile was subsequently rescaled before superimposing it onto a modeled load profile. While this was valid for non-residential consumers, this normalization scheme cannot be used for residential consumers where the HF component significantly contributes to the peak demand. Consequently, a different approach has to be considered for the rescaling of the HF component.

As opposed to the aforementioned normalization scheme, this framework starts from a daily peak demand $P^d_{\max}$, either measured or modeled. The research question then reduces to determining whether a combination of a LF approximation of consumer i and a rescaled detail of consumer j can consistently lead to a load profile with the predetermined peak value.

By construction, the daily consumption of the LF approximation A_j of the original daily profile is equal to the daily consumption of the original daily load profile, as the sum of the oscillating detail functions is equal to 0. Hence, it is possible to scale and translate a detail function before superimposing it on an approximated load while, if desired, keeping the inherent correlation between the daily consumption and the daily peak demand intact.

Let $A^{i,d}_n$ now denote the LF approximated load of consumer i on day d at the n-th 15-minute interval of the day while $\sum D^{j,d}_n$ is the sum of the HF detail values of consumer j on the same day at that 15-minute interval. Keeping the LF approximation of consumer i fixed allows the unambiguous determination of the scaling factor β to rescale the HF detail of consumer j.

$$P^d_{\max} = \max_{1\leq n\leq 96} A^{i,d}_n + \beta \max_{1\leq n\leq 96} \sum D^{j,d}_n \tag{7.14}$$

The daily HF detail of consumer j can now correctly be rescaled to yield the desired peak demand in the combined load profile. However, there is no guarantee that the maximum values of the approximation signal and the rescaled detail function take place at the same time of the day.

The time period with the maxima of $A_n^{i,d}$ tends to contain the time of the daily peak demand of consumer i. Adding a HF profile with its maximum during this time block thus ensures consistent timing of the daily peak demand between consumer i and the synthetic profile. Therefore, a circular shift is performed on the HF profile of consumer j in order to align its maximum value with the peak values of the approximation function. Analytically this shift can be represented by the operator σ.

Given an obtained HF detail profile D as a function of time measured at T timesteps, let D_k denote the load at timestep k. For the considered dataset, the detail function is an array of 96 values: $D = [D_1, D_2, \ldots D_{96}]$. A circular shift is now defined as follows, where the operator σ_κ denotes a shift of κ steps.

$$\sigma_\kappa(k) = k - \kappa \pmod{96}, \forall k \in \{1, 2, \ldots 96\} \tag{7.15}$$

For example, $\sigma_4 D_k$ becomes $[D_{93}, D_{94}, \ldots D_{92}]$. The circularly shifted HF signal can now be superimposed on the LF approximation of a different consumer, once the maximum of both coincide in time. The choice for a shifted HF detail is logical from the point of view of load modeling, as it does not matter whether the variability from stochastic peaks occurs at e.g., 6 PM or at 9 PM, as long as it is realistic.

This framework is illustrated for the construction of a single load profile in Figure 7.7, where the stochastic peaks of household j during the evening are rescaled and shifted to the morning to construct a profile with a daily maximum demand equal to that of household i.

The choice of only considering HF profiles of the same day d by including a circular shift is somewhat arbitrary. An alternative would be to create a taxonomy of daily variability profiles and making abstraction of the day d it corresponds with, and it is no longer required to superimpose a detail function of the same day d on a LF approximation. Similar considerations were made in [195] to model on the day-by-day or year-by-year approach. Given the interest in generating daily load profiles in this work, the circular shift approach was introduced to generate viable synthetic profiles for each day d while only considering inputs from the same day.

Two additional aspects of this framework need to be discussed as they have a computational impact. First, the HF detail has negative values as an oscillating function. Consequently, it is possible that combinations of rescaled HF and LF functions will have negative load values. In the absence of local generation, this is a non-viable solution. Accordingly, only synthetic profiles that have non-negative loads will be considered. Should this occur, a different HF detail and/or LF approximation need to be selected. The impact of this constraint is discussed in-depth in Section 7.8.

Second, the LF approximations of the considered load profiles have a resolution of 2 hours, and as such they exhibit 8 identical maximum values. Consequently, there are 8 possible solutions for the operator σ to align the maximum values of the LF and the HF arrays. Therefore, the obtained synthetic profile will not be a unique combination of the two arrays. If a single profile needs to be obtained, we can randomly select one of the viable solutions.

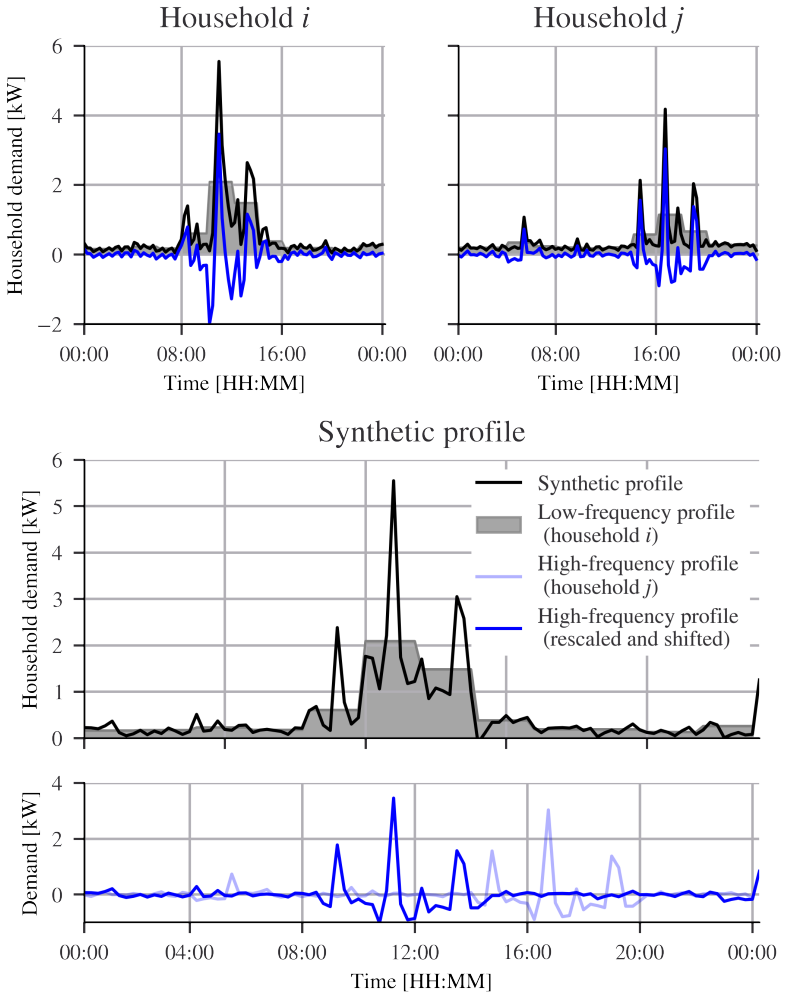

Figure 7.7: Illustration of the proposed framework. The LF component (grey patch) of household i is combined with the rescaled and shifted HF component (blue) of household j to obtain a synthetic profile (black) with daily peak demand equal to household i.

7.6 Stochastic load profile generator

This methodology for superimposing residential variability can now be used to construct a larger dataset of synthetic profiles. As suggested in Section 7.1, one of the main shortfalls of traditional load modeling of residential consumers is related to the second step. This step entails the generation of typical load curves for different types of consumers by using statistical measures.

We instead propose a two-step method that avoids the usage of the mean or median value, and subsequent smoothing process. We aim to approximate the original dataset in a distributional sense. Thus, the intended output is a dataset that approaches the real distribution of daily consumptions and peak demands in the original dataset for each individual day. The input is a dataset (or clustered subset) of daily load profiles for a given day d.

In the first step, the distribution is defined that will be approximated and sampled. This distribution is estimated based on a dataset of daily load profiles. A bivariate distribution is selected, as this offers a twofold advantage. First, the inherent correlation between the consumption and the peak demand can be included. Second, sampling a 2D-distribution yields two parameters, allowing for the rescaling of both the HF and LF profile.

The first parameter is the daily consumption, as the LF approximation can be rescaled to this value. While it is not strictly necessary to rescale the approximation function, it does allow for the introduction of a continuous spectrum of values for the daily consumption, instead of limiting the constructed profiles to the discrete values in the historical dataset. Furthermore, by taking the approximation profile closest to the sampled value, the realism of the obtained profile is assured. For example, a profile with 4 kWh daily consumption will not be upscaled to an unrealistic 30 kWh profile.

Second, a metric needs to be chosen to rescale the HF component. Intuitively, it seems logical to use the daily peak demand of the original data as metric to rescale the HF detail. However, computational concerns have lead to the choice of the daily load factor as second metric. The daily load factor of consumer i on day d is denoted as $L_F^{i,d}$, and is given by the ratio of its mean daily load to its maximum load on day d. Let $P_k^{i,d}$ denote the measured load of consumer i on day d at the k-th 15-minute interval of the day. The daily load factor is then given by:

$$L_F^{i,d} = \frac{\frac{1}{96}\sum_{k=1}^{96} P_k^{i,d}}{\max\limits_{1\leq k\leq 96} P_k^{i,d}} \tag{7.16}$$

The load factor for residential consumers was introduced in [90] as a way to describe the 'peakyness' of their consumption behavior. A high L_F corresponds to demand that is distributed evenly throughout the day, while a low L_F is indicative for intervals of high demand compared to its base load. From Equation 7.16, the daily peak demand can be calculated if both the daily load factor and the daily consumption are known.

Figure 7.8 illustratively displays the probability distributions of the daily load factor and the peak demands for the full dataset for a single winter day. Both parameters are non-normally distributed and display a long tail. A continuous PDF can be estimated through kernel density estimation (KDE) using a Gaussian kernel. However, the main drawback of this KDE approach occurs for densities exhibiting long tails, as the KDE technique tends to oversmooth the longer tails [198].

Consequently, the load factor is preferable over the peak demand for this approach, as stochastic peak demands occur for lower L_F in the bulk of the distribution, whereas they are located in the long tail of the peak demand distributions and their density will be underestimated if this parameter were to be used.

Furthermore, the load factor was used in [199] as a metric for the comparison between a bottom-up load model and a historical dataset. They concluded that the synthetic data underestimated the seasonal effect in both the load factor and peak demand behavior, and that '*limited treatment of seasonal variation in load modeling can lead to inaccurate predictions of its effects*'.

Therefore, by limiting the construction of the daily load profiles to each separate day d based on its 2D KDE, the seasonality of the dataset is preserved. As displayed in Figure 7.9, the KDE shows significant differences depending on the season, despite showing similar values of maximum probability.

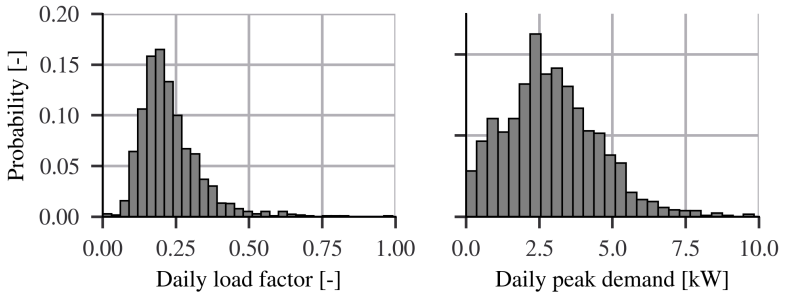

Figure 7.8: Distribution of the load factor (left) and the peak demand (right) for the full dataset on a winter day (December 20).

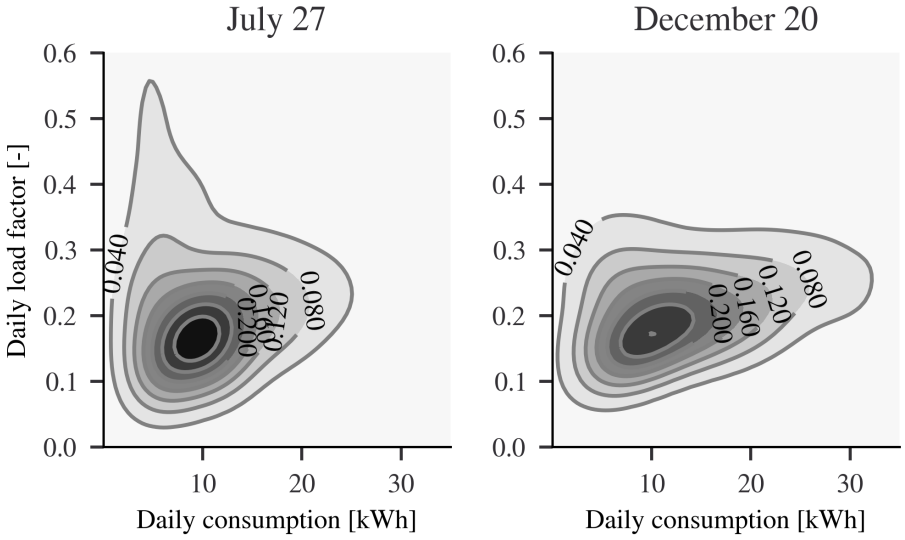

Figure 7.9: Example of the 2D kernel density estimations of the daily consumption (x-axis) and the daily load factor (y-axis) for a summer (left) and a winter day (right).

The left-hand side of Figure 7.9 displays a summer day, while the right-hand side shows a winter day. The KDE of the summer day exhibits a long tail towards high load factors at lower consumption that is absent during winter days, which indicates the presence of vacant households where families are on holiday.

The block diagram to construct a single stochastic daily load profile of day d is given in Figure 7.10. This method can be repeated as many times as desired to obtain a dataset of daily load profiles of a predetermined size. In summary, based on the measured load profiles, a 2-dimensional probability density function (PDF) is estimated between the daily load factor and the daily consumption. The sampled consumption is used to rescale a LF profile of consumer i, while the load factor determines how the HF profile of consumer j has to be reshaped and shifted.

For the remainder of this work, the proposed decomposition - recombination methodology is applied to the full dataset of consumers unless otherwise stated. However, an identical approach can be used if only a subset of the complete dataset is considered.

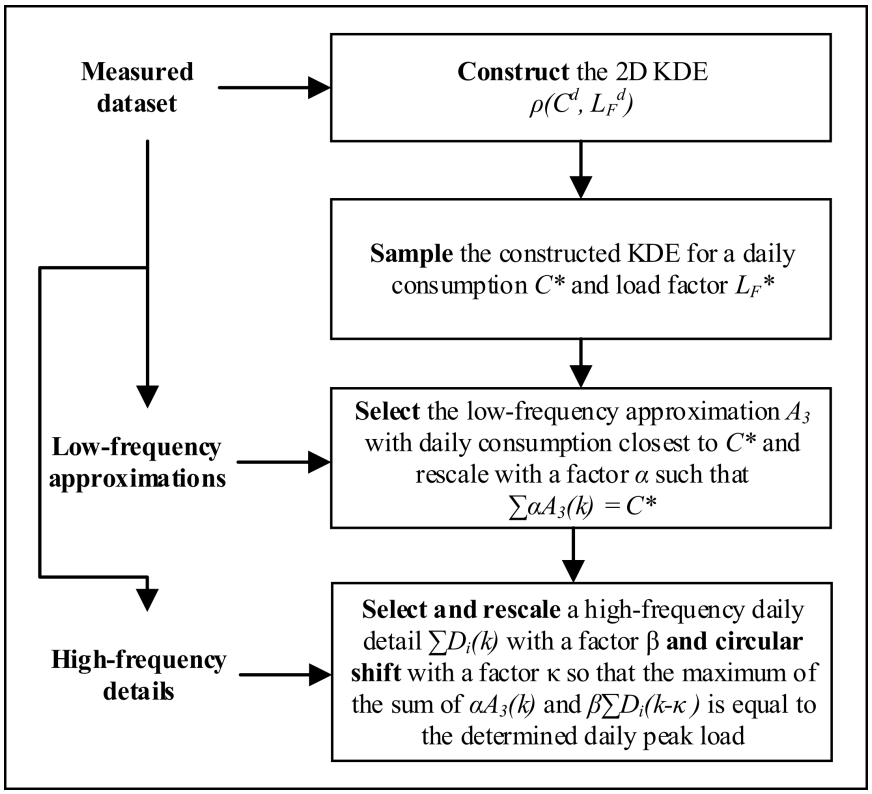

Figure 7.10: Block diagram of the proposed framework to generate a single daily load profile for day d.

For example, a possible refinement of the proposed decomposition - recombination scheme is using a preprocessed dataset as input, with labels corresponding to clusters obtained through unsupervised learning on features or metadata, as was done in Chapter 6.

By limiting both the 2D KDE construction and sampling to the L_F and daily consumption to elements contained in the cluster, and constricting the sampling of both the LF approximation and HF details from elements within that cluster, more targeted output can be obtained.

When looking at the 2D KDE of individual clusters compared to the full dataset, significant differences can be observed. We reconstruct Figure 7.9 for two individual clusters. Clusters 4 and 10 from Chapter 6 are compared in Figure 7.11. As a reminder, cluster 4 contains households with a significant amount of their consumption and peak demands at night, whereas cluster 10 shows a large number of peaks during mornings and afternoon.

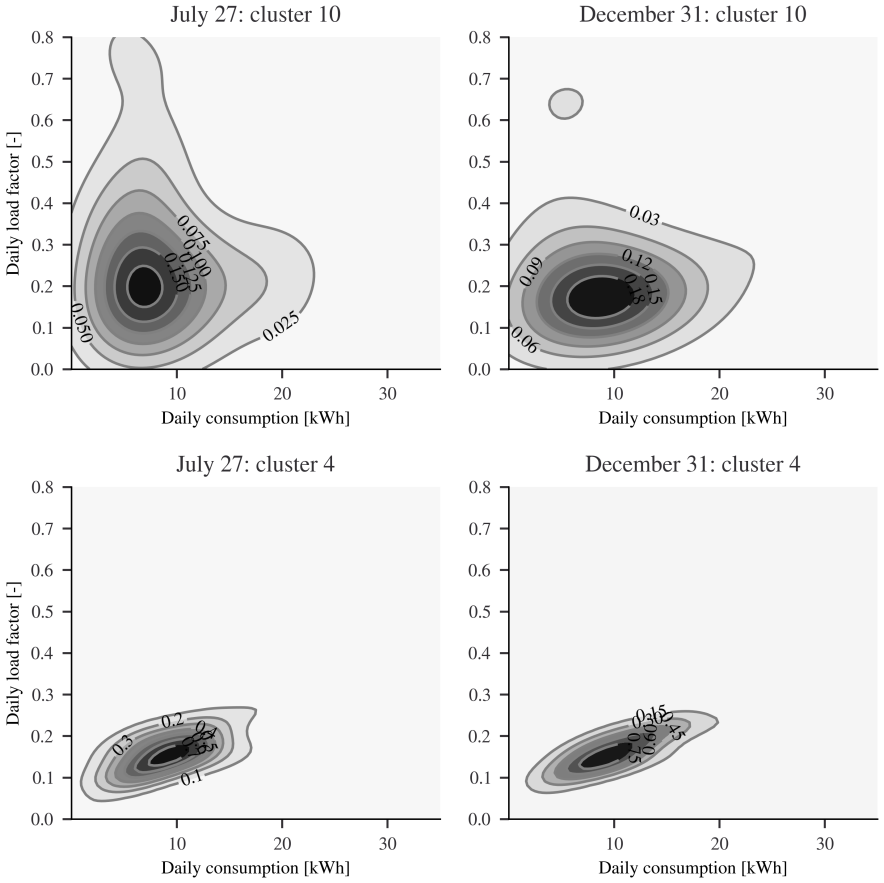

Figure 7.11: 2D KDE plots for a summer (left) and a winter day (right), for two distinct clusters determined in Chapter 6: clusters 10 (top) and 4 (bottom).

While cluster 4 is a very compact clusters in the 2D plane, cluster 10 shows a large variation in the daily load factor. Consequently, the 2D KDE approach when restricted to individual clusters can lead to more targeted output that is similar to elements within the cluster, as long as relevant metadata was used to construct coherent clusters.

7.7 Results

The output of the stochastic load profile generator are as much individual daily load profiles as desired for a given day d. Eight randomly generated profiles are shown in Figure 7.12 for an initial visual inspection. The LF approximations are once again shaded in gray, while the superposition of the LF and HF profiles is displayed in blue.

The majority of the generated profiles qualitatively behave as expected, with the HF component significantly contributing to the daily peak demand, featuring the stochastic behavior that is often associated with residential consumers, with morning, noon or evening peaks clearly visible in the HF component of the consumption profiles. By construction, the HF component is able to capture a large fraction of the daily peak demand for the highly stochastic profiles.

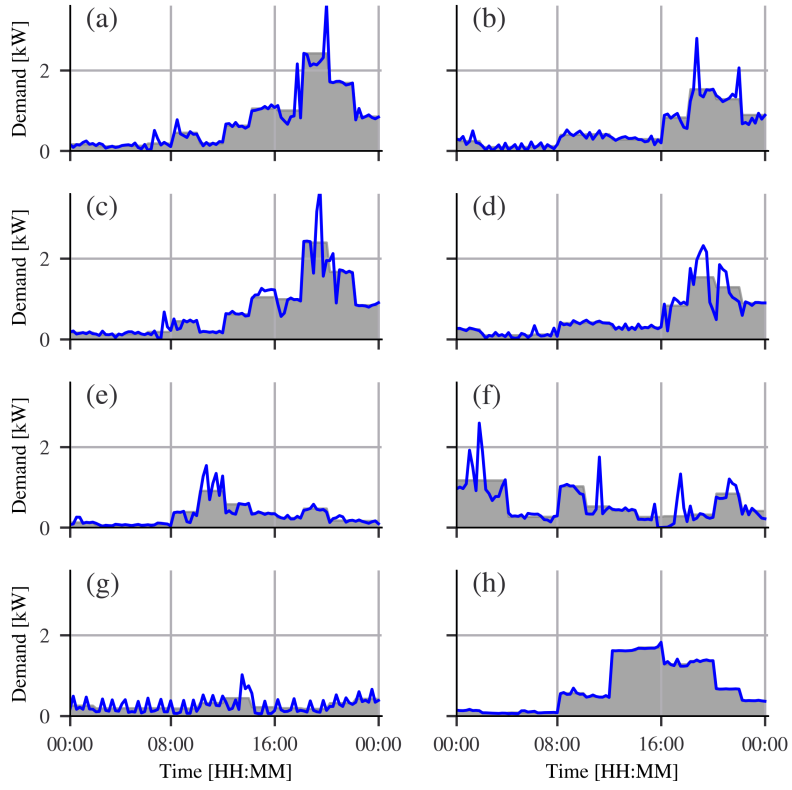

Figure 7.12: Illustrative examples of eight daily load profiles generated by the stochastic load generator.

However, the introduced methodology exhibits drawbacks which can lead to unrealistic profiles for two specific cases, more precisely cases (g) and (h) included in Figure 7.12.

First, subplot (g) displays a vacant household with a low base load. However, the inclusion of the HF details can subsequently lead to an unrealistic cyclic behavior on top of the small base load for these households.

Second, profile (h) displays a profile where the LF approximation already captures more than 95% of the daily peak demand. Therefore, the addition of a HF detail has very little impact on the overall shape of the resulting profile, and the synthetic profile displays a stable consumption for every block of two hours. This is a result of the choice of wavelet and decomposition level. A lower decomposition level or a more variable wavelet for the DWT-MRA would lead to a less constant consumption profile for this case.

In order to investigate whether the implemented methodology behaves as expected at the level of the population, 200 profiles were generated using a subset of households with annual consumption between 2,750 and 3,750 kWh. This consumption range is centered around the mode of the distribution of the yearly consumption and is representative for typical Flemish households without electric heating [143].

The distributions of the daily peak demands throughout the year are visualized in Figure 7.13 for the original dataset, as well as for the synthetic dataset. Both the mean value of each day, as well as the 25-75 percentiles for the original and the synthetic dataset are displayed.

In order to evaluate the performance of the synthetic dataset, the time series of the mean daily peak demand is compared with that of the original dataset. Table 7.1 displays three metrics. The mean absolute error (MAE) is 0.09 kW, while the mean percentage error (MPE) is -2.9%. The proposed methodology yields the expected results, with only a slight deviation from the original dataset.

However, as visible in Figure 7.13 and the negative MPE, the synthetic model leads to a slight underestimation of the mean daily peak demand. This underestimation can be attributed to original profiles with high daily peak demand. These profiles are harder to reconstruct, which is discussed in the following section on limitations of the proposed methodology.

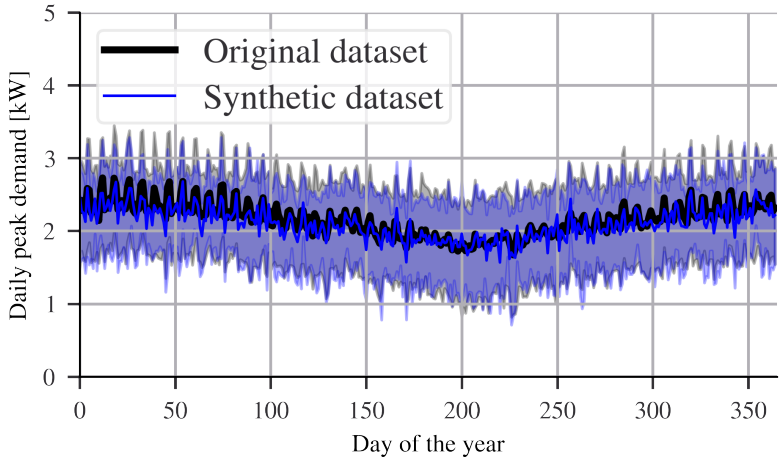

Figure 7.13: Daily peak demands of the households with annual consumption between 2,750 kWh and 3,750 kWh contained in (black) the original dataset, and (blue) a constructed synthetic dataset. The full line denotes the mean value of the distribution for each day, while the shaded band displays the 25-75 percentiles.

Table 7.1: Performance evaluation of the synthetic data in Figure 7.13.

Mean Absolute Error	MAE	0.09 kW
Mean Absolute Percentage Error	MAPE	4.1%
Mean Percentage Error	MPE	-2.9%

7.8 Limitations

Two limitations of the daily load profile generator need to be addressed.

First, the generator hinges on combining LF and HF profiles of different households. However, as discussed in Section 7.6, not all combinations under the proposed methodology yield valid synthetic profiles. As the HF profile has negative values, the rescaling of this component can lead to negative values in the synthetic profile, which is a non-viable solution.

In order to quantify the number of viable combinations, the original daily load profiles in the historical dataset are considered. For each household i, the daily load profile on day d is decomposed in a LF approximation and HF detail function for a given decomposition level k.

The detail functions of each other household j are rescaled and shifted according to Section 7.5, and subsequently superimposed on the approximation of household i to obtain synthetic profiles with the same daily consumption and peak demand. The percentage of other households j that lead to synthetic profiles without negative values was determined for each household i and day d.

The results are visualized on Figure 7.14. The mean of the distribution for each day d is displayed together with the 25-75 percentile band.

On average, 25–30% of the households in the dataset can be used to reconstruct a synthetic load profile according to the aforementioned boundary conditions if the DWT decomposition level is 3, while this is nearly 45% for decomposition level 1. A seasonal effect is visible, which can be attributed to more households being vacant during the holiday period and thus less stochastic.

While Figure 7.14 is informative on the general behavior of the number of valid combinations, the variance on the displayed distributions is significant, as seen from the large interquartile range. Consequently, it is necessary to investigate which profiles have a lower number of valid combinations and whether this has an impact on the proposed methodology.

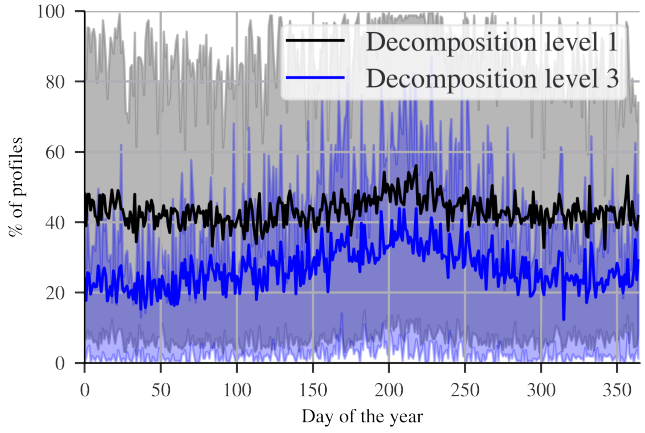

Figure 7.14: Distribution of the proportion of households that lead to viable synthetic combinations for each day d and different decomposition levels. The full line denotes the mean value of the distribution for each day, while the shaded band displays the 25-75 percentiles.

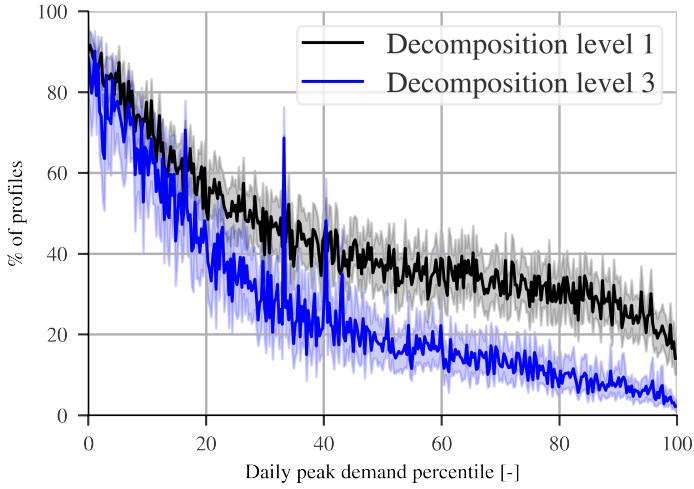

Figure 7.15: Distribution of the proportion of households that lead to viable synthetic combinations for each day d and different decomposition levels as a function of the daily peak demand percentile. The full line denotes the mean value of the distribution, while the shaded band displays the 25-75 percentiles.

This is done in Figure 7.15. Therein, the percentage of profiles that lead to a valid synthetic profile is displayed as a function of the percentile of the daily peak demand of each individual consumer. The maximum daily peak demand of each consumer corresponds to 100, while the minimum daily peak demand corresponds to 0.

As the daily peak demand percentile increases, fewer profiles tend to lead to valid combinations. This is consistent with the assumption that higher peak demands are the result of more stochastic behavior, and thus need to be captured by a larger HF detail. The daily load profile of each household that exhibits the maximum peak demand can only be reconstructed by 15% of the other profiles for a decomposition level 1, while this drops to 4% for a decomposition level 3.

Figures 7.14 and 7.15 highlight the trade-off between anonymization and retaining sufficient profiles to be able to form enough valid combinations. The higher the decomposition level of the MRA-DWT, the more anonymized the coarse LF approximations will be. However, the price for this increased anonymization is a diminishing number of HF details of other households that can be combined with this approximation to yield valid profiles.

However, not only the decomposition level has an impact on this anonymization process. Note that the displayed results up to now were obtained based on the Haar wavelet. In [195], the choice of wavelet in the DWT-MRA decomposition was investigated for 34 different wavelets. Therein it was found that the Haar wavelet, there denoted as the Daubechies 1 (db1) wavelet, tends to lead to a decomposition with the highest variability in the HF profile.

To illustrate this trade-off, Figure 7.16 recapitulates the findings from Figure 7.6, but this time for three different types of wavelets: the Haar, sym5 and db5 wavelets. These wavelets have previously been visualized in Figure 7.2. According to [195], the sym5 wavelet contains less variability in the HF detail than the Haar wavelet, while the db5 contains even less than the sym5 wavelet. This means that the LF approximation contains more variability for the db5 than the sym5 wavelet.

The consequence of this is shown in Figure 7.16. Wavelets with less variability contained in the HF detail, tend to capture the peak demands better in their LF approximation.

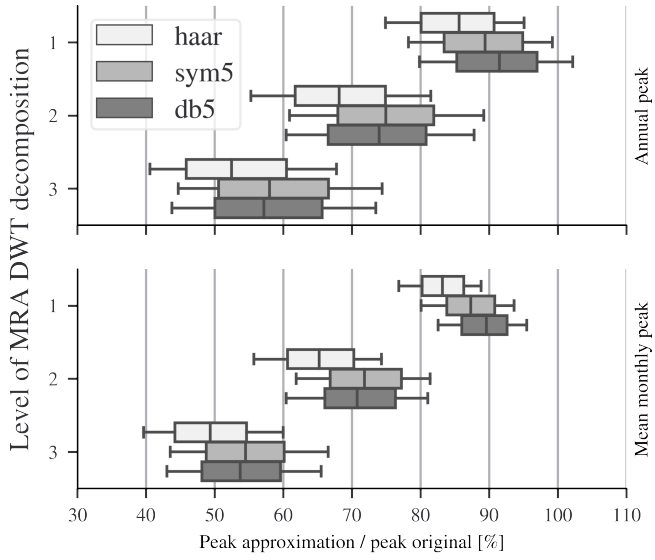

Figure 7.16: Boxplot of the annual (top) and mean monthly (bottom) peak demands of the k-level LF approximation compared to the original load profiles for the Haar, sym5 and db5 wavelets. The whiskers denote the [10, 90] percentiles.

The limiting factor for the representativity of the output of the stochastic profile generator is the number of valid combinations for the largest percentiles of the daily peak demands. In the absence of obtaining a larger dataset, the methodology initially points toward reducing the decomposition level in the MRA-DWT. However, using a more variable wavelet than the traditional Haar wavelet can also lead to more valid combinations.

A second limitation of our current approach is that we only generate daily load profiles. Simplistically combining these stochastic daily load profiles to create annual load profiles can lead to inaccuracies, particularly in terms of preserving autocorrelation. This issue is evident in Figure 7.17, which presents the autocorrelation function (ACF) of the time series of daily consumptions for both a household from the original dataset and a corresponding synthetic profile.

The ACF plot for the measured profile shows significant autocorrelation for lags up to 7-14 days, reflecting the typical weekly patterns observed in residential electricity consumption. In contrast, the synthetic profile's ACF lacks significant autocorrelation at these lags, indicating that the random combination of daily profiles fails to capture the temporal dependencies inherent in real data.

To address this, Chapters 8 and 9 focus on developing methods to combine stochastic daily profiles while preserving realistic autocorrelation. In the next chapter, we explore the generation of synthetic annual profiles of daily consumption values with realistic multiscale temporal dynamics. Subsequently, Chapter 9 inserts the synthetic daily load profiles from the current chapter into the annual profiles generated in Chapter 8.

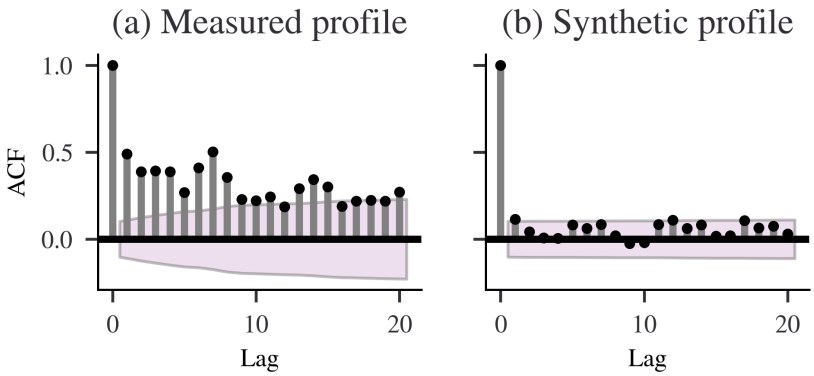

Figure 7.17: ACF of the daily consumptions in (a) a measured profile in the original dataset, and (b) a synthetic profile obtained by randomly combining stochastically generated daily load profiles.

7.9 Conclusion

The primary methodological contribution of this chapter to the state-of-the-art lies in advancing the wavelet-based decomposition-recombination method originally developed for daily load modeling of non-residential consumers [195–197]. We adapted and expanded this framework to make it usable for residential consumers, addressing the limitations of the original normalization schemes, which are ineffective due to the significant influence the high-frequency components have on residential peak demands.

This chapter presented a data-driven stochastic load profile generator for residential consumers. First, privacy-sensitive load profiles can be decomposed into high-frequency details and a low-frequency approximation using the discrete wavelet transformation and a multi-resolution analysis. Second, the high-frequency component of the load profile corresponding with one household can be rescaled and shifted, and subsequently combined with the approximated profile of a different household. This yields a synthetic load profile with a given daily peak demand and daily consumption.

By sampling the two-dimensional distribution of (i) the daily consumption and (ii) the daily load factor of the original dataset, their relation is preserved in the resulting stochastically generated profiles. The generated profiles were subsequently benchmarked with respect to the original dataset for the daily peak demand behavior. The seasonal behavior in the original data was found to be preserved by limiting the generator to the daily level. Furthermore, the distribution of the synthetic daily peak demands showed a MAE of 0.09 kW, corresponding to a MAPE of 4.1%.

Two restrictions of the introduced methodology were discussed. The first limitation concerns the reconstruction of daily load profiles with a high peak demand. These profiles tend to have a significant contribution of the high-frequency component to their peaks, which leads to fewer other households which can be used for constructing a synthetic profile with a similar peak demand. This highlights the trade-off between anonymization of the privacy-sensitive data and the computational process involved in the load profile generation. Second, daily load profiles as output cannot be used for all hosting capacity studies. The autocorrelation of a random sequence of synthetic daily profiles is inaccurate, as measured profiles contain a high degree of autocorrelation between their daily consumptions. This second limitation will be addressed in the next chapter.

8

Construction of annual profiles via Generative Adversarial Networks (GANs)

Generative models are a key enabler of machine creativity, allowing machines to go beyond what they've seen before and create something new.
Ian Goodfellow

In this chapter, we present a novel methodology that uses the potential of Generative Adversarial Networks (GANs) as state-of-the-art technique in time series modeling to capture the long-term correlation inherent to residential load profiles at the annual level.

We first discuss annual profiles and their temporal correlations in Section 8.1. We subsequently present GANs as state-of-the-art for synthetic data generation in Section 8.2, after which we present DoppelGANger as the architecture of choice. The remainder of this chapter is subsequently dedicated to the training of the GAN models and evaluation of their outputs.

Sections 8.3 to 8.5 of this chapter are derived from:

- R. Claeys, R. Cleenwerck, J. Knockaert, and J. Desmet, "Capturing multiscale temporal dynamics in synthetic residential load profiles through Generative Adversarial Networks (GANs)", *Applied Energy*, vol. 360, p. 122831, 2024. [200]

8.1 Descriptions of annual load profiles

For this chapter, annual load profiles represent the total electricity consumption of a household or building over a year, broken down into daily measurements. At this level, load profiles of daily consumptions form a time series of 365 (or 366 for leap years) values. Both at the level of the individual day and on a day-to-day basis, these time series exhibit high variability and are challenging to predict due to stochastic behavior. However, these daily consumptions become more consistent and correlated behaviors emerge at longer time scales.

This consistency is expected, as the observed time correlations reflect the inherent calendrical patterns of household activities. Figure 8.1 illustrates the averaged ACF of all households in the dataset. The plot reveals a clear seasonal dependence in the ACF, with significant lags corresponding to weekly intervals, highlighting regular human activities aligned with the societal seven-day calendar cycle.

Seasonal correlations are particularly influenced by the use of electric heating for space or domestic hot water, which follows the heating and cooling seasons. Additionally, intraday, weekly, and biweekly correlations are evident, reflecting regular consumption patterns throughout different timescales [201].

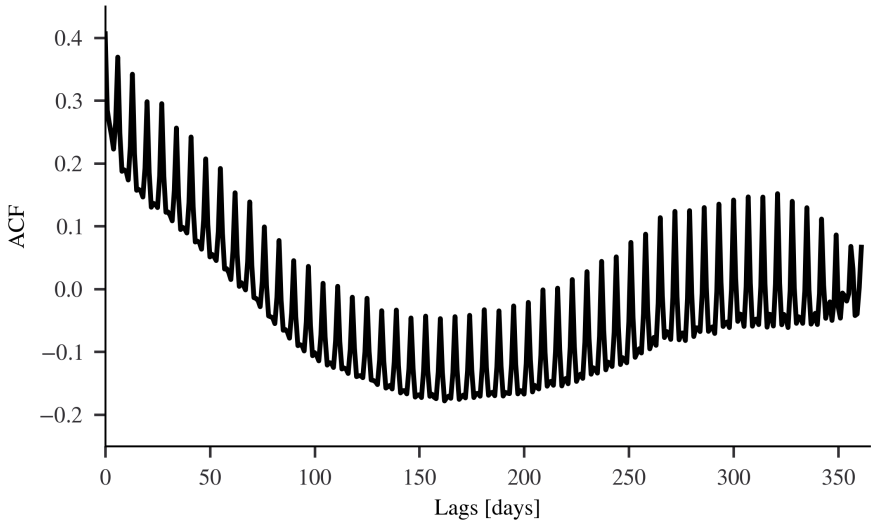

Figure 8.1: Autocorrelation function (ACF), averaged over all households in the considered dataset.

8.2 GANs as state of the art for time series modeling

Introduced in 2014 by Goodfellow et al., GANs are a data-driven modeling technique that takes real samples as input and, when trained until convergence, outputs a model that can generate new samples from the same distribution as the original data [202]. A general block diagram of GANs is given in Figure 8.2.

At a basic level, a GAN consists of two neural networks: a generator (G) and a discriminator (D) with learnable parameters θ_G and θ_D respectively. The generator accepts random noise vectors $z \in \mathcal{Z}$ drawn from a predefined distribution, commonly Gaussian or uniform distributions, and maps it to a learned distribution $\hat{p}$. Concurrently, the discriminator is trained to distinguish between real and generated data samples, respectively x and $G(z)$. It is a binary classifier trained by minimizing the classification error by feeding samples from both the real and generated dataset as input.

Errors in the binary classification are subsequently used to train both G and D through backpropagation as the networks are trained adversarially: D aims to maximize its probability of assigning the correct labels to its input samples, while G aims to minimize this probability. This is expressed in the following loss function:

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{x\sim p} \left[\log D(x)\right] + \mathbb{E}_{z\sim \hat{p}} \left[\log \left(1 - D(G(z))\right)\right] \tag{8.1}$$

The basic GAN architecture is tailored and expanded upon depending on the considered application. In the case of load profiling, they are adapted to suit the temporal nature of the demand data. A single load profile measured during T timesteps can generally be described as $x_{1:T} = (x_1, \dots x_T)$.

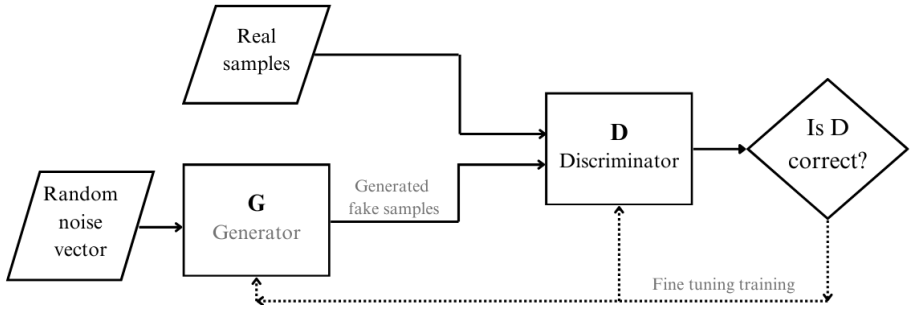

Figure 8.2: High-level block diagram of the GAN framework.

As the considered data are time series with their own temporal dynamics, the neural networks are traditionally chosen to generate data recurrently, where the output at $t-1$ is taken as input for the generation of data at t.

Several illustrative examples from research using basic GAN architectures for load profile generation are discussed in the following paragraphs.

In [203], load profiles for residential consumers were generated using a convolution GAN architecture. The load profiles were generated for a period of one week with a 30 minute resolution. The authors of [204] used recurrent neural networks (RNNs) to generate load data with a time step of 10 minutes, using a sliding window of 60 time steps to synthesize the data. In [205], energy use profiles for retail and commercial office buildings were generated at a 15-minute resolution, using daily data sequences. Their considered GAN architecture used a Long Short-Term Memory (LSTM) for the generator to capture the long-term dependencies in the data in addition to short-term correlations.

Concurrently, efforts have been made to include metadata in the generation process. For example, both [206] and [207] generated load profiles at a 30 minute resolution with corresponding sociodemographic information such as the age of the chief income earner, the number of bedrooms or whether the load profile corresponded to a single-person household. However, the generation process remained limited to profiles describing the load behavior for a single week. As such, no long-term monthly or seasonal correlations were included in the aforementioned studies.

In addition to using neural networks such as RNNs, LSTMs, or temporal convolutions that capture the sequential nature of time series, previous models have tailored the architecture itself to explicitly incorporate the time correlation. Considering this, the two most relevant architectures for our work are TimeGAN and DoppelGANger.

TimeGAN, introduced in 2018 by Yoon et al. [208], is the current state-of-the-art methodology for time series generation. Its diagram and training scheme is shown in Figure 8.3. In addition to using an auto-encoding component, TimeGAN introduces the temporal dynamics directly in the learning process by including a supervised autoregressive loss that incorporates the conditional time dependence, i.e., at each timestep t the architecture includes learning the density $\hat{p}\left(x_t|x_{1:t-1}\right)$ that best approximates the true density $p\left(x_t|x_{1:t-1}\right)$. Furthermore, the architecture explicitly enables the use of metadata during the learning process, allowing the networks to discover links between fixed attributes and their corresponding time series.

TimeGAN was used in [209] to generate both residential and commercial load profiles based on sparse datasets with 12 buildings each. In this study, the generator and discriminator were additionally conditioned on mean outdoor temperature, leading to significant performance gains in the generation process. However, the authors concluded that TimeGAN was unable to sufficiently capture the weekly dynamics of the commercial load profiles with the observed significant differences between weekdays and weekends being absent in the synthetic data.

In addition to the aforementioned study on load modeling, the authors of [210] found that TimeGAN failed to adequately capture weekly and annual correlations in longer time series and concluded TimeGAN is unable to sufficiently capture long-term dependencies or complex temporal dynamics. As previous GAN-based work on load modeling did not consider longer time periods, instead often focusing on daily or weekly load profiles, this aspect has been underexposed in academic literature on load modeling.

Regardless of the issues with respect to capturing the long-term temporal dynamics, the applicability of TimeGAN for time-sensitive applications is severely limited because it uses a sliding window to shift-sample a larger dataset. The obtained sequences are subsequently randomly mixed to make them independent and identically distributed. While this allows for the creation of more sets of sequences to be used in the training process, it comes at the cost of not knowing exactly to what times the sequences correspond to [209].

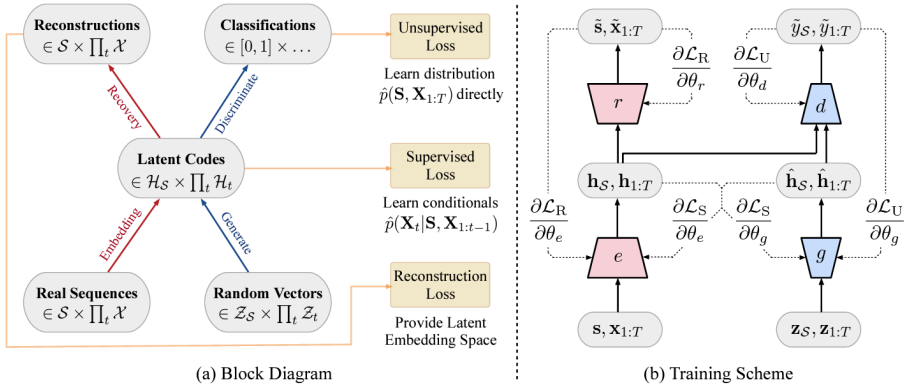

Figure 8.3: TimeGAN block diagram and training scheme as reproduced from [208].

8.3 DoppelGANger

Because of the issues surrounding TimeGAN mentioned in Section 8.2, the GAN architecture of choice for this work is DoppelGANger (DGAN). DGAN is an emerging architecture specifically tailored to generate time series that are influenced by metadata, first introduced in 2020 by Lin et al. [210]. The architecture itself is shown in Figure 8.4.

DGAN offers four main advantages compared to TimeGAN for the generation of residential load profiles:

1. A separate metadata generator is included, decoupling the generation of metadata and time series.
2. It introduces a separate generator that outputs synthetic minima and maxima of time series.
3. Instead of sequentially generating a single value at a time, S batched samples are generated at the same time.
4. The metadata and min-max generator are used in an auxiliary discriminator.

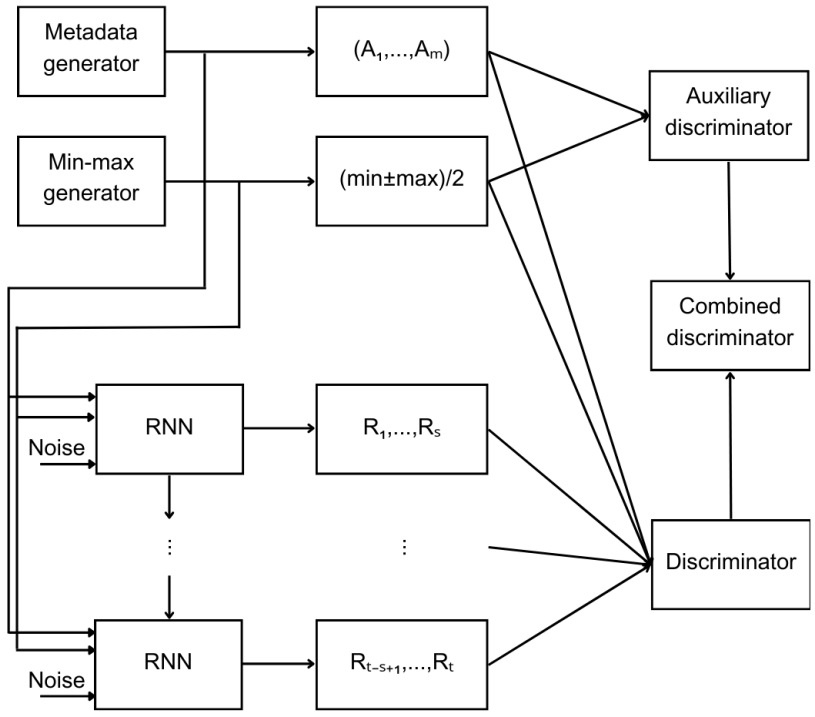

Figure 8.4: DoppelGANger block diagram.

The introduction of the min-max generator combined with the auxiliary discriminator is particularly relevant in the context of residential load modeling. One of the major hurdles during the training of GANs is so-called mode collapse. This occurs when the output does not reflect the diversity of the input samples, instead producing repetitive or limited outputs. Datasets with a highly variable range appear to be more prone to mode collapse [211].

The data under consideration in this dissertation exhibits variable ranges, as they are residential consumers. Individual household annual consumption can range from several hundreds to tens of thousands of kWh. Relevant metadata that are closely linked to the consumption behavior is the presence of, e.g., electric heating or an electric vehicle, as well as dwelling size and type, number of inhabitants and total household income.

Previous GAN-based load modeling either focused on generating daily or weekly time series, or used a single architecture to generate sequences for a longer time period, leading to a loss in long-term correlation and fidelity. In contrast, we maintain the proposed methodology for load modeling based on a decoupled approach where the stochastic behavior at the daily level is modeled separately from the dynamics at the annual level. Consequently, DGAN is used to generate annual load profiles of daily consumptions.

By limiting the GAN generation to a time series of daily values, traditional issues of longer time series modeling through GANs can be avoided. However, the choice of modeling technique, either at the annual or daily level, is somewhat arbitrary. While we opt for a data-driven decomposition-recombination method at the daily level, a GAN-based approach towards generating daily load profiles would also be valid. However, the proposed method allows for fine-grained control on both the daily consumption as well as peak demand compared to a black-box GAN-based approach, which can prove useful, e.g., to model the impact of weather events absent in the original dataset.

The primary objective of this dissertation is to provide a proof of concept for the efficacy of the proposed two-step methodology in generating residential load profiles that are practical for various downstream applications. However, evaluating the fidelity of GAN outputs in a rigorous manner is difficult with no separate objective function for the generator to evaluate the generated data sequences, nor standard agreed-upon evaluation criteria [212, 213].

Consequently, we refrain from asserting that the presented GAN model is entirely optimized, as achieving full optimization would require a rigorous investigation encompassing an exploration of all hyperparameters, i.e., the set of parameters that are not learned during the training process but are set prior to it in order to control and fine-tune both the GAN architecture and training process.

Instead, a suitable GAN-model is selected based on three qualitative microbenchmarks intended to compare the fidelity of the synthetic data with respect to the original data. The microbenchmarks are chosen to be indicative of mode collapse, either in the amplitude or frequency domain, or be a measure for how well the temporal dynamics are recreated.

Overlap coefficient: In the generation process, sequences can be generated that, while exhibiting realistic relative day-to-day variations, show unrealistic absolute values. Hence, the overlap between the distributions of the real and synthetic annual consumptions is chosen as microbenchmark.

Variability coefficient: the overlap between the distribution of the standard deviations of daily consumptions was selected as microbenchmark as it offers a metric that can indicate mode collapse in the frequency domain.

Autocorrelation: The autocorrelation is included as a way to verify both the expected short-term and long-term behavior of the generated samples.

These benchmarks were selected based on common failures observed during the generation process of load profiles, where good results on one benchmark tended to correspond with bad results on another. This trade-off is illustrated on Figure 8.5 for two obtained models.

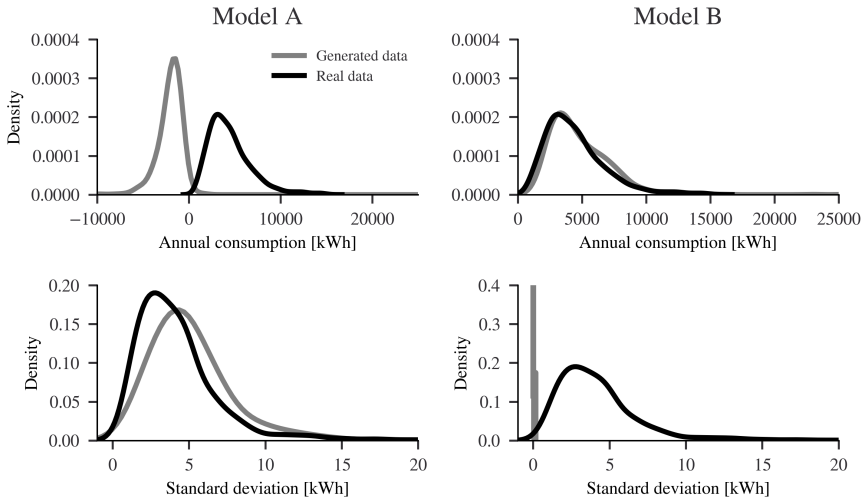

Figure 8.5: Example of failing microbenchmarks concerning the probability density functions of the annual consumption and the standard deviation for the real (black) and generated (gray) datasets. Model A is unable to output realistic absolute values, while mode collapse in the frequency domain materializes in Model B.

To wrap up the discussion in this section, we focus on two additional modeling-related design choices in the DGAN architecture, while the remaining configuration parameters will be swept in search of a suitable GAN model. First, the min-max generator and auxiliary discriminator are both enabled during the generation process. The output of the generator are normalized time series between 0 and 1, which are subsequently rescaled back to realistic ranges.

Second, as shown in Figure 8.4, DGAN allows for batch generation of sequences during each RNN pass. In the context of modeling daily consumptions of residential consumers, a batch size of 7 is deemed to be most suitable as this reflects the weekly calendrical cadence inherent to household lifestyles.

In addition to this content-specific justification for the batch size, this choice can be backed up by a computational argument. The authors of [210] empirically examined the relation between batch size S, sequence length T and their obtained results. They empirically found that setting S so that T/S is around 50 yields good results.

Instead of considering a full year of 365 days, we opt for 364 as sequence length T as this is dividable by the batch size $S = 7$. For the considered time series of 364 daily values, the value of $S = 7$ days therefore yields a ratio of 52, which is consistent with their proposed value to obtain good results with DGAN.

8.4 Results

Before moving to the results and analyses, the hardware and software set-up is first presented in Section 8.4.1. Second, the generation of annual profiles by means of the GAN methodology presented in Section 8.3 is investigated through a sensitivity analysis of the various hyperparameters influencing the generation process. The trained GAN models are evaluated based on the performance of the generated profiles compared to the original dataset through the three introduced qualitative microbenchmarks (variability, annual consumption, and autocorrelation).

Based on the presented sensitivity analysis, a suitable model is selected to investigate the fidelity of a synthetic dataset on downstream applications. Section 8.4.3 presents this selected model that are used for the applications in the following chapter.

8.4.1 Hardware and software set-up

The training process was performed on a hardware environment comprising an NVIDIA RTX 3080 Ti GPU, an Intel i7-12700H (2.3 GHz) processor, and 32 GB RAM. We use the Python-based DGAN algorithm as implemented in `Gretel Synthetics`, version 0.20.0, combined with PyTorch 1.13.1, Tensorflow 2.8.0 and the CUDA toolkit 11.7 [214].

8.4.2 GAN sensitivity analysis

The identification of a suitable GAN model is a twofold process. First, multiple GAN models are trained a total of 500 epochs for different hyperparameter configurations. After training the individual models, 2,000 individual profiles are generated per model. All trained models are subsequently evaluated qualitatively, where we rely on their performance across the three microbenchmarks to determine a suitable model.

Five hyperparameters are included for this sensitivity analysis. These parameters and their considered values are given in Table 8.1. The chosen values of the hyperparameters lead to 1,350 different model configurations. The nomenclature as used by `Gretel Synthetics`'s DGAN is maintained for consistency. The feature and attribute layers denote the number of LSTM layers in the GAN generator network and the number of layers in the discriminator network respectively. For the remainder of this dissertation, the learning rate is abbreviated as LR.

For each combination of hyperparameters, multiple models are trained in order to take the stochastic nature of GAN training into account. Bearing in mind the computational load, we opted to train five different models for each hyperparameter configuration set. This leads to 6,750 GAN models that were trained and evaluated.

Table 8.1: Hyperparameters considered for the sensitivity analysis with their corresponding values.

Parameter	Values
`Batch size`	$[6, 8, 10, 12, 14, 16] \times 10$
`Feature layers`	$[1, 2, 3]$
`Attribute layers`	$[1, 2, 3]$
`Generator learning rate`	$[1, 3, 5, 7, 9] \times 10^{-5}$
`Discriminator learning rate`	$[1, 3, 5, 7, 9] \times 10^{-5}$

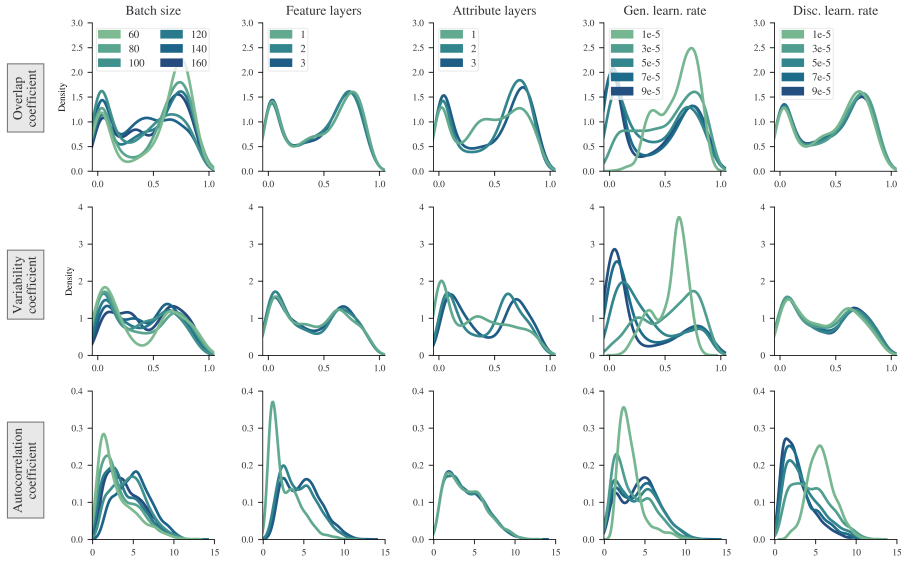

Figure 8.6: One-dimensional marginals of the multivariate probability density function of the microbenchmarks (rows). The five hyperparameters under consideration are displayed in the columns.

The first step of the sensitivity analysis is investigating how different hyperparameters influence the microbenchmarks. Consequently, the marginals of the multivariate probability density functions are displayed in Figure 8.6, which isolates the behavior of each parameter independently in a 1D KDE of their probability density function.

Table 8.2 summarizes the correlations between hyperparameter and microbenchmark as seen in Figure 8.6.

Table 8.2: Summary of the correlations between a hyperparameter and the considered microbenchmarks. $+$ denotes a positive correlation with increasing hyperparameter; $-$ corresponds to a negative correlation; "0" signals no impact of a change in hyperparameter; "?" indicates an ambiguous relationship.

	Overlap	**Variability**	**Autocorrelation**
Batch size	$-$	$+$	$-$
Feature layers	0	0	$-$
Attribute layers	$+$	$+$	0
Generator LR	?	?	?
Discriminator LR	0	0	?

For example, when considering the batch size, both the overlap and variability coefficient exhibit a bimodal distribution. Increasing the batch size negatively impacts the overlap coefficient (more mode collapse as indicated by higher density at minimal values of the coefficient), while it positively impact the variability coefficient. Similarly, the autocorrelation coefficient is negatively impacted by higher batch size.

This isolated behavior with respect to the batch size is expected: larger batch sizes hinder the model's ability to capture granular temporal details in the data, leading to worsening autocorrelation coefficients. However, larger batch sizes lead to faster convergence and by averaging over more samples yields less occurrences of mode collapse for the variability coefficient.

The impact of the LRs is ambiguous. While the discriminator LR has no impact on the overlap or variability coefficient, all coefficients are highly unpredictable under varying generator LR. However, instead of considering the absolute values of the LRs, it is necessary to consider their relative difference. The networks are trained simultaneously in an adversarial manner, so isolated results lead to an inadequate portrayal.

Therefore, we introduce ΔLR, the difference between the learning rates of the generator and discriminator. Figure 8.7 now depicts the behavior of the microbenchmarks under increasing ΔLR.

In contrast to the individual learning rates, a coherent understanding emerges by considering ΔLR: the data generation process heavily favors a negative ΔLR. A strongly negative ΔLR tends to correspond to higher density at high overlap and variability coefficient, and minimal autocorrelation coefficients.

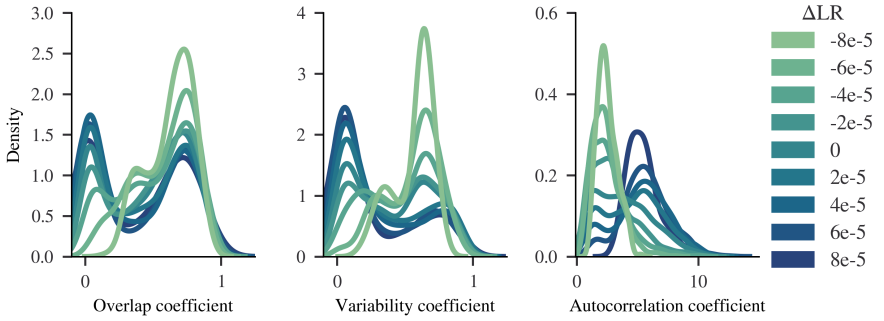

Figure 8.7: One-dimensional marginals of the multivariate probability density function of the microbenchmarks as a function of ΔLR.

The preliminary analysis based on the marginal distributions presents intriguing, albeit incomplete, insights. Bimodal distributions arise for the relation between both the batch size and ΔLR on the one hand, and the microbenchmarks on the other hand.

Nevertheless, the correspondence between zones exhibiting high overlap coefficients and their relationship to either high or low variability coefficients remains ambiguous. To elucidate this association, an exploration into the 2D probability density functions, as displayed in Figure 8.8, is warranted.

The 2D KDE as observed in Figure 8.8 exhibit distinctly different behavior depending on the value of ΔLR. We focus on regions of positive and negative ΔLR respectively.

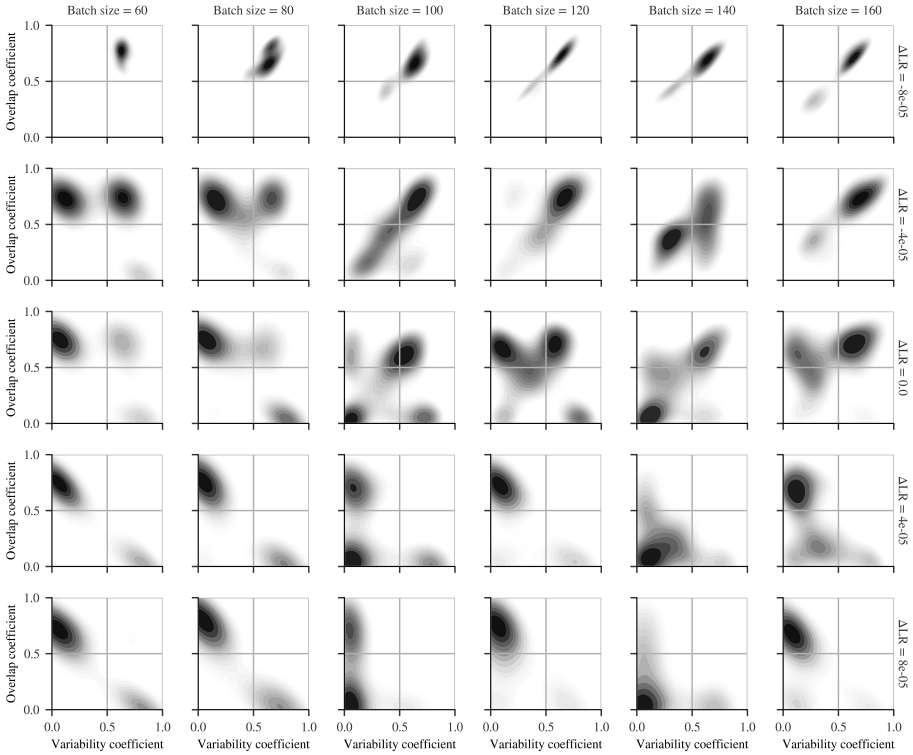

Figure 8.8: 2D KDE between the variability (x-axis) and overlap (y-axis) coefficient, as a function of the batch size (columns) and ΔLR (rows).

For the 2D KDEs with $\Delta LR > 0$, a significant and unwanted trade-off is apparent. Regions exhibiting high overlap coefficients tend to align with a high probability of observing low variability coefficients, and vice versa. This relationship is visually depicted by the probability densities distributed along the principal diagonal of the 2D graphical representations, where minimal probability density is observed within the upper right quadrant. This quadrant signifies scenarios where both microbenchmarks exhibit satisfactory scores. Consistent occurrences of mode collapse are observed, manifesting either (i) in the amplitude of generated load profiles, resulting in a minimal overlap coefficient, or (ii) in the frequency domain, leading to persistently uniform profiles with a zero variability coefficient.

However, this unwanted trade-off disappears when ΔLR becomes negative. For $\Delta LR < 0$, the converged model consistently produces datasets that demonstrates satisfactory performance across both microbenchmark assessments. Simultaneous reduction of the batch size for these learning rates serves to diminish the likelihood of mode collapse during the generative process, as visible in the columns of Figure 8.8. Consequently, a negative ΔLR emerges as the optimal configuration of hyperparameters facilitating the consistent generation of datasets featuring load profiles that exhibit satisfactory performance across all the three microbenchmarks.

In conclusion, the analysis conducted in this section has provided valuable insight into the responsiveness of the proposed GAN model to variations in several key hyperparameters. The following sections delve into the performance of the proposed approach across several downstream applications. Therefore, for the remainder of this dissertation, we exclusively consider a single GAN model. The hyperparameters for this model, listed in Table 8.3, represent an informed selection based on the insights garnered from the sensitivity analysis.

Table 8.3: Hyperparameter configuration set considered for the remainder of this dissertation.

Hyperparameter	**Value**
`Batch size`	160
`Feature layers`	1
`Attribute layers`	3
`Generator learning rate`	1×10^{-5}
`Discriminator learning rate`	9×10^{-5}

8.4.3 Selected GAN model

This section presents an examination of a single GAN model intended for utilization in the following chapter's downstream applications. The discussion of the selected model is twofold. First, the microbenchmarks of the considered GAN model are discussed. Second, we investigate potential memorization effects that may have transpired throughout the training process.

The benchmarks are displayed in Figure 8.9, while their performance is evaluated in Table 8.4. The Mean Average Error (MAE) and Root Mean Square Error (RMSE) are the metrics considered for the performance evaluation. From these comparisons, it is clear that DGAN is able to output a model that generates annual load profiles which (i) display the same distribution in their total annual consumption, and (ii) exhibit similar variability.

The batched generation ensures the weekly correlations, while the seasonal correlation remains accurate due to the smaller number of individual RNN generation steps. Furthermore, the hyperparameter configuration set all but eliminated mode collapse in the amplitude or frequency domain occurring during the training process, leading to consistent results.

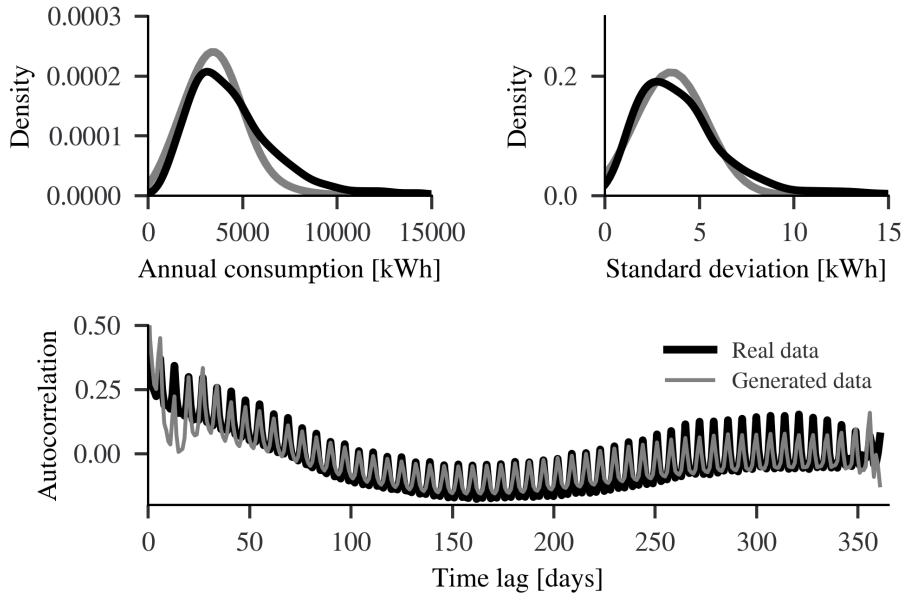

Figure 8.9: Microbenchmarks of the considered GAN model compared to the real dataset.

Table 8.4: Evaluation of the generated dataset shown in Figure 8.9.

	MAE	RMSE
Annual consumption	0.0000143	0.0000197
Standard deviation	0.0000158	0.000438
Autocorrelation	0.0297	0.0419

The autocorrelation plot does reveal minor discrepancies within time lags of two weeks. We will elaborate on the limitations of both the proposed methodology and the three microbenchmarks used to evaluate the trained GAN models in a separate section.

A major concern of GAN is overfitting during the training process, especially if the training dataset is limited. This overfitting leads to memorization of individual profiles, defeating the purpose of providing generated synthetic data as a feasible alternative to privacy-sensitive metering data. To evaluate the memorization effect, individual synthetic profiles are compared with their nearest samples in the training data. This is illustrated in Figure 8.10 for three randomly selected profiles. Significant differences can be observed, both quantitatively in the square error as well as qualitatively in the trends of the load profiles, which suggests that no memorization occurred.

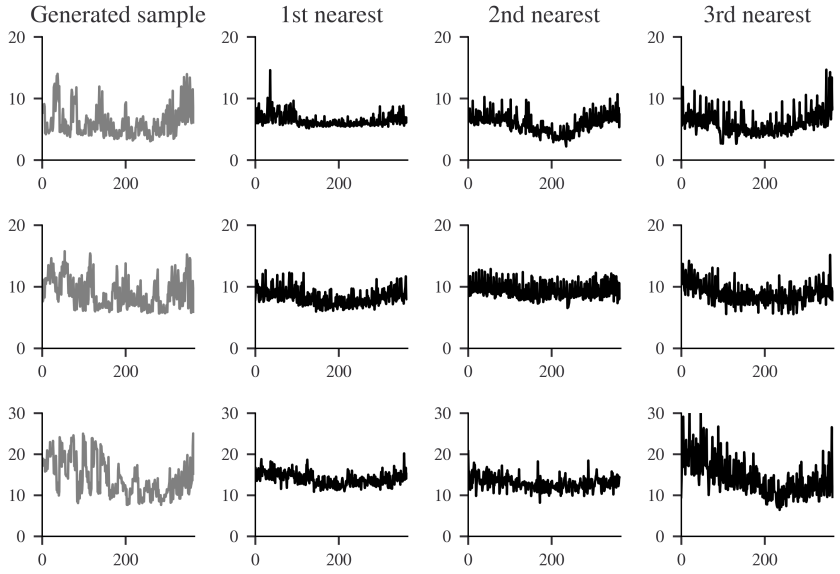

Figure 8.10: Three profiles randomly selected from the generated dataset and their corresponding three nearest neighbors.

In the interest of reproducibility, the complete configuration of the 29 DGAN parameters, in addition to the five hyperparameters that were included in the sensitivity analysis presented in Section 8.4.2, that resulted in the discussed output is given in Table 8.5.

Table 8.5: DGAN configuration settings for `Gretel synthetics` that yielded the presented GAN model.

Parameter	**value**
max_sequence_len	364
sample_len	7
attribute_noise_dim	10
feature_noise_dim	10
attribute_num_layers	3
attribute_num_units	100
feature_num_layers	1
feature_num_units	100
use_attribute_discriminator	True
normalization	ZERO_ONE
apply_feature_scaling	True
apply_example_scaling	True
binary_encoder_cutoff	True
forget_bias	False
gradient_penalty_coef	10.0
attribute_gradient_penalty_coef	10.0
attribute_loss_coef	1.0
generator_learning_rate	1E-5
generator_beta1	0.5
discriminator_learning_rate	9E-5
discriminator_beta1	0.5
attribute_discriminator_learning_rate	0.001
attribute_discriminator_beta1	0.5
batch_size	160
epochs	500
discriminator_rounds	1
generator_rounds	1
cuda	True
mixed_precision_training	False

8.5 Limitations

The autocorrelation plot in Figure 8.9 reveals minor discrepancies within time lags of up to two weeks. These deviations between real and generated data at the weekly and biweekly levels can be attributed to the challenges of accurately capturing the general weekly calendrical cadence, the variable behavior during holiday periods, and the presence of electric heating for a subset of households in the dataset, which skews the autocorrelation during heating seasons.

This calendrical dependence is a significant issue not only for the microbenchmarks but also for the generation process itself. For instance, when using data from 2013, the generated profiles inherently reflect the calendrical properties of that specific year, with holidays like Easter occurring at the same time. This limits the profiles' applicability for other years. Similarly, if multiple datasets from different years (e.g., 2013 and 2014) are used as input, it becomes necessary to label the days of the year for the generation process. Without this labeling, the generator cannot produce realistic load behavior for a single year due to non-synchronous holidays. One possible area of improvement is therefore labeling the datetime series to include information on holidays and periods with closed schools to enable the GAN to incorporate this information in the learning process to increase the utility of the proposed methodology.

Furthermore, the autocorrelation behavior of households with and without electric heating differs significantly during heating and non-heating seasons. Grouping these two types of households together for the evaluation of a single autocorrelation-based microbenchmark that spans an entire year introduces additional difficulties. The divergent seasonal consumption patterns result in mixed autocorrelation signals, complicating the assessment of the synthetic data fidelity. Limiting the generation process to a single cluster is expected to alleviate this concern.

8.6 Conclusion

The primary methodological contribution of this chapter to the state-of-the-art of load modeling lies in using GANs as state-of-the-art technique for synthetic data modeling to generate annual load profiles. Based on a review of literature on time series generation, we selected the DoppelGANger architecture for this dissertation due to its proven ability to capture long-term time correlations and its capability to batch-generate short-term intraday variations at the weekly level.

In the absence of established evaluation criteria for the output of GANs, our contribution extends to the introduction of three microbenchmarks which assess the synthetic dataset's fidelity in terms of (i) annual consumption values, (ii) daily consumption variability, and (iii) time correlations within the generated load profiles. The first two microbenchmarks are specific for smart meter data and are able to measure mode collapse in the amplitude and frequency domain respectively. Furthermore, to the best of our knowledge we are the first to publish an in-depth sensitivity analysis of GAN hyperparameters and show how they influence the generation process. We finally identified a set of hyperparameters that effectively eliminated mode collapse in both the frequency and amplitude domains.

However, the evaluation of the DGAN-generated profiles using the proposed microbenchmarks should not be viewed as conclusive evidence of the methodology's effectiveness in producing realistic profiles. Instead, these benchmarks primarily validate that the GAN model has successfully converged and is functioning as intended according to the microbenchmarks. Therefore, it is essential to further validate the fidelity of the high-frequency profiles on an annual scale and assess their practical utility. This comprehensive validation is performed in the following chapter, where we consider the performance of the generated data across various use cases relevant to residential consumers.

9

Benchmarking of the synthetic high-frequency annual load profiles

The proof of the pudding is in the eating.
English proverb, earliest known written reference by William Camden.

In this penultimate chapter, we validate the high-resolution annual load profiles using a series of benchmarks. Drawing inspiration from the overview of use cases for residential smart meters as discussed in Chapter 2, we benchmark the constructed synthetic dataset at both the individual and aggregated level.

At the individual level, the evaluation is threefold. First, we compare the LDCs of the synthetic data to those of the real data to assess distributional similarity. Second, we validate whether the synthetic data accurately reconstructs different elements of the electricity invoice, with a focus on the mean monthly peak demand behavior for peak-based pricing and the commodity cost under dynamic pricing. Third, we demonstrate that the synthetic data can replicate the performance of real data in applications involving both PV and PV-BESS systems. At the aggregated level, hosting capacity (HC) studies involving EVs and heat pumps serve as validation.

The benchmarks presented in Subsection 9.2.4 to 9.2.6 were previously published in:

- R. Claeys, R. Cleenwerck, J. Knockaert, and J. Desmet, "Capturing multiscale temporal dynamics in synthetic residential load profiles through Generative Adversarial Networks (GANs)", *Applied Energy*, vol. 360, p. 122831, 2024. [200]

Similarly, the benchmark discussed in Subsection 9.3.1 is derived from:

- R. Claeys, R. Cleenwerck, J. Knockaert, and J. Desmet, "Stochastic generation of residential load profiles with realistic variability based on wavelet-decomposed smart meter data", *Applied Energy*, vol. 350, p. 121750, 2023. [188]

9.1 Combining daily and annual profiles

Our methodology to generate annual data at high resolution was introduced in Chapter 3. We decoupled the generation process as visualized in Figure 9.1, opting to model each timescale with the appropriate technique.

The annual profiles obtained through GANs as described in Chapter 9 and daily profiles from Chapter 8 are combined to obtain annual load profiles with high-frequency variability and realistic temporal short-term and long-term dynamics. For each day of a generated profile of daily consumptions, the stochastic daily load profile with the closest consumption is rescaled to the value obtained during the GAN modeling and appended to the new high-frequency profile. For the proposed validations, 2,000 individual profiles at the annual level are generated.

It is important to note that, unless otherwise stated in the benchmark description, no metadata is taken into account for (i) the generation process, and (ii) the linking process, where the daily load profiles are added together based on the annual load profile. The inferred clusters from Chapter 6 could theoretically serve as metadata. However, this would lead to subsets of the considered dataset that are too small to reliably generate synthetic data without memorization effects occurring. To provide context for the limitation related to cluster size, the publication that introduced DGAN considered a minimum training data size of 500 samples [210].

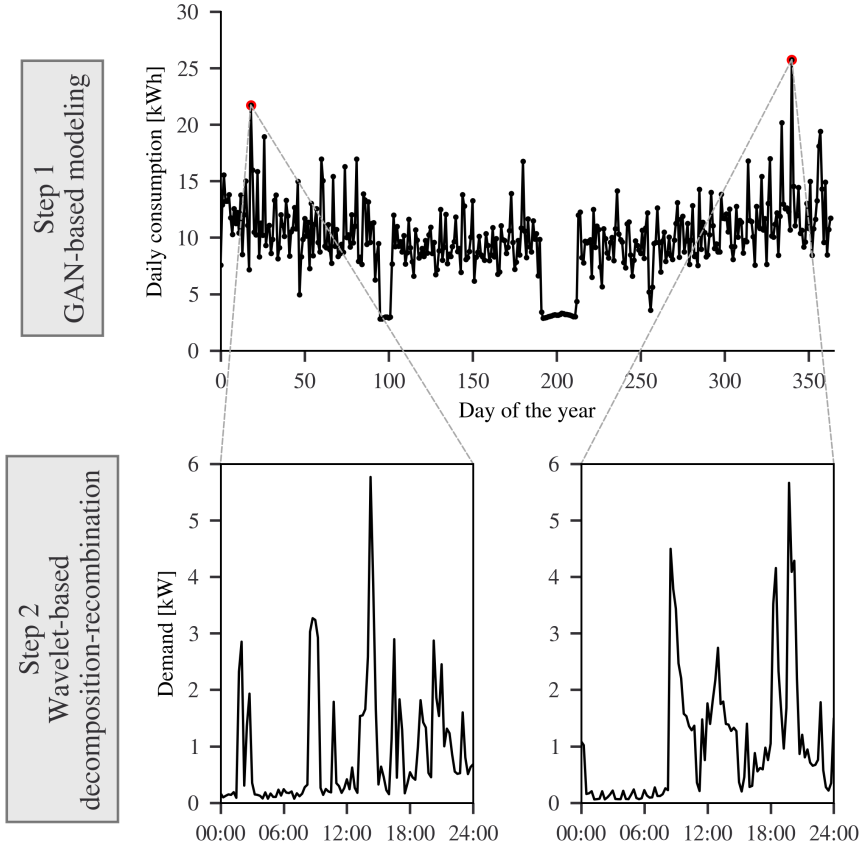

Figure 9.1: Two-step methodology for generating an annual load profile at high resolution.

9.2 Benchmarks at the individual level

9.2.1 LDC parameters

The first considered benchmark at the individual level concerns the 5-parameter LDC of individual consumers as introduced in Equation (5.2) in Chapter 5. As such, in this section we verify whether the dataset of generated synthetic consumption data exhibits the same distributional behavior for all parameters that shape the load-duration curve of the consumers in the original, measured dataset.

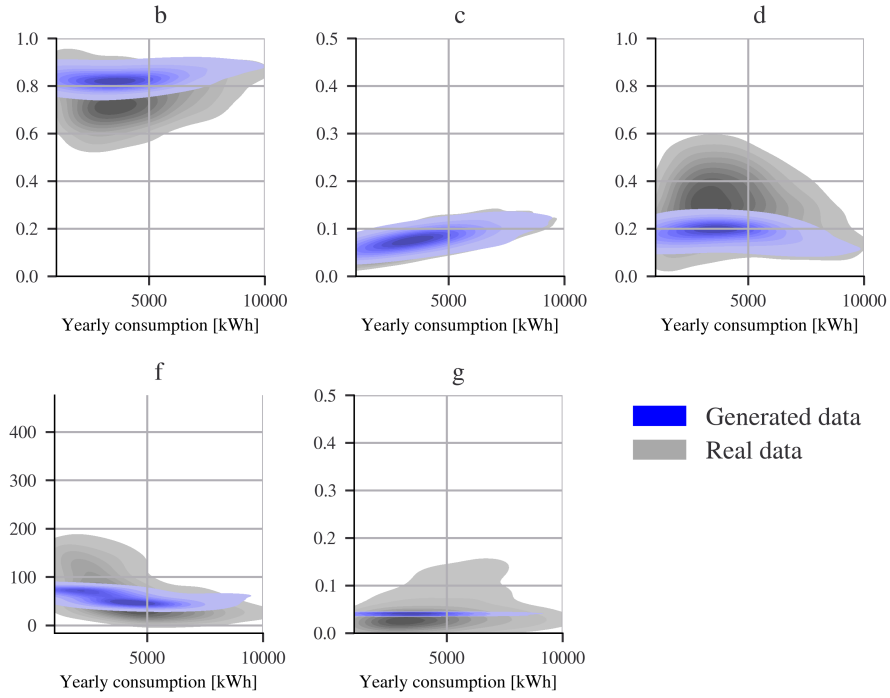

Figure 9.2: Comparison between the 2D KDEs of the five LDC parameters.

The 2D KDE of the parameter values obtained from the fitting procedure are given in Figure 9.2 for both the real (gray) and synthetic (blue) data. The parameter values are given as a function of the annual consumption, as was done in Figure 5.2. Furthermore, Figure 9.3 on the following page reiterates Figure 5.3 from Chapter 5 for additional context to these results. This plot displays the generic LDC shape and the relation between the five parameters and the shape of the fit.

Although the synthetic data generation process successfully replicates the broad trends of the parametric distributions observed in the real dataset, it falls short in capturing the variability in certain parameters, notably b, d, f, and g. However, as illustrated in Figure 9.3, these parameters show inherent correlations due to the positioning and magnitude of the exponential decay in the LDC. For instance, an overestimation of the b parameter tends to coincide with an underestimation of the d parameter.

Hence, we hypothesize that the underlying factor(s) contributing to the deviation in these three parameters are shared. In Chapter 8, we discussed several limitations associated with the Generative Adversarial Network (GAN) generation process.

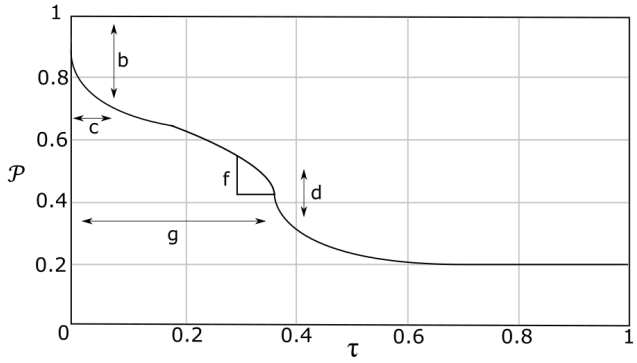

Figure 9.3: Shape of the 5-parameter LDC.

First, due to the constraints of sample size, both in terms of the number of consumers and temporal extent, the generation process struggles to accurately capture the general weekly calendrical patterns while accommodating deviations during holiday periods and the heating season. These variations also significantly impact the autocorrelation behavior, one of the considered microbenchmarks.

Second, the absence of metadata in the generation process poses another limitation. The limited number of consumers within each cluster prevents the inclusion of metadata as a training set for the generative AI. However, it is important to bear in mind that the step in the LDC function associated with the d, f, and g parameters is influenced by consistently high consumption attributed to larger appliances like electric heating or boilers.

Therefore, our current working hypothesis suggests the necessity of introducing clusters based on the presence of larger electrical appliances within households to replicate the observed distributions accurately. This approach is expected to yield coherent clusters of sufficient size that can effectively guide the GAN generation process for more realistic interday behavior.

9.2.2 Mean monthly peak demand

Sections 9.2.4, 9.2.5, and 9.2.6 will investigate the applicability of synthetic data for PV(-BESS) installations. Systems designed to maximize PV self-consumption and those controlled to both increase PV self-consumption and perform peak shaving are considered.

However, to accurately perform an economic optimization for the hybrid control strategy, it is necessary to first verify if the synthetic data can accurately reconstruct all aspects of the electricity bill of residential consumers. In this section, we focus on the mean monthly peak demand, as the non-volumetric component of the grid tariffs for residential consumers in Flanders is based on this value.

Until the end of 2022, network costs in the Flemish distribution grid tariffs were calculated purely on a volumetric basis. However, starting in 2023, the operator's network costs are no longer billed volumetrically to households. Smart meters log the monthly peak demands, which are then averaged over the past twelve months to determine a mean monthly peak demand. Additionally, a minimum monthly peak of 2.5 kW is assumed, meaning that if a household's monthly peak demand is less than 2.5 kW, it will still be billed the cost equivalent of 2.5 kW.

The comparison between the mean monthly peak demand of the real and synthetic datasets is given in Figure 9.4. The distributions are visualized in boxplots binned per MWh annual consumption. As annual consumptions higher than 9 MWh are underrepresented in the synthetic dataset, this allows for a fair comparison between the real and synthetic dataset. The whiskers of the boxplots denote the [10, 90] percentiles.

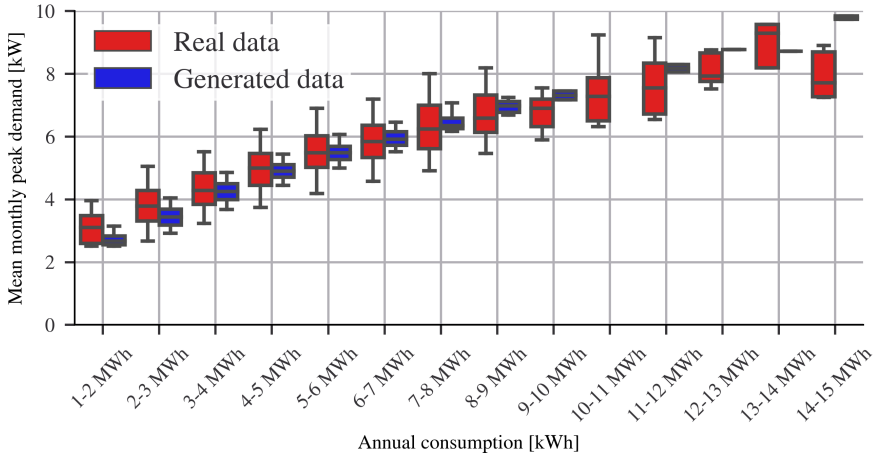

Figure 9.4: Binned boxplots visualizing the distributions of the (corrected) mean monthly peak demands for different annual consumption ranges.

Figure 9.4 reveals two conclusions about the peak behavior of the synthetic dataset. First, the general trend is accurately captured by the synthetic data. However, the synthetic data does not accurately reconstruct the diversity in mean monthly peak demands as seen from the interquartile range and whiskers of the boxplots.

To investigate the possible cause of this underestimated diversity, Figure 9.5 considers the distributions of the standard deviations of the 12 individual monthly peak demands used to calculate the mean monthly peak demand. The monthly peaks of the synthetic individual consumers exhibit a significantly higher standard deviation.

This observation is expected and not an immediate cause for concern, as the linking process of the considered daily profiles is currently completely random, unguided by additional information. Instead of consistently combining either high or low monthly peak demands, consistent with real-life habits, the random combinations currently regress towards the mean value of the binned distributions, leading to a lack of diversity in the averaged monthly peaks. However, even despite this random combination, the observations from Figure 9.4 are encouraging. Limiting the generation process to consumers showing similar behavior in their daily consumption versus daily peak demands is expected to further increase the consistency and therefore the diversity in the mean monthly peak demand.

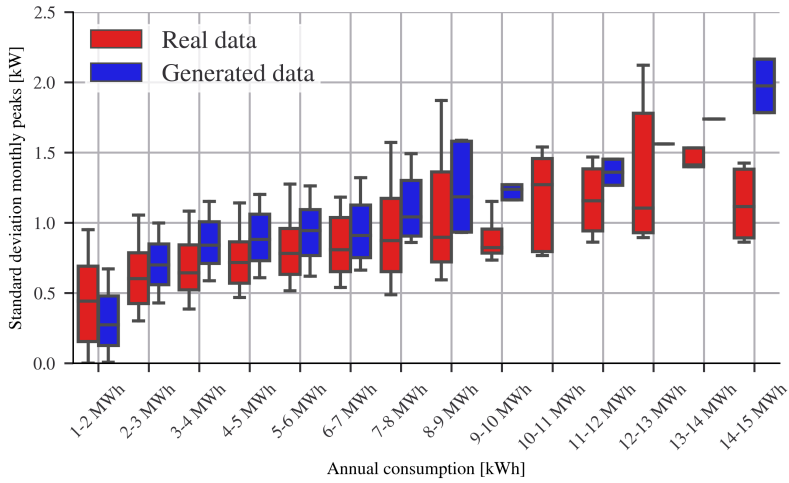

Figure 9.5: Distributions of the standard deviation of the individual monthly peak demands, per dataset and consumption range.

9.2.3 Annual commodity price under dynamic tariffs

While the previous section examined the average monthly peak demands for peak-based grid tariffs, this section shifts focus to the commodity prices under a dynamic tariff system. To calculate the total commodity cost for each consumer, we used the hourly energy prices from 2023 provided by Eneco, an energy supplier. The hourly cost is calculated as follows:

$$(0.102 \times \text{Belpex-H} + 1) \times 1.06, \tag{9.1}$$

where Belpex-H indicates the wholesale Day Ahead price.

The distributions of the calculated annual commodity cost under a dynamic tariff are given in Figure 9.6. An excellent agreement is observed between the real and generated datasets for annual consumptions lower than 9 MWh. This deviation is expected, as the higher annual consumptions are underrepresented in the output of the chosen GAN model, as previously discussed in Section 8.4.3.

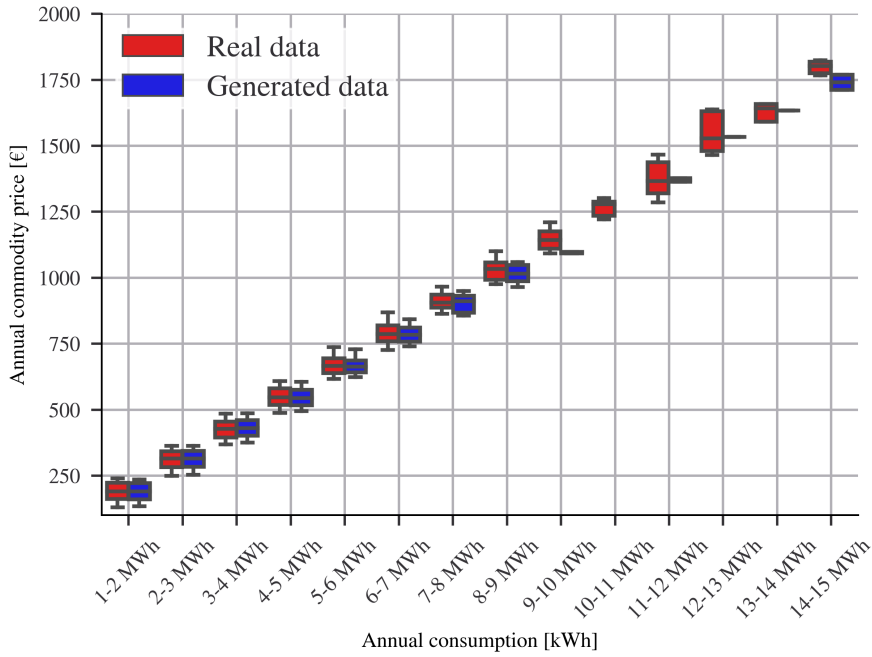

Figure 9.6: Distributions of the annual commodity price under a dynamic tariff, using the Belpex Day Ahead Market prices of 2023, for real (red) and synthetic (blue) datasets.

9.2.4 PV installation

The self-consumption ratio (SCR) is considered as an evaluation metric of interest. The SCR is defined as the ratio between the instantaneous consumed PV energy and the total produced PV energy, for our case over the entire year [12]. Distributional deviations in the observed SCR for similar irradiation profiles can be indicative for consumption which is modeled at the wrong time period, either at the seasonal or intraday level.

To account for various sizes of photovoltaic (PV) installations, we adopt the conventional sizing scheme in Flanders as the baseline. This baseline is established at annual yield = annual consumption. We introduce the parameter k_{PV} as a relative size factor in relation to the base case, where k_{PV} ranges from 0.5 to 1.5. This sweeping range encompasses installations undersized (with $k_{PV} < 1$) and oversized (with $k_{PV} > 1$) relative to the established base case, enabling a comprehensive exploration of different installation sizes for each consumer.

First, a visual comparison of the performance of the synthetic dataset for the PV self-consumption is shown in Figure 9.7 on the next page. The left subplot highlights the probability density function for the SCR for the conventional PV sizing scheme, while the right subplot displays the results for the relative sizings.

This benchmark highlights challenges in accurately modeling consumption behavior. The generated dataset exhibits limitations in capturing the low SCRs that are present in the original dataset. Notably, the SCR distribution of the generated data is concentrated around the most prevalent value observed in the real dataset. This discrepancy can be observed both for the conventional sizing scheme, as well as for different sizings of the considered PV installations. The following paragraphs explore the underlying cause of this discrepancy, as well as how to alleviate this problem.

So far, the generation process of the load profiles has not yet taken any information related to metadata into account, as this is often unavailable due to privacy concerns. However, previous research has shown that metadata such as dwelling type, inhabitant properties as well as appliance ownership are intrinsically linked to household consumption behavior [90, 215, 216]. For example, the presence of an electric heating unit significantly contributes to the consumption behavior, both intraday and seasonally [217].

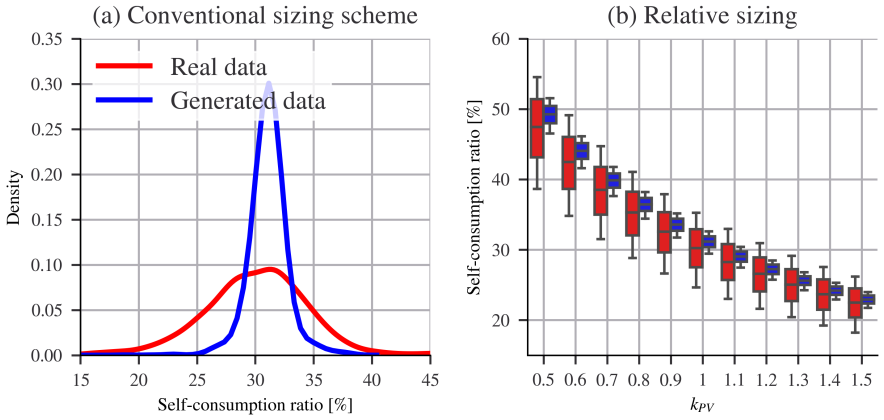

Figure 9.7: Results of the PV SCR for the real (red) and synthetic (blue) datasets, where metadata is not yet taken into account. The whiskers of the boxplots denote the [10, 90] percentiles.

Therefore, the discrepancy in Figure 9.7 needs to be attributed to the exclusion of metadata during the generation and linking process. This can be grasped intuitively by considering the step described in Section 9.1. Therein, high-resolution profiles at the daily level are linked to a consumption value of day d for a GAN-generated time series. However, this allows for the assignment of a load profile with LF approximation belonging to a household with electric heating at day d, while day $d+1$ of the same annual load profile gets assigned a LF profile without electric heating. While the consumption at the daily level matches by design, the interday behavior is inconsistent and unrealistic.

The joining of daily LF approximations belonging to all types of consumers leads to the tendency of the distribution of the synthetic SCR to regress to the mean of the real SCR distribution, without reaching the variety present in the original dataset. Therefore, the proof of concept presented in this dissertation needs to be extended for further validation by taking into account metadata during the generation and linking process. However, as no metadata are available for the considered dataset, the unsupervised learning methodology from Chapter 6 is used to partition the households comprising this dataset into clusters with similar properties.

For the intended proof of concept validation of this chapter, we limit the number of clusters to three, as one of these already groups the households with a significantly high consumption during the night pointing towards the presence of electric heating in these dwellings [152].

This cluster contains 102 households, which is consistent with 7% of households in Flanders that use electricity as energy source for space heating at the time of the data gathering [218].

This clustered subset is subsequently used to generate a new synthetic dataset of households with high nighttime consumption, both in DGAN as well as in the wavelet-based recombination scheme. As the clustered dataset only contains 102 entries, the results should be interpreted with caution. Nonetheless, the results of including metadata are encouraging.

Figure 9.8 displays the SCR of the clustered subset, as well as the previous results for the SCR shown in Figure 9.7. As intuitively expected, limiting the generation process to households with significant nighttime consumption leads to a distribution of the SCR which is significantly lower and captures the lower bound of the SCR observed in the original dataset for all considered sizing schemes, which was absent when excluding metadata during the process.

It is clear that the inclusion of metadata is necessary to obtain annual profiles with high-frequency resolution which exhibit realistic interday variability consistent with their metadata. This has been illustrated for a clustered subset with high nighttime consumption, but could analogously have been proven for households with high daytime consumption, e.g., corresponding to retirees, leading to significantly higher SCR values than obtained in the dataset constructed without metadata.

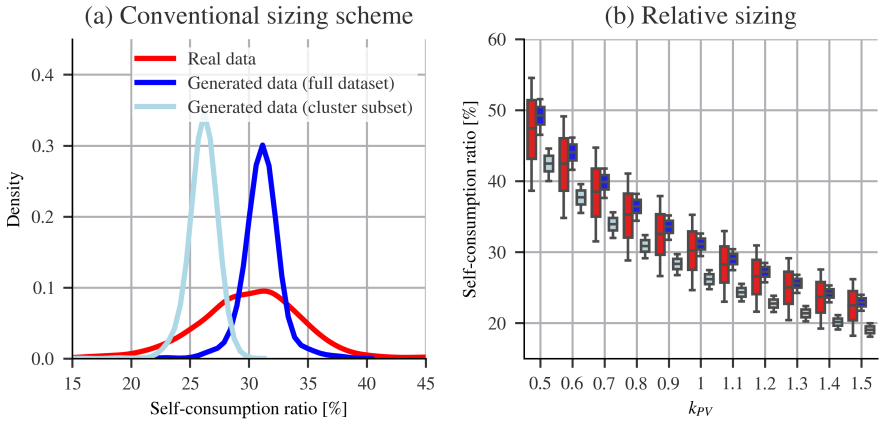

Figure 9.8: SCR of the real dataset (red), a synthetic dataset generated without metadata (blue) and a dataset generated on a subset with high nighttime consumption (light blue).

9.2.5 PV-BESS installation (self-consumption)

The second evaluation metric under consideration is the increase in PV self-consumption realized by adding a BESS to a regular PV system. Residential BESS traditionally operate on an intraday basis, e.g., charging during the day and discharging during the evening. Deviations between real and synthetic data therefore can indicate unrealistic modeling of the daily consumption behavior and variability.

The determination of the BESS size parallels the methodology used for sizing the PV installation: a conventional sizing criterion is established for BESS before introducing a relative size factor. We initially consider a lithium-ion battery size of 1 kWh per MWh annual consumption for residential applications. This sizing methodology aligns with common practices in Western European countries [219, 220].

Introducing the parameter k_{BESS} as the relative BESS size factor, the range of k_{BESS} values varies between 0.5 and 1.5, allowing for an examination of diverse BESS sizes relative to the base case.

The performance of the synthetic dataset for this application is investigated in Figure 9.9 on the following page. The upper subplot illustrates the increase in self-consumption achieved through the conventional sizing scheme of 1 kWh BESS capacity per MWh annual consumption, in conjunction with 1 MWh annual production per MWh of consumption. The synthetic dataset successfully emulates the observed behavior in real data for the conventional sizing scheme.

Given the observed linear trend, an Ordinary Least Squares (OLS) estimation was conducted, providing both slope and intercept values, as included in the legend of subplot (a). Subplots (b) and (c) respectively depict the variations in slope and intercept for relative sizings of both PV and BESS components.

While the generated dataset accurately reproduces the slope, indicated by the congruence of the surface plots in subplot (b), a consistent underestimation of the intercept is evident. Notably, this deviation becomes more pronounced for larger PV installations. However, given the magnitude of the increase in self-consumption, the deviation in the OLS intercept can be considered negligible.

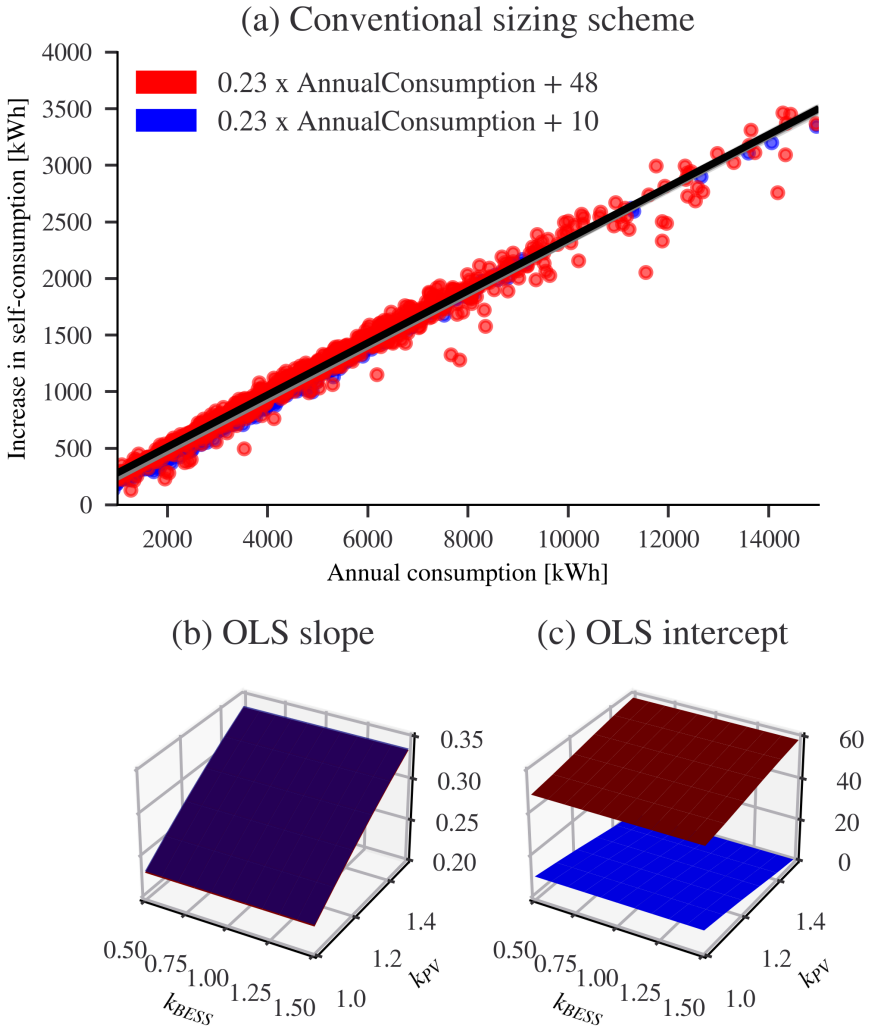

Figure 9.9: Results of the PV-BESS application, where the storage system is solely used to improve the PV self-consumption, for the real (red) and synthetic (blue) datasets, where metadata is not yet taken into account. Subplot (a) depicts the increase in self-consumption for the conventional sizing scheme with OLS fits for both datasets, while subplots (b) and (c) respectively present the slope and intercept of the OLS estimation for relative sizes of both PV and BESS components.

9.2.6 PV-BESS installation (hybrid)

The third considered metric is related to a PV-BESS system with a hybrid control strategy. The optimal control strategy between increasing the PV self-consumption and peak shaving when the demand exceeds a fixed threshold is determined for an economic optimization. The peak shaving behavior serves as a validation for the realistic modeling of the daily peak demands and variability.

Figure 9.10 schematically displays the operation of the BESS for this case study. Five distinct modes can be observed:

(a) At any time t, the first priority is maintaining the minimally reserved capacity for peak shaving at time $t+1$. If $\text{SoC}(t) < \text{SoC}_{th}$, the BESS charges from the grid as long as $P(t) - PV(t) + P_{\text{charge}} < P_{th}$.

(b) If $\text{SoC}(t) > \text{SoC}_{th}$, the BESS only charges with self-produced PV energy. Once the BESS is full, it remains idle and the excess PV energy is injected into the grid and sold at a feed-in tariff.

(c) Charging with PV-produced energy occurs bearing in mind the involved efficiency and maximum inverter power.

(d) If the net household demand exceeds the predetermined threshold P_{th}, the BESS performs peak shaving, bearing in mind the maximum inverter power, available SoC, E-rate, and inverter efficiency.

(e) If the net household demand does not exceed P_{th}, the BESS is allowed to discharge while keeping its SoC above SoC_{th}.

P_{th} and SoC_{th} are determined iteratively at the level of the individual households to find the combination which minimizes the annual electricity bill, given a fixed electricity price, feed-in tariff, and peak-based price. For the peak shaving aspect of the hybrid control strategy, a demand threshold has to be determined above which the BESS will discharge. Furthermore, a percentage of the BESS capacity is reserved for the purpose of peak shaving at all times as no forecasting is assumed. SoC_{th} is the State of Charge (SoC) corresponding to this reserved capacity.

Table 9.1 lists the specifications of the considered battery energy storage systems, as well as tariff properties used for the iterative optimization.

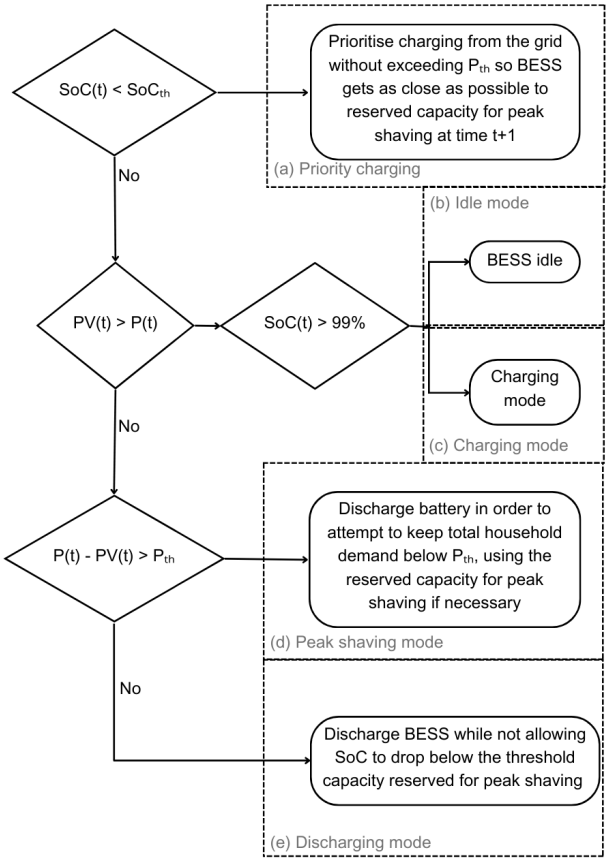

Figure 9.10: Flowchart of the BESS operating scheme under the hybrid strategy combining self-consumption with peak shaving.

Table 9.1: Considered BESS and tariff properties.

Parameter	Value
Type	Lithium-ion battery
Capacity (reference)	1 kWh / MWh consumption
E-rate	1
Inverter power	$\min\{5\text{ kW}, \max(\text{demand})\}$
Self-discharge rate	3% per month
Depth of discharge (max)	80%
Electricity price	30 c€/kWh
Feed-in tariff	10 c€/kWh
Capacity-based grid tariff	50 €/kW

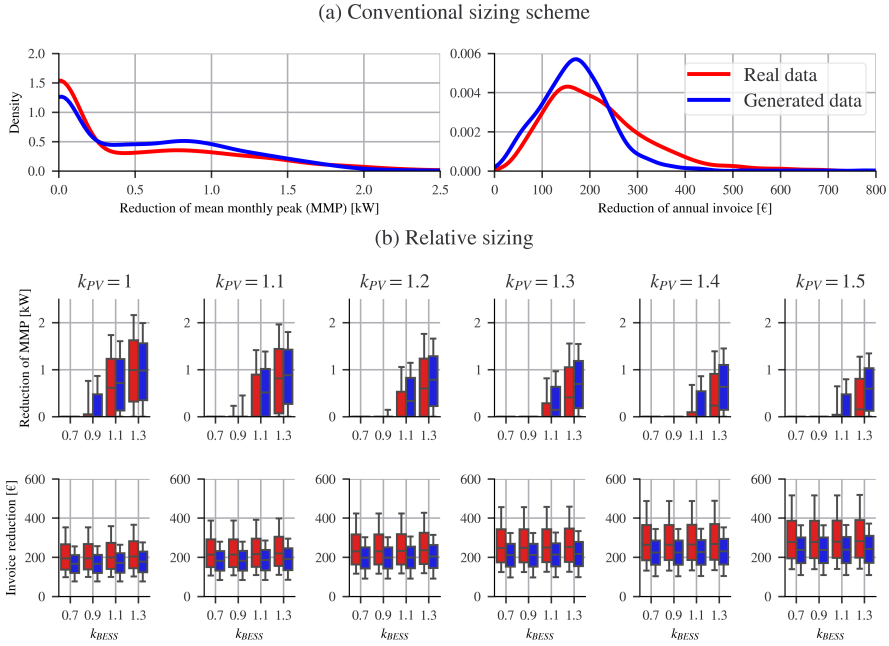

Figure 9.11: Results of PV-BESS configurations where the storage system is used to reduce the mean monthly peak (MMP) demand and improve the PV self-consumption, for the real (red) and synthetic (blue) datasets, where metadata is not yet taken into account. Subplot (a) depicts the results for the conventional sizing scheme, while subplots (b) present the reduction of the MMP and annual invoice for relative sizes of both PV and BESS components.

Figure 9.11 presents the performance of the generated dataset benchmarked with respect to the real dataset for the final, individual downstream application. Two values are displayed both for the conventional sizing scheme and the relative sizings: the reduction of the mean monthly peak demand and the annual invoice reduction.

While the distributions of the mean monthly peak demand reduction show an agreeable match for the conventional sizing scheme, the generated data encounters challenges in accurately reproducing the extended tail observed in the annual invoice reduction distribution. This discrepancy can be attributed to the incongruity in the annual consumptions within the generated dataset. Figure 8.9 displays the distribution of the annual consumption in the real and generated dataset. The generated dataset exhibits a reduced representation of profiles with annual consumptions above 5,000 kWh.

The leads to an underestimation of higher savings on the annual invoice due to the chosen sizing schemes of both PV and BESS components with respect to the consumer's annual consumption, which is evident in the tail of the distributions in Figure 9.11 (a). A higher overlap coefficient could reduce this mismatch, although possibly at a cost of performance on other downstream applications if other microbenchmarks score worse.

We can conclude that the synthetic dataset is able to reproduce the real data behavior for the benchmarks comprising the applicability of PV-BESS systems, e.g., the cases serving as proxy to benchmark the behavior of the peak demands and the interday and intraday variability. Furthermore, the generated data matches the performance of the real data for a wide variety of relative component sizes, illustrating its suitability for these applications. However, here we once more observe that the synthetic data is not yet able to capture the full diversity in the real data, as metadata is not taken into account during the generation and concatenation process.

9.3 Benchmarks at the aggregated level

Two case studies related to the HC of the distribution grid are considered as benchmarks at the aggregated level: a study related to the voltage drop due to the charging of electric vehicles, and a study related to the impact on the LV distribution grid caused by widespread integration of heat pumps.

Section 9.3.1 discusses the impact of EV charging, comparing the performance of the synthetic dataset to the real data. In contrast to the case studies at the individual level, we use this case study to only investigate the usability of the generated daily load profiles from Chapter 7. If the usability at the daily level needs to be validated, a case study where the use of (disjoint) daily profiles is a viable option is necessary. An EV HC study was deemed to be suitable, as the problems related to EV charging happen at a daily level. Furthermore, HC studies typically involve multiple runs with a Monte Carlo allocation of consumers on nodes in the network. Hence, the disjoint nature of the daily profiles on each node is of lesser importance for this benchmark.

Finally, Section 9.3.2 concludes the benchmarking analysis with a study focused on the integration of heat pumps. This benchmark considers the same LV distribution grid from Section 9.3.1, alongside a future scenario in which the residential building stock is sufficiently insulated and equipped with heat pumps. In this scenario, the aggregated peak demands serve as a validation metric for the synthetic data, which simultaneously allows us to quantify the impact of including high-frequency components in the synthetic profiles.

The HC study of Section 9.3.1 involves comparing voltage profiles across a network using both real and generated consumption data assigned to individual nodes. To determine the voltages at each node for each time step, a steady-state load flow calculation is performed, starting with the active power consumed at each bus.

Given that distribution networks tend to be radial networks which exhibit higher resistance-to-reactance (R/X) ratios than transmission lines, an iterative Backward-Forward Sweep (BFS) method is chosen to perform the load flow calculations at each time step [221, 222]. Conceptually, this method involves two main iterative steps after initializing all nodes on the network:

1. **Initialization**: The initial voltage at the substation, a slack bus, is set to its specified value, typically its nominal value of 1 p.u. with zero phase angle. All other bus voltages are initially set to this value as well.

2. **Backward sweep**: Starting at the end nodes of the network with no downstream connections, the branch currents flowing into each load are calculated. For a bus j connected to bus i (downstream to upstream), the current I_{ij} is calculated as follows, with S_j the complex power demand at bus j and V_j the voltage at bus j:

$$I_{ij} = \frac{S_j^*}{V_j^*} \tag{9.2}$$

   The current flowing through downstream branches can subsequently be summed to obtain the total current flowing into each node.

3. **Forward sweep**: Starting from the substation, the currents calculated in the backward sweep are used to recalculate the voltages for each bus downstream, with Z_{ij} as the impedance of the line between bus i and bus j:

$$V_j = V_i - I_{ij} Z_{ij} \tag{9.3}$$

4. **Check convergence**: The procedure stops after the mismatch of the calculated and the specified voltages at the substation is less than a convergence tolerance.

5. **Repeat steps**: If the calculated change is not within a specified tolerance, the newly calculated bus voltages are used as the initial voltages for the next iteration of the backward sweep, followed by a forward sweep.

9.3.1 EV hosting capacity

The considered network is a representative network for Flanders of a semi-urban area with a balanced mix between detached and semi-detached housing units. The considered grid is shown in Figure 9.12. A summary of its specifications is given in Table 9.2.

Each household is assigned a load profile and an EV charging profile. Similar to Section 7.7, only households with annual consumption between 2.750 kWh and 3.750 kWh are considered for the original dataset. These profiles are subsequently used to construct a dataset of similar size with synthetic load profiles.

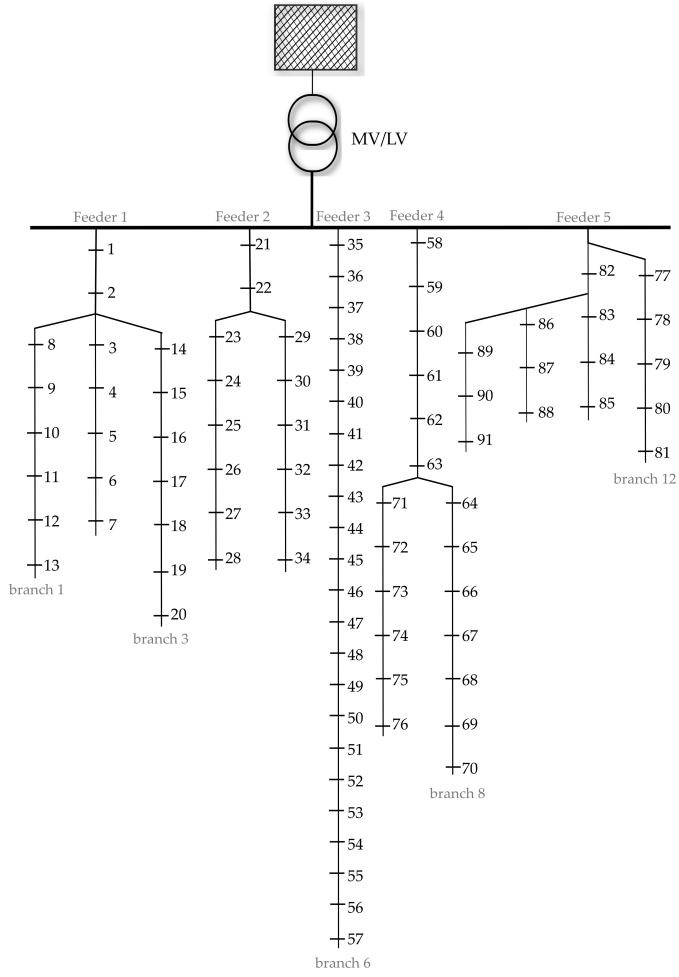

Figure 9.12: Line diagram of the considered LV distribution network.

Table 9.2: Summary of the considered grid specifications.

Description	Values
Transformer rating	250 kVA
Grid voltage	3x400 V + N
Distribution cable	EAXeVB 4x 150 mm^2
Connection cable	EXVB 4x 16 mm^2
No. of housing units	91
Max. feeder length	400 m
Distance to junction	[8,..., 15] m

The EV charging is assumed to occur at 3.7 kW, and the EV charging profiles are constructed based on the data sources and methodology presented in Section 4.3. These constructed charging profiles are subsequently superimposed on the load profile assigned to that household. Furthermore, it is assumed that the load distribution of the house-units is symmetrically connected to the distribution cable (i.e., housing units 1 is connected to L_1-N, 2 to L_2-N, 3 to L_3-N, 4 again to L_1-N etc.).

Simulations are performed within an OpenDSS-Python environment, therefore the distribution network is modeled in OpenDSS [223] while the actual power flow analysis is performed in Python through the OpenDSS COM interface, as presented in [60, 224]. The method adopted for the modeling of the cables is described in [225]. Results are obtained through a steady-state power flow analysis performed every timestep.

To compare the performance of the measured and synthetic dataset, the voltage level is calculated for every 15 minutes, for each individual household over the course of one year.

Figure 9.13 displays the resulting distributions of the voltage levels throughout the day, where the distributions contain the voltage levels of all 91 households and the 365 considered days of the year. The median value, 10^{th} and 90^{th} percentile are displayed in the figure together with the minimum and maximum values of the voltage. The performance is evaluated in Table 9.3 by the MAPE and MAE.

As seen in Figure 9.13 by visual inspection, as well as in Table 9.3, the synthetic dataset yields nearly identical results to the original data, proving its suitability for EV-related hosting capacity studies. The synthetic dataset adequately captures the simultaneity of the residential peak demands and the EV charging behavior.

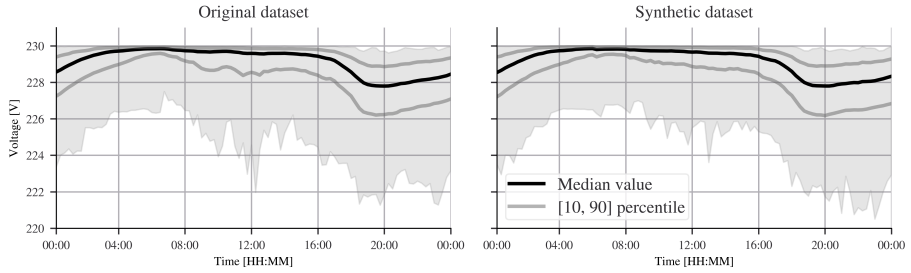

Figure 9.13: Comparison of the distributions of the daily voltage profile for the considered case study.

Table 9.3: Performance evaluation of the synthetic data compared to the original data in the voltage drop profile of the EV HC case study as shown in Figure 9.13. P_i denotes the i^{th} percentile of the distribution.

Metric	P_{10}	median	P_{90}
MAPE [%]	0.055	0.016	0.014
MAE [V]	0.127	0.037	0.031

9.3.2 HP hosting capacity

In the previous benchmark, the individual EV charging profiles for each day exhibited a high degree of non-simultaneity due to the stochastic nature of arrival and departure times, as well as the energy charged per session. In contrast, the consumption of heat pumps is characterized by a high degree of simultaneity. Therefore, for this validation, the voltage profile is not our primary focus, as the synthetic data accurately reproduce it. Instead, we concentrate on the LDC of the aggregated demand of 91 households equipped with air-source heat pumps (ASHPs).

To maintain consistency with the EV charging case study, we randomly selected 91 households. The methodology for constructing the ASHP demand profiles was detailed in Section 4.4. In this study, the 91 residential consumers are assumed to reside in dwellings of 240 m² with an E-level between 0 and 20, indicating a well-insulated building stock where heat pumps are expected to be most effective and prevalent. We then superimposed the electric load profiles of these heat pumps onto the residential smart meter data, both real and synthetic, to quantify differences in the resulting aggregated demand profile.

The purpose of this case study is to investigate whether the inclusion of high-frequency components in the synthetic data is necessary for accurately modeling aggregated demand. HF details refer to short-term fluctuations in energy usage, which can be significant at an individual level but may average out when data is aggregated across many households. By focusing on the LDCs of the synthetic data with and without HF details, we aim to assess the impact of these HF components on the overall demand patterns and determine their relevance in aggregated scenarios.

The LDCs of the aggregated demands are visualized in Figure 9.14. An excellent match between the real and synthetic data in the aggregated LDC can be observed over the course of the full year, regardless of the inclusion of HF components in the synthetic data. However, the inset plot displays the LDC for the highest 50 values of the aggregated demand. The synthetic data without HF components underestimates the maximum peak demand by 8.2 kW, or 4.2%, although this deviation is only observed for less than 0.01% of the time. In contrast, the synthetic data with HF components captures this coincident peak more accurately, with a deviation of only 1.7% between the real and synthetic data.

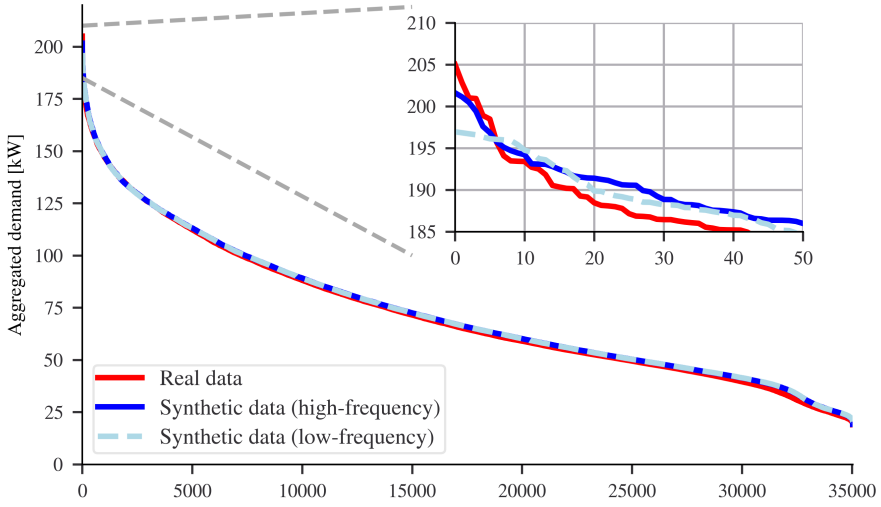

Figure 9.14: LDCs of the aggregated demand of 91 residential consumers equipped with ASHPs for the real (red) and synthetic (blue) data. The aggregated synthetic data with (full line) and without (dashed line) high-frequency components are included in the figure. The inset plot displays the first 50 values of the LDC.

9.4 Conclusion

In this chapter, we validated the constructed synthetic data through a series of downstream applications involving residential smart meter data. This approach contrasts with the validations conducted in Chapters 7 and 8, which focused solely on the similarity between the synthetic data and the original data. The validation process in this chapter followed a two-track approach, first focusing on downstream applications at the individual level and subsequently discussing applications at the aggregated level.

At the individual level, the validation was threefold. First, we compared the LDCs of the synthetic data to those of the real data to assess distributional similarity. While the general behavior was reproduced in the parametric distributions, the synthetic data was unable to replicate the step behavior observed in the LDCs. We suggested that limiting the generation to clusters of households with similar larger electrical appliances is necessary to reproduce this behavior. Second, we investigated whether the synthetic data is able to reproduce different elements of the electricity invoice, with a focus on the mean monthly peak demand behavior and the commodity cost under dynamic pricing. The synthetic data shows an excellent match for the general trend over a wide range of annual consumption values. However, they fail to capture the diversity present in the real data without metadata to guide the generation and concatenation process. Third, we demonstrated that the synthetic data can effectively replicate the performance of real data in applications involving both PV and PV-BESS systems. However, the need for metadata again became apparent in these downstream applications. Without metadata to obtain consistent interday behavior, the synthetic data failed to capture the diversity of the input samples.

At the aggregated level, our focus was on quantifying the impact of the widespread integration of low-carbon appliances, such as electric vehicles and heat pumps, on a typical Flemish LV network. The EV hosting capacity validation demonstrated that our proposed approach accurately reconstructs the impact of voltage drop caused by residential EV charging. Additionally, we investigated the necessity of including high-frequency components in the synthetic data for accurately modeling aggregated demand, specifically in the case study involving the integration of heat pumps. The load duration curves of the aggregate demand showed an excellent match between the real and synthetic data, even without the HF components. However, the aggregated peak demand was underestimated by 4.2%. If this level of accuracy is critical for an application, our results suggest the inclusion of HF details is necessary. Otherwise, the HF components at the individual level tend to average out when aggregating over many residential consumers.

The absence of metadata during the generation and linking process highlights a key limitation of the current validation. The input dataset used throughout this work remains limited and metadata had to be inferred, resulting in distinct clusters, as discussed in Chapter 6, that are too small to serve as effective inputs for the GAN-based generation of annual profiles in Chapter 8.

However, our validations in this chapter underscore the importance of incorporating metadata into the generation process at multiple stages. This inclusion is crucial in the linking stage to ensure that synthetic profiles exhibit interday variability that consistently aligns with the patterns observed in real data. Metadata provides context such as appliance usage patterns during heating and cooling seasons, which is necessary for producing realistic synthetic data. For example, without metadata, the generation process is unable to (i) reproduce the step in the LDCs of the synthetic data, and (ii) replicate the wide range of self-consumption values resulting from the addition of PV installations to residential consumers.

Despite this limitation stemming from the proof-of-concept research phase and the limited dataset, the results from the diverse benchmarks included in this chapter demonstrate that the proposed proof-of-concept for synthetic residential load modeling is able to capture the different temporal behaviors inherent to real smart meter data, leading to acceptable performance for downstream applications over a wide range of consumption values. This validates the potential of our proposed two-step methodology to support various downstream applications at both the individual and aggregated levels. These applications include the integration of low-carbon technologies such as electric vehicles and heat pumps, peak shaving initiatives by PV-BESS installations, as well as grid planning at the aggregated level.

10

Conclusions and perspectives

I may not have gone where I intended to go,
but I think I have ended up where I intended to be.
Douglas Adams

This thesis has developed a new methodology for generating synthetic residential load data over a full year that accurately reproduces the peak demand behavior. In this concluding chapter, we first contextualize the work presented in this dissertation. Following this, the key methodology and main findings are summarized, highlighting the contributions and implications of the research. Finally, we discuss recommendations for future improvements and potential directions for continuing the development of this methodology, aiming to improve its accuracy, applicability, and integration into practical applications.

10.1 Motivation and objectives

The residential sector is a significant contributor to greenhouse gas emissions, being responsible for approximately 20% of the total emissions in the European Union. Given the increasingly urgent need for decarbonization across all segments of society to mitigate the impacts of climate change, this sector faces a critical transformation within a relatively short time frame of less than three decades.

A widespread deployment of PV systems, coupled with a rapid transition to heat pumps and electric vehicles, is essential to reduce emissions. While these transitions are driven by the urgent need to decarbonize, it is crucial that distribution grids can accommodate the increasing number of these appliances. Grid operators are therefore actively reinforcing their networks, and various stakeholders are introducing economic incentives to encourage consumers to spread their energy consumption. Smart meters and digitalization play a key role in these efforts, enabling more effective grid management through demand-side response and flexibility initiatives.

Smart meters record consumption behavior at sub-hourly timescales, providing unprecedented insight into consumer habits. This detailed data has paved the way for innovative, data-driven digital business models, rapidly evolving into a billion-dollar market. However, due to privacy concerns, public access to residential smart meter data has quickly been restricted.

The research presented in this dissertation resides within the field of load modeling, a research discipline dedicated at leveraging privacy-sensitive smart meter data to produce synthetic datasets suitable for unrestricted sharing. Research within this domain endeavors to unlock the vast potential of smart meter data for innovative, data-driven research while upholding the privacy rights of individuals.

Traditional load modeling techniques typically rely on unsupervised machine learning to group consumers with comparable characteristics, followed by the averaging and smoothing of their smart meter data to produce synthetic load profiles. However, contemporary applications of smart meter data focusing on flexibility and demand response initiatives emphasize the need for information on both the timing and amplitude of individual peak demands, rendering traditional smoothened curves inadequate.

Given the complex interplay of factors that collectively shape the smart meter output of residential consumers, we have argued that accurate synthetic load models at this timescale and resolution cannot be achieved with a single modeling method. Recognizing this complexity, we advocated for the decoupling of the problem and the adoption of scale-appropriate modeling techniques for different timescales, each tailored to capture the unique dynamics inherent to the timescale under consideration.

To systematically address these methodological challenges and refine the primary research objective, we formulated several additional secondary research questions. These secondary objectives, listed on the following page, served as critical milestones, collectively contributing to achieving the primary research question of obtaining synthetic residential metering data with realistic peak demand properties.

Secondary research objectives

- ➢ To identify the use cases for granular smart meter data, both at the individual and aggregated level. These downstream applications will serve as benchmarks of the synthetic data throughout this dissertation.
- ➢ To define a rigorous mathematical framework to classify which values can be considered '*peak demands*'.
- ➢ To construct a feature set capable of incorporating the peak demand behavior, and to use this feature set in a clustering methodology to investigate and quantify the difference in peak demand behavior for different consumer categories.
- ➢ To construct daily load profiles with sufficiently stochastic peak demands.
- ➢ To generate annual profiles with realistic multiscale time correlations.

10.2 Summary of methodology and conclusions

Mathematical framework to identify peak demands

To achieve the first objective of this thesis, we utilized the concept of the Load Duration Curve (LDC). The LDC is derived by ordering power consumption measurements in descending order rather than chronologically. Historically, LDCs have been extensively used for network planning and power plant operations.

We began by validating a 5-parameter analytical expression for the LDC of individual residential consumers. This proposed equation is characterized by (i) a steep exponential decline representing stochastic peak demands, followed by (ii) a step reflecting the consistent concurrent consumption behavior of larger household appliances, before slowly transitioning into (iii) the residential base load.

The point of maximum curvature in the exponential decay of the LDC was identified as the threshold for peak demands. Any value above this unique consumer-specific threshold was classified as a peak demand. We derived an expression for this point of maximum curvature based on the LDC equation. Finally, applying this threshold to our dataset of residential consumers resulted in, on average, 2.8% of the consumption values being labeled as peak demands on an annual basis.

Constructing a feature set with peak-based properties

Armed with a rigorous method to define peak demands at the level of individual consumers, we subsequently constructed a new feature set designed to capture and quantify the stochastic behavior of these peak demands. Two types of features were developed: first, the fraction of total consumption occurring within specific daily or weekly periods; second, the fractions of peak demands occurring during these same periods.

The constructed feature set was used in a hierarchical clustering process to build 10 clusters from the dataset under consideration. The clustering algorithm produced compact clusters with clear connections to real-life applications involving residential peak demands, such as demand response initiatives and the applicability of battery storage systems for peak shaving. The timing of consumption was the primary factor driving the separation of the full dataset into three distinct clusters, aligning with the current synthetic load profiles provided by the Flemish regulator. Subsequently, the timing of the demand peaks emerged as the secondary factor for transitioning from the initial 3 clusters into 10 distinct clusters.

In the final analysis of this secondary research objective, the stochastic nature of peak demands was investigated by examining the relationship between consumption and the presence of peak demands in the same time period. The disproportionate presence of peak demands in certain periods was quantified, revealing a linear relationship between the fraction of consumption and peak demands in each time period. The spread in these results quantified the stochasticity of the peak demands, highlighting the limitations in the general applicability of the observed relations.

However, the constructed clusters demonstrated a clear connection to the predictability and variability of consumption and peak behavior, effectively reducing the stochasticity of these peak demands and their timing. This alignment underscores the robustness and coherence of the clustering approach, as it successfully grouped consumers with similar consumption patterns and peak demand behaviors.

The clusters and metadata resulting from this secondary research objective were used throughout the remainder of this work to highlight the necessity of incorporating metadata for producing realistic data. However, due to the size limitations of the initial dataset, the resulting clusters were relatively small, which posed challenges for performing cluster-specific synthetic data generation. Consequently, the data generation processes developed for both daily and annual levels as discussed in the following sections should be regarded as proof of concept rather than fully scalable solutions.

Daily load profiles through decomposition-recombination

The fourth research objective of this thesis was to explore the use of time-frequency analysis to construct daily load profiles, with a particular focus on addressing the stochastic nature of peak demands. Time-frequency analysis was explicitly proposed as a method because it allows us to decompose the load profiles into their constituent frequency components. By doing so, we can treat the stochastic peak demands as high-frequency components superimposed on a more slowly varying, low-frequency base load.

A multi-resolution analysis with the Haar wavelet as mother wavelet was used to decompose daily smart meter data in their low-frequency and high-frequency components. The high-frequency component of the load profile corresponding with one household can be rescaled and shifted, and subsequently superimposed on the approximated profile of a different household. This yields a synthetic load profile with a given daily peak demand and daily consumption. By sampling the two-dimensional distribution of (i) the daily consumption and (ii) the daily load factor of the original dataset, their relation is preserved in the resulting generated profiles.

The generated profiles were benchmarked with respect to the original dataset for the daily peak demand behavior. The seasonal behavior in the original data is preserved by limiting the generator to the daily level. Furthermore, the distribution of the synthetic daily peak demands showed a MAE of 0.09 kW, corresponding to a MAPE of 4.1%, compared to the original dataset.

A limitation of the proposed methodology was identified, concerning the reconstruction of daily load profiles with a high peak demand. These profiles tend to have a significant contribution of the high-frequency component to their peaks, which leads to fewer other households which can be used for constructing a synthetic profile with a similar peak demand. This highlighted the trade-off between anonymization of the privacy-sensitive data and the computational process involved in the load profile generation.

Annual load profiles through Generative Adversarial Networks

At the annual level, a different modeling technique is suggested, as annual consumption profiles exhibit more consistent and correlated behaviors over longer timescales compared to daily load profiles. The complex interplay of these behaviors across interday, weekly, monthly, and seasonal levels necessitates the use of advanced modeling techniques. Generative Adversarial Networks (GANs) are particularly well-suited for this task, as they have been shown to reproduce the patterns and correlations present in long-term data.

After reviewing literature on time series generation, we opted for the DoppelGANger architecture for its established capability to capture long-term time correlations and its capacity to generate short-term intraday variations efficiently, particularly at the weekly level. To evaluate the output of the trained GANs, we introduced three application-specific microbenchmarks. These benchmarks assessed the synthetic dataset's fidelity in terms of (i) annual consumption values, (ii) daily consumption variability, and (iii) time correlations within the dataset.

In a sensitivity analysis of five key parameters influencing the generation process, 6,750 GAN models were trained. Through this in-depth analysis, we managed to identify a set of hyperparameters that effectively eliminated mode collapse in both the frequency and amplitude domains, where the difference in learning rates between the generator and discriminator proved to be instrumental.

Drawing from the insights gained in the sensitivity analysis and the performance of the converged model on the three microbenchmarks, we affirmed that DoppelGANger can generate annual load profiles of daily consumptions that (i) align with the distribution of total annual consumption and (ii) exhibit similar variability. The batched generation ensures accurate weekly correlations, while the seasonal correlation remains robust due to the smaller number of individual RNN generation steps.

Validation through downstream applications

In the final segment of this thesis, we evaluated the performance of our constructed synthetic dataset across a diverse array of practical downstream applications, selected through a literature review on the uses of residential smart meter data. The annual profiles generated through GANs were combined with daily profiles obtained through a wavelet-based decomposition-recombination scheme to obtain annual load profiles with high-frequency variability and realistic temporal short-term and long-term dynamics.

At the individual level, the validation was threefold. We compared the distributional similarity of the 5 parameters describing the individual LDCs, followed by a comparison of the mean monthly peak demand as proxy for capacity-based grid tariffs. Finally, we demonstrated that the synthetic data can effectively replicate the performance of real data in applications involving both PV and PV-BESS systems. However, without metadata to guide the generation and linking process to obtain consistent intraday behavior, the synthetic data failed to capture the diversity of the input samples.

At the aggregated level, our focus was on quantifying the impact of the widespread integration of low-carbon appliances, such as electric vehicles and heat pumps. The EV hosting capacity validation demonstrated that our proposed approach accurately reconstructs the impact of voltage drop caused by residential EV charging. Additionally, we considered a case study involving heat pumps to investigate whether it is necessary to include high-frequency components in the synthetic data for accurately modeling aggregated demand. Indeed, the HF components at the individual level tend to average out when aggregating over many residential consumers. The load duration curves of the aggregate demand showed an excellent match between the real and synthetic data, even without the high-frequency components. However, our results showed that the aggregated peak demand was underestimated by 4.2% by omitting the high-frequency components.

Overall, the outcomes of this thesis confirm the viability of the proposed two-step methodology for generating synthetic, high-frequency annual load models as a viable alternative for real, privacy-sensitive residential smart meter data across a broad spectrum of practical applications. Nonetheless, it is important to acknowledge the limitations identified throughout this dissertation, which stem from its proof-of-concept nature. Moving forward, we expect that improving the quality of the metadata, the size of the training data, labeling the time series for GAN training, and more diversification in the microbenchmarks evaluating the GAN models will further increase the coherence and fidelity of the synthetic data. These recommendations will be further elaborated upon in the following section.

10.3 Recommendations and further research

The introduced methodology for generating synthetic smart meter data was trained and validated on a limited dataset, consisting of one year of metering data from 1,422 consumers. An evident area for improvement is to use a larger dataset that spans multiple years and includes a diverse range of residential consumers for the training process. A larger dataset is particularly important when partitioning into clusters, as clusters need to be of sufficient size for effective GAN-based modeling of annual profiles of daily consumption. Additionally, incorporating multiple years of data would enable better integration of seasonal effects into the synthetic data, increasing its accuracy and applicability. Consequently, discussions are ongoing with Fluvius, the data manager of smart meter data in Flanders, to obtain a larger dataset spanning a longer time period, as this would allow us to advance the introduced methodology beyond its current proof-of-concept phase.

Two future research directions for the stochastic generation of daily load profiles, as introduced in Chapter 7, are currently being considered. First, we currently use a time resolution of 2 hours for the low-frequency approximations, with each consumer having the same LF time blocks (e.g., 00:00 to 02:00, 02:00 to 04:00, etc.). A refinement to this approach is to introduce more diversity by varying the start time for the LF approximation for each consumer. Second, we are considering other time resolutions that align better with specific practical applications, such as the time-of-use periods for grid tariffs, or the 4-hour time blocks used for aFRR frequency response.

Furthermore, the main limitations of the methodology were identified during the GAN-based annual load modeling presented in Chapter 8. As discussed, evaluating the fidelity of GAN outputs in a rigorous manner is difficult with no separate objective function for the generator to evaluate the generated data sequences, nor standard agreed-upon evaluation criteria. We introduced three microbenchmarks to evaluate the output dataset. However, there is a degree of arbitrariness in the selection of these microbenchmarks. They were chosen to be application-specific, gauging properties relevant in the context of smart meter data analytics.

However, several areas for improvement were identified. First, in the autocorrelation microbenchmark, minor discrepancies were observed within time lags of up to two weeks. These deviations were traced back to the challenge of simultaneously capturing both the weekly calendrical cadence and variable behavior during holiday periods. We suggest that labeling the datetime series to include information on holidays and school closures would enable the GAN to better incorporate this information, thereby improving the quality of the synthetic data. Second, the presence of electric heating significantly skews the autocorrelation depending on whether it is the heating or cooling season. Limiting the generation process to a cluster containing only households with electric heating would make this microbenchmark more informative. Additionally, a promising future research direction is to split the generation process between heating and non-heating seasons.

A final recommendation is related to the downstream applications selected for this thesis. While only a limited subset of the numerous potential applications for residential smart meter data was chosen for the validation of the synthetic data, it is evident that additional use cases are necessary. This broader validation will help confirm the robustness and versatility of the synthetic data across a wider range of practical applications.

Beyond future research directions, we are considering the dissemination of our work through the development of an open-source Python package that incorporates all aspects of the proposed two-step methodology for synthetic smart data generation. Making our research more easily accessible in this manner would enable broader use and further refinement by both the research community and industry practitioners. However, this endeavor would only be pursued subsequent to validating our proof-of-concept methodology on a larger and more contemporary dataset of residential metering data.

Part II

Published Papers

A

List of Publications

Publications in international peer–reviewed journals

1. **A novel feature set for low-voltage consumers, based on the temporal dependence of consumption and peak demands**
   R. Claeys, H. Azaioud, R. Cleenwerck, J. Knockaert, J. Desmet
   Energies, **2021**, 14 (1) 139
   IF: 3.252
2. **A low-voltage DC backbone with aggregated RES and BESS : benefits compared to a traditional low-voltage AC system**
   H. Azaioud, R. Claeys, J. Knockaert, L. Vandevelde, J. Desmet
   Energies, **2021**, 14 (5) 140
   IF: 3.252
3. **An approach to the impedance modelling of low-voltage cables in digital twins**
   R. Cleenwerck, H. Azaioud, R. Claeys, T. Coosemans, J. Knockaert, J. Desmet
   Electric Power Systems Research, **2022**, 210, 108075
   IF: 3.414
4. **Stochastic generation of residential load profiles with realistic variability based on wavelet-decomposed smart meter data**
   R. Claeys, R. Cleenwerck, J. Knockaert, J. Desmet
   Applied Energy, **2023**, 350, 121750
   IF: 11.446

5. **Capturing multiscale temporal dynamics in synthetic residential load profiles through Generative Adversarial Networks (GANs)**
   R. Claeys, R. Cleenwerck, J. Knockaert, J. Desmet
   Applied Energy, **2024**, 360, 122831
   IF: 11.446

Conference contributions

1. **Assessing the influence of the aggregation level of residential consumers through load duration curves**
   R. Claeys, T. Delerue, J. Desmet
   IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe 2019), September 29 – October 2, 2019
2. **A data-driven approach to assessing and improving stochastic residential load modeling for district-Level simulations and PV integration**
   R. Claeys, C. Protopapadaki, D. Saelens, J. Desmet
   International Conference on Probabilistic Methods Applied to Power Systems (PMAPS 2020), Liege (Online), Belgium, August 19–August 21, 2020
3. **Sizing BESS for a peak shaving and valley filling control strategy for residential consumers based on their load-duration curves**
   R. Claeys, G. De Greve, J. Knockaert, J. Desmet
   Conference on Sustainable Energy Supply and Energy Storage Systems (NEIS 2021), Hamburg (Online), Germany, September 13–14, 2021
4. **Peak demand dynamics of low-voltage consumers under aggregation and its impact on upstream PV injection**
   R. Claeys, H. Azaioud, J. Desmet
   26th International Conference and Exhibition on Electricity Distribution (CIRED 2021), Online, September 20–23, 2021
5. **Self-sufficiency and lifetime improvement of community BESS on an LVDC backbone compared to individual BESS**
   H. Azaioud, R. Claeys, J. Knockaert, L. Vandevelde, J. Desmet
   27th International Conference on Electricity Distribution (CIRED 2023), Rome, Italy, June 12–15, 2023
6. **Volumetric or capacity-based grid tariffs: a case study for residential consumers in Flanders**
   R. Claeys, R. Cleenwerck, J. Knockaert, L. Vandevelde, J. Desmet
   27th International Conference on Electricity Distribution (CIRED 2023), Rome, Italy, June 12–15, 2023
7. **Optimizing Grid Integration of Heat Pumps for Balancing Hosting Capacity and Customer Satisfaction** (Accepted)
   K. Jalilpoor, M. Asadi, R. Claeys, J. Desmet
   7th International Conference on Smart Energy Systems and Technologies (SEST 2024), Torino, Italy, September 10–12, 2024

B

List of Software Packages

In the context of this PhD dissertation, one simulation model was specifically developed using results of our research. Furthermore, several software packages were used to facilitate our simulations. Below we first provide a description of the developed software model, after which an alphabetical overview is given of software packages which have been used extensively without alterations.

Developed software model

Simulator home batteries

In 2021, EELab/Lemcko developed a publicly available simulator to assess the economic feasibility of residential battery systems, commissioned by the Flemish Energy and Climate Agency. This simulator incorporated elements from this PhD thesis related to the characterization of residential consumers and peak shaving algorithms.

From its launch in April 2021 until its retirement in March 2024, the home battery simulator was visited 265,092 times. This underscores the significant market potential for accessible, data-driven, digital tools. The high engagement rate highlights the growing interest among consumers in sustainable energy solutions and the importance of providing user-friendly and objective resources to guide their decision-making.

Used software packages

`Gretel Synthetics`

Gretel Synthetics is a library designed for generating synthetic data using advanced machine learning techniques. It is part of the Gretel.ai platform, which focuses on privacy-preserving machine learning and synthetic data generation. We used the implemented version of the DoppelGANger algorithm as optimized by Gretel.ai as GAN architecturue of choice throughout Chapter 8.

`lmfit`

The fitting procedure described in Chapter 5 to fit an analytic expression of the LDC to the raw data was performed through the lmfit package. This package is designed for nonlinear optimization and curve fitting, providing a high-level interface to implement custom equation shapes and parameter constraints.

`PyWavelets`

We made extensive use of PyWavelets (pywt) in Chapter 7. The PyWavelets package is library dedicated to wavelet transforms in Python. As discussed in Section 7.4, we made extensive use of the 1D Discrete Wavelet Transform and wavelet-based decompositions to decompose the high-frequency daily load profiles in a low-frequency approximations and a high-frequent oscillating detail function.

`SciPy`

SciPy is a package that is widely used in the scientific community for a wide array of tasks involving scientific computing. In Chapter 6, we made use of the agglomerative clustering algorithm with Ward's linkage method as implemented. Furthermore, the Wasserstein-1 distance was used in the same chapter as distance metric for the distributional analysis.

Bibliography

[1] Eurostat. Energy consumption in households. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy_consumption_in_households, 2023. Accessed: 2024-04-11.

[2] Belgian Interregional Environment Agency (CELINE - IRCEL). National inventory report: Belgium's greenhouse gas inventory (1990-2021), 2023.

[3] European Commission - Directorate-General for Environment - Unit 01: Strategy, Digitalization, Better Regulation and Economic Analysis Unit. A toolbox for reforming environmentally harmful subsidies in europe: detailed annexes, 2022.

[4] Katherine Calvin, Dipak Dasgupta, Gerhard Krinner, Aditi Mukherji, Peter W. Thorne, Christopher Trisos, José Romero, Paulina Aldunce, Ko Barret, Gabriel Blanco, William W.L. Cheung, Sarah L. Connors, Fatima Denton, Aïda Diongue-Niang, David Dodman, Matthias Garschagen, Oliver Geden, Bronwyn Hayward, Christopher Jones, Frank Jotzo, Thelma Krug, Rodel Lasco, Yune-Yi Lee, Valérie Masson-Delmotte, Malte Meinshausen, Katja Mintenbeck, Abdalah Mokssit, Friederike E.L. Otto, Minal Pathak, Anna Pirani, Elvira Poloczanska, Hans-Otto Pörtner, Aromar Revi, Debra C. Roberts, Joyashree Roy, Alex C. Ruane, Jim Skea, Priyadarshi R. Shukla, Raphael Slade, Aimée Slangen, Youba Sokona, Anna A. Sörensson, Melinda Tignor, Detlef van Vuuren, Yi-Ming Wei, Harald Winkler, Panmao Zhai, Zinta Zommers, Jean-Charles Hourcade, Francis X. Johnson, Shonali Pachauri, Nicholas P. Simpson, Chandni Singh, Adelle Thomas, Edmond Totin, Andrés Alegría, Kyle Armour, Birgit Bednar-Friedl, Kornelis Blok, Guéladio Cissé, Frank Dentener, Siri Eriksen, Erich Fischer, Gregory Garner, Céline Guivarch, Marjolijn Haasnoot, Gerrit Hansen, Mathias Hauser, Ed Hawkins, Tim Hermans, Robert Kopp, Noëmie

Leprince-Ringuet, Jared Lewis, Debora Ley, Chloé Ludden, Leila Niamir, Zebedee Nicholls, Shreya Some, Sophie Szopa, Blair Trewin, Kaj-Ivar van der Wijst, Gundula Winter, Maximilian Witting, Arlene Birt, and Meeyoung Ha. *IPCC, 2023: Climate Change 2023: Synthesis Report, Summary for Policymakers. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.* July 2023. doi:10.59327/ipcc/ar6-9789291691647.001. URL http://dx.doi.org/10.59327/IPCC/AR6-9789291691647.001.

[5] Flemish Energy and Climate Agency. Installed PV capacity in Flanders (dashboard), 2024. https://apps.energiesparen.be/energiekaart/vlaanderen/zonnepanelen [Accessed: 01/05/2024].

[6] Flemish Government. Draft Flemish Energy and Climate Plan 2021-2030, 2023.

[7] Fluvius System Operator cv. Investment Plan 2024-2033 (version June 2023), 2023.

[8] Wim Clymans, Karolien Vermeiren, Dirk Vanden Boer, Leen Van Esch, Frank Meinke-Hubeny, Louis Godon, Arnaud Schils, Joris Lemmens, François Duchene, Geert Smet, and Joris Van den Bergh. BREGILAB WP3 RES Generation: Wind & PV deployment evolution and availability factor. , 2022.

[9] Statistiek Vlaanderen. Passenger Fleet Flanders, 2023. https://www.vlaanderen.be/statistiek-vlaanderen/mobiliteit/personenwagenpark [Accessed: 01/05/2024].

[10] Statistiek Vlaanderen. Housing stock of the Flemish Region, 2023. https://www.vlaanderen.be/statistiek-vlaanderen/bouwen-en-wonen/woningvoorraad [Accessed: 01/05/2024].

[11] International Energy Agency (IEA). Global early-stage venture capital investments in digital energy-efficiency and demand-side flexibility start-ups, by type of new business model, 2015-2021 – charts – data & statistics, 2021. URL https://www.iea.org/data-and-statistics/charts/global-early-stage-venture-capital-investments-in-digital-energy-efficiency-and-demand-side-flexibility-start-ups-by-type-of-new-business-model-2015-2021.

[12] Hakim Azaioud, Jan Desmet, and Lieven Vandevelde. Benefit Evaluation of PV Orientation for Individual Residential Consumers. *En-*

ergies, 13(19), 2020. ISSN 1996-1073. doi:10.3390/en13195122. URL https://www.mdpi.com/1996-1073/13/19/5122.

[13] Altti Meriläinen, Pietari Puranen, Antti Kosonen, and Jero Ahola. Optimization of rooftop photovoltaic installations to maximize revenue in Finland based on customer class load profiles and simulated generation. *Solar Energy*, 240:422–434, 2022. ISSN 0038-092X. doi:https://doi.org/10.1016/j.solener.2022.05.057. URL https://www.sciencedirect.com/science/article/pii/S0038092X22004030.

[14] Lai Zhou, Yongjun Zhang, Xiaoming Lin, Canbing Li, Zexiang Cai, and Ping Yang. Optimal Sizing of PV and BESS for a Smart Household Considering Different Price Mechanisms. *IEEE Access*, 6:41050–41059, 2018. doi:10.1109/ACCESS.2018.2845900.

[15] Joern Hoppmann, Jonas Volland, Tobias S. Schmidt, and Volker H. Hoffmann. The economic viability of battery storage for residential solar photovoltaic systems – a review and a simulation model. *Renewable and Sustainable Energy Reviews*, 39:1101–1118, 2014. ISSN 1364-0321. doi:https://doi.org/10.1016/j.rser.2014.07.068. URL https://www.sciencedirect.com/science/article/pii/S1364032114005206.

[16] V. Muenzel, I. Mareels, J. de Hoog, A. Vishwanath, S. Kalyanaraman, and A. Gort. PV generation and demand mismatch: Evaluating the potential of residential storage. In *2015 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pages 1–5, 2015. doi:10.1109/ISGT.2015.7131849.

[17] Yijie Zhang, Tao Ma, and Hongxing Yang. A review on capacity sizing and operation strategy of grid-connected photovoltaic battery systems. *Energy and Built Environment*, 5(4):500–516, 2024. ISSN 2666-1233. doi:https://doi.org/10.1016/j.enbenv.2023.04.001. URL https://www.sciencedirect.com/science/article/pii/S2666123323000156.

[18] B. Yildiz, J.I. Bilbao, J. Dore, and A.B. Sproul. Recent advances in the analysis of residential electricity consumption and applications of smart meter data. *Applied Energy*, 208:402–427, 2017. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2017.10.014. URL https://www.sciencedirect.com/science/article/pii/S0306261917314265.

[19] T. Beck, H. Kondziella, G. Huard, and T. Bruckner. Assessing the influence of the temporal resolution of electrical load and pv generation profiles on self-consumption and sizing of pv-battery systems. *Applied Energy*, 173:331–342, 2016. ISSN 0306-

2619. doi:https://doi.org/10.1016/j.apenergy.2016.04.050. URL https://www.sciencedirect.com/science/article/pii/S0306261916305104.

[20] Sunliang Cao and Kai Sirén. Impact of simulation time-resolution on the matching of pv production and household electric demand. *Applied Energy*, 128:192–208, 2014. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2014.04.075. URL https://www.sciencedirect.com/science/article/pii/S0306261914004383.

[21] Andrew Wright and Steven Firth. The nature of domestic electricity-loads and effects of time averaging on statistics and on-site generation calculations. *Applied Energy*, 84(4):389–403, 2007. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2006.09.008. URL https://www.sciencedirect.com/science/article/pii/S0306261906001097.

[22] Marc Beaudin and Hamidreza Zareipour. Home energy management systems: A review of modelling and complexity. *Renewable and Sustainable Energy Reviews*, 45:318–335, 2015. ISSN 1364-0321. doi:https://doi.org/10.1016/j.rser.2015.01.046. URL https://www.sciencedirect.com/science/article/pii/S1364032115000568.

[23] Khalid Abdulla, Julian de Hoog, Valentin Muenzel, Frank Suits, Kent Steer, Andrew Wirth, and Saman Halgamuge. Optimal operation of energy storage systems considering forecasts and battery degradation. *IEEE Transactions on Smart Grid*, 9(3):2086–2096, 2018. doi:10.1109/TSG.2016.2606490.

[24] Faeza Hafiz, M. A. Awal, Anderson Rodrigo de Queiroz, and Iqbal Husain. Real-Time Stochastic Optimization of Energy Storage Management Using Deep Learning-Based Forecasts for Residential PV Applications. *IEEE Transactions on Industry Applications*, 56(3):2216–2226, 2020. doi:10.1109/TIA.2020.2968534.

[25] Sara Abedi and Soongeol Kwon. Rolling-horizon optimization integrated with recurrent neural network-driven forecasting for residential battery energy storage operations. *International Journal of Electrical Power & Energy Systems*, 145:108589, 2023.

[26] Sima Aznavi, Poria Fajri, Reza Sabzehgar, and Arash Asrari. Optimal management of residential energy storage systems in presence of intermittencies. *Journal of Building Engineering*, 29:101149, 2020.

[27] Sangyoon Lee and Dae-Hyun Choi. Reinforcement learning-based energy management of smart home with rooftop solar photovoltaic

system, energy storage system, and home appliances. *Sensors*, 19(18): 3937, 2019.

[28] Zhiqiang Wan, Hepeng Li, and Haibo He. Residential energy management with deep reinforcement learning. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2018. doi:10.1109/IJCNN.2018.8489210.

[29] I. P. Panapakidis, M. C. Alexiadis, and G. K. Papagiannis. Load profiling in the deregulated electricity markets: A review of the applications. In *2012 9th International Conference on the European Energy Market*, pages 1–8, 2012. doi:10.1109/EEM.2012.6254762.

[30] Elisavet Proedrou. A comprehensive review of residential electricity load profile models. *IEEE Access*, 9:12114–12133, 2021. doi:10.1109/ACCESS.2021.3050074.

[31] Gianfranco Chicco. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy*, 42(1):68–80, 2012. ISSN 0360-5442. doi:https://doi.org/10.1016/j.energy.2011.12.031. URL https://www.sciencedirect.com/science/article/pii/S0360544211008565. 8th World Energy System Conference, WESC 2010.

[32] Yi Wang, Qixin Chen, Tao Hong, and Chongqing Kang. Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Transactions on Smart Grid*, 10(3):3125–3148, 2019. doi:10.1109/TSG.2018.2818167.

[33] Ramon Granell, Colin J. Axon, and David C. H. Wallom. Impacts of raw data temporal resolution using selected clustering methods on residential electricity load profiles. *IEEE Transactions on Power Systems*, 30(6):3217–3224, 2015. doi:10.1109/TPWRS.2014.2377213.

[34] Fanidhar Dewangan, Almoataz Y. Abdelaziz, and Monalisa Biswal. Load forecasting models in smart grid using smart meter information: A review. *Energies*, 16(3), 2023. ISSN 1996-1073. doi:10.3390/en16031404. URL https://www.mdpi.com/1996-1073/16/3/1404.

[35] Corentin Kuster, Yacine Rezgui, and Monjur Mourshed. Electrical load forecasting models: A critical systematic review. *Sustainable Cities and Society*, 35:257–270, 2017. ISSN 2210-6707. doi:https://doi.org/10.1016/j.scs.2017.08.009. URL https://www.sciencedirect.com/science/article/pii/S2210670717305899.

[36] Bastian Dietrich, Jessica Walther, Matthias Weigold, and Eberhard Abele. Machine learning based very short term load forecasting of machine tools. *Applied Energy*, 276:115440, 2020. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2020.115440. URL https://www.sciencedirect.com/science/article/pii/S0306261920309521.

[37] Che Guan, Peter B. Luh, Laurent D. Michel, Yuting Wang, and Peter B. Friedland. Very short-term load forecasting: Wavelet neural networks with data pre-filtering. *IEEE Transactions on Power Systems*, 28(1):30–41, 2013. doi:10.1109/TPWRS.2012.2197639.

[38] Nazih Abu-Shikhah and Fawwaz Elkarmi. Medium-term electric load forecasting using singular value decomposition. *Energy*, 36(7):4259–4271, 2011. ISSN 0360-5442. doi:https://doi.org/10.1016/j.energy.2011.04.017. URL https://www.sciencedirect.com/science/article/pii/S0360544211002660.

[39] Nima Amjady and Ali Daraeepour. Midterm demand prediction of electrical power systems using a new hybrid forecast technique. *IEEE Transactions on Power Systems*, 26(2):755–765, 2011. doi:10.1109/TPWRS.2010.2055902.

[40] Math H. J. Bollen and Sarah K. Rönnberg. Hosting capacity of the power grid for renewable electricity production and new large consumption equipment. *Energies*, 10(9), 2017. ISSN 1996-1073. doi:10.3390/en10091325. URL https://www.mdpi.com/1996-1073/10/9/1325.

[41] Mohammad Zain ul Abideen, Omar Ellabban, and Luluwah Al-Fagih. A review of the tools and methods for distribution networks' hosting capacity calculation. *Energies*, 13(11), 2020. ISSN 1996-1073. doi:10.3390/en13112758. URL https://www.mdpi.com/1996-1073/13/11/2758.

[42] Arpan Koirala, Tom Van Acker, Reinhilde D'hulst, and Dirk Van Hertem. Hosting capacity of photovoltaic systems in low voltage distribution systems: A benchmark of deterministic and stochastic approaches. *Renewable and Sustainable Energy Reviews*, 155:111899, 2022. ISSN 1364-0321. doi:https://doi.org/10.1016/j.rser.2021.111899. URL https://www.sciencedirect.com/science/article/pii/S1364032121011655.

[43] Mansoor Alturki, Amin Khodaei, Aleksi Paaso, and Shay Bahramirad. Optimization-based distribution grid hosting capacity cal-

culations. *Applied Energy*, 219:350–360, 2018. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2017.10.127. URL https://www.sciencedirect.com/science/article/pii/S030626191731574X.

[44] Rong-Ceng Leou, Chun-Lien Su, and Chan-Nan Lu. Stochastic analyses of electric vehicle charging impacts on distribution network. *IEEE Transactions on Power Systems*, 29(3):1055–1063, 2013.

[45] Matthew K. Gray and Walid G. Morsi. Power quality assessment in distribution systems embedded with plug-in hybrid and battery electric vehicles. *IEEE Transactions on Power Systems*, 30(2):663–671, 2015. doi:10.1109/TPWRS.2014.2332058.

[46] Christina Protopapadaki and Dirk Saelens. Heat pump and PV impact on residential low-voltage distribution grids as a function of building and district properties. *Applied Energy*, 192:268–281, 2017. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2016.11.103. URL https://www.sciencedirect.com/science/article/pii/S0306261916317329.

[47] Calum Edmunds, Stuart Galloway, James Dixon, Waqquas Bukhsh, and Ian Elders. Hosting capacity assessment of heat pumps and optimised electric vehicle charging on low voltage networks. *Applied Energy*, 298:117093, 2021. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2021.117093. URL https://www.sciencedirect.com/science/article/pii/S0306261921005420.

[48] M.K. Gray and W.G. Morsi. Probabilistic quantification of voltage unbalance and neutral current in secondary distribution systems due to plug-in battery electric vehicles charging. *Electric Power Systems Research*, 133:249–256, 2016. ISSN 0378-7796. doi:https://doi.org/10.1016/j.epsr.2015.12.022. URL https://www.sciencedirect.com/science/article/pii/S0378779615003995.

[49] Farhad Shahnia, Ritwik Majumder, Arindam Ghosh, Gerard Ledwich, and Firuz Zare. Voltage imbalance analysis in residential low voltage distribution networks with rooftop PVs. *Electric Power Systems Research*, 81(9):1805–1814, 2011. ISSN 0378-7796. doi:https://doi.org/10.1016/j.epsr.2011.05.001. URL https://www.sciencedirect.com/science/article/pii/S0378779611001040.

[50] Mohammad Zain ul Abideen, Omar Ellabban, and Luluwah Al-Fagih. A review of the tools and methods for distribution networks' hosting capacity calculation. *Energies*, 13(11), 2020. ISSN 1996-1073. doi:10.3390/en13112758. URL https://www.mdpi.com/1996-1073/13/11/2758.

[51] Enock Mulenga, Math H.J. Bollen, and Nicholas Etherden. A review of hosting capacity quantification methods for photovoltaics in low-voltage distribution grids. *International Journal of Electrical Power & Energy Systems*, 115:105445, 2020. ISSN 0142-0615. doi:https://doi.org/10.1016/j.ijepes.2019.105445. URL https://www.sciencedirect.com/science/article/pii/S0142061519306490.

[52] Tri Kurniawan Wijaya, Matteo Vasirani, Samuel Humeau, and Karl Aberer. Cluster-based aggregate forecasting for residential electricity demand using smart meter data. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 879–887, 2015. doi:10.1109/BigData.2015.7363836.

[53] Dejan Ilić, Per Goncalves da Silva, Stamatis Karnouskos, and Malte Jacobi. Impact assessment of smart meter grouping on the accuracy of forecasting algorithms. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 673–679, 2013.

[54] Hakim Azaioud, Robbert Claeys, Jos Knockaert, Lieven Vandevelde, and Jan Desmet. A Low-Voltage DC Backbone with Aggregated RES and BESS: Benefits Compared to a Traditional Low-Voltage AC System. *Energies*, 14(5), 2021. ISSN 1996-1073. doi:10.3390/en14051420. URL https://www.mdpi.com/1996-1073/14/5/1420.

[55] R. Claeys, H. Azaioud, and J. Desmet. Peak demand dynamics of low-voltage consumers under aggregation and its impact on upstream PV injection. In *CIRED 2021 - The 26th International Conference and Exhibition on Electricity Distribution*, volume 2021, pages 2153–2157, 2021. doi:10.1049/icp.2021.1718.

[56] Mike B Roberts, Anna Bruce, and Iain MacGill. PV for apartment buildings: Which side of the meter? In *Asia Pacific Solar Research Conference, Melbourne*, 2017.

[57] H Azaioud, R Claeys, J Knockaert, L Vandevelde, and J Desmet. Self-sufficiency and lifetime improvement of community BESS on an LVDC backbone compared to individual BESS. In *27th International Conference on Electricity Distribution (CIRED 2023)*, volume 2023, pages 2099–2103. IET, 2023.

[58] Edward Barbour, David Parra, Zeyad Awwad, and Marta C. González. Community energy storage: A smart choice for the smart grid? *Applied Energy*, 212:489–497, 2018. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2017.12.056. URL https://www.sciencedirect.com/science/article/pii/S0306261917317713.

[59] Jonas Schlund, Noah Pflugradt, David Steber, Urs Muntwyler, and Reinhard German. Benefits of virtual community energy storages compared to individual batteries based on behaviour based synthetic load profiles. In *2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*, pages 1–6, 2018. doi:10.1109/ISGTEurope.2018.8571506.

[60] Rémy Cleenwerck, Hakim Azaioud, Maarten Messagie, Jos Knockaert, Thierry Coosemans, and Jan Desmet. Enabling EV Charging by introducing LVDC Backbones in Low-Voltage Distribution Networks. In *35th Electric Vehicle Symposium (EVS35)*, pages 1–11, 2022.

[61] Rémy Cleenwerck, Hakim Azaioud, Majid Vafaeipour, Thierry Coosemans, and Jan Desmet. Impact Assessment of Electric Vehicle Charging in an AC and DC Microgrid: A Comparative Study. *Energies*, 16(7), 2023. ISSN 1996-1073. doi:10.3390/en16073205. URL https://www.mdpi.com/1996-1073/16/7/3205.

[62] European Court of Human Rights and Council of the European Union. European Convention on Human Rights, 2021. URL https://www.echr.coe.int/documents/d/echr/Convention_ENG.

[63] European Parliament and Council of the European Union. Charter of Fundamental Rights of the European Union, 2012. URL https://eur-lex.europa.eu/eli/treaty/char_2012/oj.

[64] European Parliament and Council of the European Union. Directive (EU) 2019/944 of the European Parliament and of the Council of 5 June 2019 on common rules for the internal market for electricity and amending Directive 2012/27/EU, 2019. URL https://eur-lex.europa.eu/eli/dir/2019/944/2022-06-23.

[65] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016. URL https://data.europa.eu/eli/reg/2016/679/oj.

[66] Francesco Fusco, Michael Wurst, and JiWon Yoon. Mining residential household information from low-resolution smart meter data. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 3545–3548, 2012.

[67] Chinthaka Dinesh, Buddhika W. Nettasinghe, Roshan Indika Godaliyadda, Mervyn Parakrama B. Ekanayake, Janaka Ekanayake, and Janaka V. Wijayakulasooriya. Residential appliance identification based on spectral information of low frequency smart meter measurements. *IEEE Transactions on Smart Grid*, 7(6):2781–2792, 2016. doi:10.1109/TSG.2015.2484258.

[68] Markus Weiss, Adrian Helfenstein, Friedemann Mattern, and Thorsten Staake. Leveraging smart meter data to recognize home appliances. In *2012 IEEE International Conference on Pervasive Computing and Communications*, pages 190–197, 2012. doi:10.1109/PerCom.2012.6199866.

[69] Thomas Bier, Djaffar Ould Abdeslam, Jean Merckle, and Dirk Benyoucef. Smart meter systems detection & classification using artificial neural networks. In *IECON 2012-38th Annual Conference on IEEE Industrial Electronics Society*, pages 3324–3329. IEEE, 2012.

[70] Wouter Labeeuw. Characterisation and modelling of residential electricity demand. *PhD diss., KU Leuven*, 2013.

[71] Hongliang Fei, Younghun Kim, Sambit Sahu, Milind Naphade, Sanjay K Mamidipalli, and John Hutchinson. Heat pump detection from coarse grained smart meter data with positive and unlabeled learning. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1330–1338, 2013.

[72] Tobias Brudermueller, Florian Wirth, Andreas Weigert, and Thorsten Staake. Automatic differentiation of variable and fixed speed heat pumps with smart meter data. In *2022 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 412–418, 2022. doi:10.1109/SmartGridComm52983.2022.9961055.

[73] Andreas Weigert, Konstantin Hopf, Nicolai Weinig, and Thorsten Staake. Detection of heat pumps from smart meter and open data. *Energy Informatics*, 3(Suppl 1):21, 2020.

[74] Tobias Brudermueller, Markus Kreft, Elgar Fleisch, and Thorsten Staake. Large-scale monitoring of residential heat pump cycling using smart meter data. *Applied Energy*, 350:121734, 2023. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2023.121734. URL https://www.sciencedirect.com/science/article/pii/S030626192301098X.

[75] Martin Neubert, Oliver Gnepper, Oliver Mey, and André Schneider. Detection of electric vehicles and photovoltaic systems in smart meter

data. *Energies*, 15(13), 2022. ISSN 1996-1073. doi:10.3390/en15134922. URL https://www.mdpi.com/1996-1073/15/13/4922.

[76] Hui Song, Chen Liu, Mahdi Jalili, Xinghuo Yu, and Peter McTaggart. Ensemble Classification Model for EV Identification From Smart Meter Recordings. *IEEE Transactions on Industrial Informatics*, 19(3):3274–3283, 2023. doi:10.1109/TII.2022.3175750.

[77] Zhilin Zhang, Jae Hyun Son, Ying Li, Mark Trayer, Zhouyue Pi, Dong Yoon Hwang, and Joong Ki Moon. Training-free non-intrusive load monitoring of electric vehicle charging with low sampling rate. In *IECON 2014 - 40th Annual Conference of the IEEE Industrial Electronics Society*, pages 5419–5425, 2014. doi:10.1109/IECON.2014.7049328.

[78] Qiyun Dang, Yuchong Huo, and Chu Sun. Privacy Preservation Needed for Smart Meter System: A Methodology to Recognize Electric Vehicle (EV) Models. In *2018 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia)*, pages 1016–1020, 2018. doi:10.1109/ISGT-Asia.2018.8467822.

[79] Behzad Najafi, Luca Di Narzo, Fabio Rinaldi, and Reza Arghandeh. Machine learning based disaggregation of air-conditioning loads using smart meter data. *IET Generation, Transmission & Distribution*, 14(21): 4755–4762, 2020.

[80] Krystian X. Perez, Wesley J. Cole, Joshua D. Rhodes, Abigail Ondeck, Michael Webber, Michael Baldea, and Thomas F. Edgar. Nonintrusive disaggregation of residential air-conditioning loads from sub-hourly smart meter data. *Energy and Buildings*, 81:316–325, 2014. ISSN 0378-7788. doi:https://doi.org/10.1016/j.enbuild.2014.06.031. URL https://www.sciencedirect.com/science/article/pii/S0378778814005131.

[81] Shi Su, Yuting Yan, Hai Lu, Li Kangping, Sun Yujing, Wang Fei, Liu Liming, and Ren Hui. Non-intrusive load monitoring of air conditioning using low-resolution smart meter data. In *2016 IEEE International Conference on Power System Technology (POWERCON)*, pages 1–5, 2016. doi:10.1109/POWERCON.2016.7753952.

[82] Dong Chen, Sean Barker, Adarsh Subbaswamy, David Irwin, and Prashant Shenoy. Non-intrusive occupancy monitoring using smart meters. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, pages 1–8, 2013.

[83] T. Vafeiadis, S. Zikos, G. Stavropoulos, D. Ioannidis, S. Krinidis, D. Tzovaras, and K. Moustakas. Machine learning based occupancy detection

via the use of smart meters. In *2017 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*, pages 6–12, 2017. doi:10.1109/ISCSIC.2017.15.

[84] Rouzbeh Razavi, Amin Gharipour, Martin Fleury, and Ikpe Justice Akpan. Occupancy detection of residential buildings using smart meter data: A large-scale study. *Energy and Buildings*, 183:195–208, 2019. ISSN 0378-7788. doi:https://doi.org/10.1016/j.enbuild.2018.11.025. URL https://www.sciencedirect.com/science/article/pii/S0378778818316724.

[85] Alo Allik, Siim Muiste, and Heino Pihlap. Smart meter data analytics for occupancy detection of buildings with renewable energy generation. In *2020 9th International Conference on Renewable Energy Research and Application (ICRERA)*, pages 248–251, 2020. doi:10.1109/ICRERA49962.2020.9242830.

[86] Christian Beckel, Leyna Sadamori, Thorsten Staake, and Silvia Santini. Revealing household characteristics from smart meter data. *Energy*, 78:397–410, 2014. ISSN 0360-5442. doi:https://doi.org/10.1016/j.energy.2014.10.025. URL https://www.sciencedirect.com/science/article/pii/S0360544214011748.

[87] Joshua D. Rhodes, Wesley J. Cole, Charles R. Upshaw, Thomas F. Edgar, and Michael E. Webber. Clustering analysis of residential electricity demand profiles. *Applied Energy*, 135:461–471, 2014. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2014.08.111. URL https://www.sciencedirect.com/science/article/pii/S0306261914009349.

[88] Wenjun Tang, Hao Wang, Xian-Long Lee, and Hong-Tzer Yang. Machine learning approach to uncovering residential energy consumption patterns based on socioeconomic and smart meter data. *Energy*, 240:122500, 2022. ISSN 0360-5442. doi:https://doi.org/10.1016/j.energy.2021.122500. URL https://www.sciencedirect.com/science/article/pii/S0360544221027493.

[89] Rouzbeh Razavi and Amin Gharipour. Rethinking the privacy of the smart grid: What your smart meter data can reveal about your household in ireland. *Energy Research & Social Science*, 44:312–323, 2018. ISSN 2214-6296. doi:https://doi.org/10.1016/j.erss.2018.06.005. URL https://www.sciencedirect.com/science/article/pii/S2214629618305899.

[90] Fintan McLoughlin, Aidan Duffy, and Michael Conlon. Characterising domestic electricity consumption patterns by

dwelling and occupant socio-economic variables: An Irish case study. *Energy and Buildings*, 48:240–248, 2012. ISSN 0378-7788. doi:https://doi.org/10.1016/j.enbuild.2012.01.037. URL https://www.sciencedirect.com/science/article/pii/S0378778812000680.

[91] F.M. Andersen, P.A. Gunkel, H.K. Jacobsen, and L. Kitzing. Residential electricity consumption and household characteristics: An econometric analysis of danish smart-meter data. *Energy Economics*, 100:105341, 2021. ISSN 0140-9883. doi:https://doi.org/10.1016/j.eneco.2021.105341. URL https://www.sciencedirect.com/science/article/pii/S0140988321002474.

[92] Mingyang Sun, Ioannis Konstantelos, and Goran Strbac. Analysis of diversified residential demand in London using smart meter and demographic data. In *2016 IEEE Power and Energy Society General Meeting (PESGM)*, pages 1–5. IEEE, 2016.

[93] Alix Hattenstone. Almost four million smart meters not working properly. *BBC News*. URL https://www.bbc.com/news/articles/cz9zqn77ezno.

[94] Lulu Wen, Kaile Zhou, Shanlin Yang, and Lanlan Li. Compression of smart meter big data: A survey. *Renewable and Sustainable Energy Reviews*, 91:59–69, 2018. ISSN 1364-0321. doi:https://doi.org/10.1016/j.rser.2018.03.088. URL https://www.sciencedirect.com/science/article/pii/S1364032118301849.

[95] Bishnu P Bhattarai, Sumit Paudyal, Yusheng Luo, Manish Mohanpurkar, Kwok Cheung, Reinaldo Tonkoski, Rob Hovsapian, Kurt S Myers, Rui Zhang, Power Zhao, et al. Big data analytics in smart grids: state-of-the-art, challenges, opportunities, and future directions. *IET Smart Grid*, 2(2):141–154, 2019.

[96] Yang Zhang, Tao Huang, and Ettore Francesco Bompard. Big data analytics in smart grids: a review. *Energy informatics*, 1(1):1–24, 2018.

[97] Shahid Tufail, Imtiaz Parvez, Shanzeh Batool, and Arif Sarwat. A survey on cybersecurity challenges, detection, and mitigation techniques for the smart grid. *Energies*, 14(18), 2021. ISSN 1996-1073. doi:10.3390/en14185894. URL https://www.mdpi.com/1996-1073/14/18/5894.

[98] Falah Alanazi, Jinsub Kim, and Eduardo Cotilla-Sanchez. Load oscillating attacks of smart grids: Vulnerability analysis. *IEEE Access*, 11:36538–36549, 2023. doi:10.1109/ACCESS.2023.3266249.

[99] Mohammad Ghiasi, Taher Niknam, Zhanle Wang, Mehran Mehrandezh, Moslem Dehghani, and Noradin Ghadimi. A comprehensive review of cyber-attacks and defense mechanisms for improving security in smart grid energy systems: Past, present and future. *Electric Power Systems Research*, 215:108975, 2023. ISSN 0378-7796. doi:https://doi.org/10.1016/j.epsr.2022.108975. URL https://www.sciencedirect.com/science/article/pii/S0378779622010240.

[100] Azadeh Peyman, Darren Addison, Terry Mee, Cristian Goiceanu, Myron Maslanyj, and Simon Mann. Exposure to electromagnetic fields from smart utility meters in gb; part i) laboratory measurements. *Bioelectromagnetics*, 38(4):280–294, 2017.

[101] Muhammad RA Qureshi, Yasir Alfadhl, Xiaodong Chen, Azadeh Peyman, Myron Maslanyj, and Simon Mann. Assessment of exposure to radio frequency electromagnetic fields from smart utility meters in gb; part ii) numerical assessment of induced sar within the human body. *Bioelectromagnetics*, 39(3):200–216, 2018.

[102] Carolina Calderon, Darren Addison, Nishtha Chopra, Simon Mann, Myron Maslanyj, and Azadeh Peyman. Exposure to electromagnetic fields from smart utility meters in gb; part iii) on-site measurements in homes. *Bioelectromagnetics*, 40(6):434, 2019.

[103] Lukas G. Swan and V. Ismet Ugursal. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews*, 13(8):1819–1835, 2009. ISSN 1364-0321. doi:https://doi.org/10.1016/j.rser.2008.09.033. URL https://www.sciencedirect.com/science/article/pii/S1364032108001949.

[104] A. Grandjean, J. Adnot, and G. Binet. A review and an analysis of the residential electric load curve models. *Renewable and Sustainable Energy Reviews*, 16(9):6539–6565, 2012. ISSN 1364-0321. doi:https://doi.org/10.1016/j.rser.2012.08.013. URL https://www.sciencedirect.com/science/article/pii/S1364032112004820.

[105] MOSTLY AI. What is data anonymization? https://mostly.ai/what-is-data-anonymization. Accessed: 2024-04-12.

[106] Vlaams Parlement. Decreet van 8 mei 2009 houdende algemene bepalingen betreffende het energiebeleid (Energiedecreet) (in Dutch). https://codex.vlaanderen.be/Zoeken/Document.aspx?DID=1018092, 2009.

[107] Fluvius System Operator cv. Fluvius Open Data - Consumption profiles digital electricity meters: quarterly values for a full year. https://opendata.fluvius.be/explore/dataset/1_50-verbruiksprofielen-dm-elek-kwartierwaarden-voor-een-volledig-jaar/information/, 2024.

[108] Vlaamse Regulator van de Electriciteits-en Gasmarkt (VREG). Data Management Report 2023 (in Dutch). https://www.vreg.be/sites/default/files/document/rapp-2023-20.pdf, 2023.

[109] M. Turowski, B. Heidrich, L. Weingärtner, L. Springer, K. Phipps, B. Schäfer, R. Mikut, and V. Hagenmeyer. Generating synthetic energy time series: A review. *Renewable and Sustainable Energy Reviews*, 206:114842, 2024. ISSN 1364-0321. doi:https://doi.org/10.1016/j.rser.2024.114842. URL https://www.sciencedirect.com/science/article/pii/S1364032124005689.

[110] Samik Raychaudhuri. Introduction to monte carlo simulation. In *2008 Winter Simulation Conference*, pages 91–100, 2008. doi:10.1109/WSC.2008.4736059.

[111] Mehdi Mirza. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[112] Houda Bouderraoui and Mouhcine Chami. SGSim: Load Profile Generator for Smart Grid Applications. In *2018 Renewable Energies, Power Systems & Green Inclusive Economy (REPS-GIE)*, pages 1–6, 2018. doi:10.1109/REPSGIE.2018.8488843.

[113] Ian Richardson, Murray Thomson, David Infield, and Conor Clifford. Domestic electricity use: A high-resolution energy demand model. *Energy and Buildings*, 42(10):1878–1887, 2010. ISSN 0378-7788. doi:https://doi.org/10.1016/j.enbuild.2010.05.023. URL https://www.sciencedirect.com/science/article/pii/S0378778810001854.

[114] Marianne M. Armstrong, Mike C. Swinton, Hajo Ribberink, Ian Beausoleil-Morrison, and Jocelyn Millette. Synthetically derived profiles for representing occupant-driven electric loads in Canadian housing. *Journal of Building Performance Simulation*, 2(1):15–30, 2009.

[115] C. Sandels, J. Widén, and L. Nordström. Forecasting household consumer electricity load profiles with a combined physical and behavioral approach. *Applied Energy*, 131:267–278, 2014. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2014.06.048. URL https://www.sciencedirect.com/science/article/pii/S0306261914006308.

[116] M. Nijhuis, M. Gibescu, and J.F.G. Cobben. Bottom-up Markov Chain Monte Carlo approach for scenario based residential load modelling with publicly available data. *Energy and Buildings*, 112:121–129, 2016. ISSN 0378-7788. doi:https://doi.org/10.1016/j.enbuild.2015.12.004. URL https://www.sciencedirect.com/science/article/pii/S0378778815304436.

[117] Lorenzo Bottaccioli, Santa Di Cataldo, Andrea Acquaviva, and Edoardo Patti. Realistic multi-scale modeling of household electricity behaviors. *IEEE Access*, 7:2467–2489, 2019. doi:10.1109/ACCESS.2018.2886201.

[118] Runming Yao and Koen Steemers. A method of formulating energy load profile for domestic buildings in the UK. *Energy and Buildings*, 37(6):663–671, 2005. ISSN 0378-7788. doi:https://doi.org/10.1016/j.enbuild.2004.09.007. URL https://www.sciencedirect.com/science/article/pii/S037877880400307X.

[119] Filip Jorissen, Glenn Reynders, Ruben Baetens, Damien Picard, Dirk Saelens, and Lieve Helsen. Implementation and verification of the IDEAS building energy simulation library. *Journal of Building Performance Simulation*, 11(6):669–688, 2018.

[120] Ruben Baetens and Dirk Saelens. Modelling uncertainty in district energy simulations by stochastic residential occupant behaviour. *Journal of Building Performance Simulation*, 9(4):431–447, 2016.

[121] C. Sandels, D. Brodén, J. Widén, L. Nordström, and E. Andersson. Modeling office building consumer load with a combined physical and behavioral approach: Simulation and validation. *Applied Energy*, 162:472–485, 2016. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2015.10.141. URL https://www.sciencedirect.com/science/article/pii/S0306261915013835.

[122] Anna Marszal-Pomianowska, Per Heiselberg, and Olena Kalyanova Larsen. Household electricity demand profiles – a high-resolution load model to facilitate modelling of energy flexible buildings. *Energy*, 103:487–501, 2016. ISSN 0360-5442. doi:https://doi.org/10.1016/j.energy.2016.02.159. URL https://www.sciencedirect.com/science/article/pii/S0360544216302213.

[123] Eurostat. Harmonised European time use surveys. https://ec.europa.eu/eurostat/web/microdata/harmonised-european-time-use-surveys, 2024.

[124] Bureau of Labor Statistics. American Time Use Survey. https://www.bls.gov/tus/overview.htm, 2024.

[125] Enrico Dalla Maria, Mattia Secchi, and David Macii. A flexible top-down data-driven stochastic model for synthetic load profiles generation. *Energies*, 15(1), 2022. ISSN 1996-1073. doi:10.3390/en15010269. URL https://www.mdpi.com/1996-1073/15/1/269.

[126] Yigzaw G. Yohanis, Jayanta D. Mondol, Alan Wright, and Brian Norton. Real-life energy use in the UK: How occupancy and dwelling characteristics affect domestic electricity use. *Energy and Buildings*, 40(6):1053–1059, 2008. ISSN 0378-7788. doi:https://doi.org/10.1016/j.enbuild.2007.09.001. URL https://www.sciencedirect.com/science/article/pii/S037877880700223X.

[127] Tony Craig, J. Gary Polhill, Ian Dent, Carlos Galan-Diaz, and Simon Heslop. The North East Scotland Energy Monitoring Project: Exploring relationships between household occupants and energy usage. *Energy and Buildings*, 75:493–503, 2014. ISSN 0378-7788. doi:https://doi.org/10.1016/j.enbuild.2014.02.038. URL https://www.sciencedirect.com/science/article/pii/S0378778814001522.

[128] Islam Safak Bayram, Faraj Saffouri, and Muammer Koc. Generation, analysis, and applications of high resolution electricity load profiles in Qatar. *Journal of Cleaner Production*, 183:527–543, 2018. ISSN 0959-6526. doi:https://doi.org/10.1016/j.jclepro.2018.02.084. URL https://www.sciencedirect.com/science/article/pii/S0959652618303901.

[129] Vasilis Michalakopoulos, Elissaios Sarmas, Ioannis Papias, Panagiotis Skaloumpakas, Vangelis Marinakis, and Haris Doukas. A machine learning-based framework for clustering residential electricity load profiles to enhance demand response programs. *Applied Energy*, 361:122943, 2024. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2024.122943. URL https://www.sciencedirect.com/science/article/pii/S030626192400326X.

[130] Ignacio Benitez Sanchez, Ignacio Delgado Espinos, Laura Moreno Sarrion, Alfredo Quijano Lopez, and Isabel Navalon Burgos. Clients segmentation according to their domestic energy consumption by the use of self-organizing maps. In *2009 6th International Conference on the European Energy Market*, pages 1–6, 2009. doi:10.1109/EEM.2009.5207172.

[131] Martin Pullinger, Ellen Zapata-Webborn, Jonathan Kilgour, Simon Elam, Jessica Few, Nigel Goddard, Clare Hanmer, Eoghan

McKenna, Tadj Oreszczyn, and Lynda Webb. Capturing variation in daily energy demand profiles over time with cluster analysis in British homes (September 2019 – August 2022). *Applied Energy*, 360:122683, 2024. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2024.122683. URL https://www.sciencedirect.com/science/article/pii/S0306261924000667.

[132] Flora Charbonnier, Thomas Morstyn, and Malcolm McCulloch. Home electricity data generator (HEDGE): An open-access tool for the generation of electric vehicle, residential demand, and PV generation profiles. *MethodsX*, 12:102618, 2024. ISSN 2215-0161. doi:https://doi.org/10.1016/j.mex.2024.102618. URL https://www.sciencedirect.com/science/article/pii/S2215016124000724.

[133] Zhe Wang and Tianzhen Hong. Generating realistic building electrical load profiles through the Generative Adversarial Network (GAN). *Energy and Buildings*, 224:110299, 2020. ISSN 0378-7788. doi:https://doi.org/10.1016/j.enbuild.2020.110299. URL https://www.sciencedirect.com/science/article/pii/S0378778820307234.

[134] Samer El Kababji and Pirathayini Srikantha. A data-driven approach for generating synthetic load patterns and usage habits. *IEEE Transactions on Smart Grid*, 11(6):4984–4995, 2020. doi:10.1109/TSG.2020.3007984.

[135] Yize Chen, Yishen Wang, Daniel Kirschen, and Baosen Zhang. Model-free renewable scenario generation using generative adversarial networks. *IEEE Transactions on Power Systems*, 33(3):3265–3275, 2018. doi:10.1109/TPWRS.2018.2794541.

[136] Chun Fu, Hussain Kazmi, Matias Quintana, and Clayton Miller. Creating synthetic energy meter data using conditional diffusion and building metadata. *Energy and Buildings*, 312:114216, 2024. ISSN 0378-7788. doi:https://doi.org/10.1016/j.enbuild.2024.114216. URL https://www.sciencedirect.com/science/article/pii/S0378778824003323.

[137] Olivier Neu, Brónagh Sherlock, Simeon Oxizidis, Damian Flynn, and Donal Finn. Developing building archetypes for electrical load shifting assessment: Analysis of Irish residential stock. 2014.

[138] Brandon J. Johnson, Michael R. Starke, Omar A. Abdelaziz, Roderick K. Jackson, and Leon M. Tolbert. A MATLAB based occupant driven dynamic model for predicting residential power demand. In *2014 IEEE PES T&D Conference and Exposition*, pages 1–5, 2014. doi:10.1109/TDC.2014.6863381.

[139] Manu Lahariya, Dries F. Benoit, and Chris Develder. Synthetic data generator for electric vehicle charging sessions: Modeling and evaluation using real-world data. *Energies*, 13(16), 2020. ISSN 1996-1073. doi:10.3390/en13164211. URL https://www.mdpi.com/1996-1073/13/16/4211.

[140] Omid Motlagh, Phillip Paevere, Tang Sai Hong, and George Grozev. Analysis of household electricity consumption behaviours: Impact of domestic electricity generation. *Applied Mathematics and Computation*, 270:165 – 178, 2015. ISSN 0096-3003. doi:https://doi.org/10.1016/j.amc.2015.08.029. URL http://www.sciencedirect.com/science/article/pii/S009630031501084X.

[141] Eurostat. Energy statistics - electricity prices for domestic and industrial consumers, price components. https://ec.europa.eu/eurostat/cache/metadata/en/nrg_pc_204_esms.htm. Accessed: 2020-11-10.

[142] Vlaamse Regulator van de Electriciteits-en Gasmarkt (VREG). Verbruiksprofielen en productieprofielen (in Dutch). https://www.vreg.be/nl/verbruiksprofielen-en-productieprofielen, 2024.

[143] Flemish Regulator of the Electricity and Gas Market. Residential energy consumption in Flanders. https://www.vreg.be/nl/energieverbruik, 2022.

[144] John E Hay. Calculation if the solar radiation incident on inclined surfaces. In *Proceedings first Canadian Solar Radiation Data Workshop, Toronto. Ontario, Canada 1978*, 1978.

[145] William F Holmgren, Clifford W Hansen, and Mark A Mikofski. pvlib python: A python package for modeling solar energy systems. *Journal of Open Source Software*, 3(29):884, 2018.

[146] Elaad NL. Data Analytics. www.elaad.nl/research/data-analytics/ (accessed on 2022-09-23), 2020.

[147] Samuel Pelletier, Ola Jabali, Gilbert Laporte, and Marco Veneroni. Battery degradation and behaviour for electric vehicles: Review and numerical analyses of several models. *Transportation Research Part B: Methodological*, 103:158–187, 2017. doi:10.1016/j.trb.2017.01.020.

[148] Oliver Ruhnau, Lion Hirth, and Aaron Praktiknjo. Time series of heat demand and heat pump efficiency for energy system modeling. *Scientific data*, 6(1):1–10, 2019.

[149] Van Hove, Matthias and Delghust, Marc and Janssens, Arnold. Analyse naar de haalbaarheid van statistische modellen die energiegebruik in woningen kunnen voorspellen op basis van gebouwparameters [full report], 2021.

[150] Verbrugge, Silke and Delghust, Marc and Janssens, Arnold. Onderzoek naar de relatie tussen het E-peil, het berekende energiegebruik in de EPB-aangifte en het werkelijke energiegebruik op basis van enquêtegegevens en gegevens opgevraagd bij de netbeheerder [full report], 2019.

[151] R. Claeys, T. Delerue, and J. Desmet. Assessing the influence of the aggregation level of residential consumers through load duration curves. In *2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*, pages 1–5, 2019.

[152] Robbert Claeys, Hakim Azaioud, Rémy Cleenwerck, Jos Knockaert, and Jan Desmet. A Novel Feature Set for Low-Voltage Consumers, Based on the Temporal Dependence of Consumption and Peak Demands. *Energies*, 14(1), 2021. ISSN 1996-1073. doi:10.3390/en14010139. URL https://www.mdpi.com/1996-1073/14/1/139.

[153] Gilbert M Masters. *Renewable and efficient electric power systems*. John Wiley & Sons, 2013.

[154] AL Soyster and RT Eynon. The conceptual basis of the electric utility sub-model of project independence evaluation system. *Applied Mathematical Modelling*, 3(4):242–248, 1979.

[155] FH Murphy, S Sen, and AL Soyster. Electric utility capacity expansion planning with uncertain load forecasts. *IIE Transactions*, 14(1):52–59, 1982.

[156] Alain Poulin, Michel Dostie, Michaël Fournier, and Simon Sansregret. Load duration curve: A tool for technico-economic analysis of energy solutions. *Energy and Buildings*, 40(1):29 – 35, 2008. ISSN 0378-7788. doi:10.1016/j.enbuild.2007.01.020. URL http://www.sciencedirect.com/science/article/pii/S0378778807000278.

[157] Isaac Newton and John Colson. *The method of fluxions and infinite series: with its application to the geometry of curve-lines*. Nourse, 1736.

[158] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-

0427. doi:https://doi.org/10.1016/0377-0427(87)90125-7. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

[159] T. Caliński and J Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. doi:10.1080/03610927408827101.

[160] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3 (3):32–57, 1973. doi:10.1080/01969727308546046.

[161] David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1 (2):224–227, 1979. doi:10.1109/TPAMI.1979.4766909.

[162] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the Number of Clusters in a Data Set Via the Gap Statistic. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 63(2): 411–423, 01 2002. ISSN 1369-7412. doi:10.1111/1467-9868.00293.

[163] Ignacio Benítez, Alfredo Quijano, José-Luis Díez, and Ignacio Delgado. Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers. *International Journal of Electrical Power & Energy Systems*, 55:437–448, 2014. ISSN 0142-0615. doi:https://doi.org/10.1016/j.ijepes.2013.09.022. URL https://www.sciencedirect.com/science/article/pii/S0142061513004043.

[164] George Emeka Okereke, Mohamed Chaker Bali, Chisom Nneoma Okwueze, Emmanuel Chukwudi Ukekwe, Stephenson Chukwukanedu Echezona, and Celestine Ikechukwu Ugwu. K-means clustering of electricity consumers using time-domain features from smart meter data. *Journal of Electrical Systems and Information Technology*, 10(1):2, 2023.

[165] Fintan McLoughlin, Aidan Duffy, and Michael Conlon. A clustering approach to domestic electricity load profile characterisation using smart metering data. *Applied Energy*, 141:190–199, 2015. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2014.12.039. URL https://www.sciencedirect.com/science/article/pii/S0306261914012963.

[166] Saehong Park, Seunghyoung Ryu, Yohwan Choi, Jihyo Kim, and Hongseok Kim. Data-driven baseline estimation of residential buildings for demand response. *Energies*, 8(9):10239–10259, 2015. ISSN 1996-1073. doi:10.3390/en80910239. URL https://www.mdpi.com/1996-1073/8/9/10239.

[167] Joaquim L. Viegas, Susana M. Vieira, R. Melício, V.M.F. Mendes, and João M.C. Sousa. Classification of new electricity customers based on surveys and smart metering data. *Energy*, 107:804–817, 2016. ISSN 0360-5442. doi:https://doi.org/10.1016/j.energy.2016.04.065. URL https://www.sciencedirect.com/science/article/pii/S0360544216304789.

[168] Joe H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. doi:10.1080/01621459.1963.10500845.

[169] Akito Ozawa, Ryota Furusato, and Yoshikuni Yoshida. Determining the relationship between a household's lifestyle and its electricity consumption in Japan by analyzing measured electric load profiles. *Energy and Buildings*, 119:200–210, 2016. ISSN 0378-7788. doi:https://doi.org/10.1016/j.enbuild.2016.03.047. URL https://www.sciencedirect.com/science/article/pii/S0378778816302018.

[170] George J Tsekouras, Nikos D Hatziargyriou, and Evangelos N Dialynas. Two-stage pattern recognition of load curves for classification of electricity customers. *IEEE Transactions on Power Systems*, 22(3):1120–1128, 2007.

[171] Joao Pedro Gouveia and Júlia Seixas. Unraveling electricity consumption profiles in households through clusters: Combining smart meters and door-to-door surveys. *Energy and Buildings*, 116:666–676, 2016. ISSN 0378-7788. doi:https://doi.org/10.1016/j.enbuild.2016.01.043. URL https://www.sciencedirect.com/science/article/pii/S0378778816300421.

[172] Mairead L Bermingham, Ricardo Pong-Wong, Athina Spiliopoulou, Caroline Hayward, Igor Rudan, Harry Campbell, Alan F Wright, James F Wilson, Felix Agakov, Pau Navarro, et al. Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific reports*, 5:10312, 2015.

[173] Reem Al-Otaibi, Nanlin Jin, Tom Wilcox, and Peter Flach. Feature construction and calibration for clustering daily load curves from smart-meter data. *IEEE Transactions on industrial informatics*, 12(2):645–654, 2016.

[174] Sergio Valero Verdú, Mario Ortiz Garcia, Carolina Senabre, Antonio Gabaldón Marin, and Francisco J Garcıa Franco. Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps. *IEEE Transactions on Power Systems*, 21(4):1672–1682, 2006.

[175] Matteo Manera and Angelo Marzullo. Modelling the load curve of aggregate electricity consumption using principal components. *Environmental Modelling & Software*, 20(11):1389–1400, 2005.

[176] Nanlin Jin, Peter Flach, Tom Wilcox, Royston Sellman, Joshua Thumim, and Arno Knobbe. Subgroup discovery in smart electricity meter data. *IEEE Transactions on Industrial Informatics*, 10(2):1327–1336, 2014.

[177] S. Haben, C. Singleton, and P. Grindrod. Analysis and clustering of residential customers energy behavioral demand using smart meter data. *IEEE Transactions on Smart Grid*, 7(1):136–144, 2016. doi:10.1109/TSG.2015.2409786.

[178] Ian Dent, Uwe Aickelin, Tom Rodden, and Tony Craig. Finding the creatures of habit; clustering households based on their flexibility in using electricity. *Clustering Households Based on Their Flexibility in Using Electricity (January 1, 2012)*, 2012.

[179] Teemu Räsänen and Mikko Kolehmainen. Feature-based clustering for electricity use time series data. In *International conference on adaptive and natural computing algorithms*, pages 401–412. Springer, 2009.

[180] Alexander Martin Tureczek and Per Sieverts Nielsen. Structured literature review of electricity consumption classification using smart meter data. *Energies*, 10(5):584, 2017.

[181] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[182] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, and et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi:10.1038/s41592-019-0686-2.

[183] Jimyung Kang and Jee-Hyong Lee. Electricity customer clustering following experts' principle for demand response applications. *Energies*, 8(10):12242–12265, 2015.

[184] Jungsuk Kwac, June Flora, and Ram Rajagopal. Household energy consumption segmentation using hourly data. *IEEE Transactions on Smart Grid*, 5(1):420–430, 2014.

[185] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[186] François Bolley. Separability and completeness for the Wasserstein distance. In *Séminaire de probabilités XLI*, pages 371–377. Springer, 2008.

[187] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[188] Robbert Claeys, Rémy Cleenwerck, Jos Knockaert, and Jan Desmet. Stochastic generation of residential load profiles with realistic variability based on wavelet-decomposed smart meter data. *Applied Energy*, 350:121750, 2023. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2023.121750. URL https://www.sciencedirect.com/science/article/pii/S0306261923011145.

[189] Raffi Avo Sevlian and Ram Rajagopal. A model for the effect of aggregation on short term load forecasting. In *2014 IEEE PES General Meeting | Conference & Exposition*, pages 1–5, 2014. doi:10.1109/PESGM.2014.6938899.

[190] Han Li, Zhe Wang, Tianzhen Hong, Andrew Parker, and Monica Neukomm. Characterizing patterns and variability of building electric load profiles in time and frequency domains. *Applied Energy*, 291:116721, 2021. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2021.116721. URL https://www.sciencedirect.com/science/article/pii/S0306261921002397.

[191] Ran Li, Furong Li, and Nathan D. Smith. Load characterization and low-order approximation for smart metering data in the spectral domain. *IEEE Transactions on Industrial Informatics*, 13(3):976–984, 2017. doi:10.1109/TII.2016.2638319.

[192] Alfred Haar. *Zur theorie der orthogonalen funktionensysteme*. Georg-August-Universitat, Gottingen., 1910.

[193] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

[194] Martin Forsberg and Elin Johansson. Wavelets and wavelet packets. 1999.

[195] Andrew Parker, Kevin James, Dongming Peng, and Mahmoud A. Alahmad. Framework for extracting and characterizing load profile variability based on a comparative study of different wavelet functions. *IEEE Access*, 8:217483–217498, 2020. doi:10.1109/ACCESS.2020.3042125.

[196] Andrew Parker, Sam Moayedi, Kevin James, Dongming Peng, and Mahmoud A Alahmad. A case study to quantify variability in building load profiles. *IEEE Access*, 9:127799–127813, 2021.

[197] Xiangqi Zhu and Barry Mather. DWT-Based Aggregated Load Modeling and Evaluation for Quasi-Static Time-Series Simulation on Distribution Feeders. In *2018 IEEE Power Energy Society General Meeting (PESGM)*, pages 1–5, 2018. doi:10.1109/PESGM.2018.8585535.

[198] Adriano Z Zambom and Dias Ronaldo. A review of kernel density estimation with applications to econometrics. *International Econometric Review*, 5(1):20–42, 2013.

[199] Matthew Li, David Allinson, and Miaomiao He. Seasonal variation in household electricity demand: A comparison of monitored and synthetic daily load profiles. *Energy and Buildings*, 179:292–300, 2018. ISSN 0378-7788. doi:https://doi.org/10.1016/j.enbuild.2018.09.018. URL https://www.sciencedirect.com/science/article/pii/S037877881831346X.

[200] Robbert Claeys, Rémy Cleenwerck, Jos Knockaert, and Jan Desmet. Capturing multiscale temporal dynamics in synthetic residential load profiles through Generative Adversarial Networks (GANs). *Applied Energy*, 360:122831, 2024. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2024.122831. URL https://www.sciencedirect.com/science/article/pii/S0306261924002149.

[201] Stephen Haben, Colin Singleton, and Peter Grindrod. Analysis and clustering of residential customers energy behavioral demand using smart meter data. *IEEE Transactions on Smart Grid*, 7(1):136–144, 2016. doi:10.1109/TSG.2015.2409786.

[202] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks, 2014.

[203] Yuxuan Gu, Qixin Chen, Kai Liu, Le Xie, and Chongqing Kang. GAN-based Model for Residential Load Generation Considering Typical Consumption Patterns. In *2019 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pages 1–5, 2019. doi:10.1109/ISGT.2019.8791575.

[204] Mohammad Navid Fekri, Ananda Mohon Ghosh, and Katarina Grolinger. Generating Energy Data for Machine Learning with Recurrent Generative Adversarial Networks. *Energies*, 13(1), 2020. ISSN

1996-1073. doi:10.3390/en13010130. URL https://www.mdpi.com/1996-1073/13/1/130.

[205] Chenlu Tian, Chengdong Li, Guiqing Zhang, and Yisheng Lv. Data driven parallel prediction of building energy consumption using generative adversarial nets. *Energy and Buildings*, 186:230–243, 2019. ISSN 0378-7788. doi:https://doi.org/10.1016/j.enbuild.2019.01.034. URL https://www.sciencedirect.com/science/article/pii/S0378778818322965.

[206] Jianbin Li, Zhiqiang Chen, Long Cheng, and Xiufeng Liu. Energy data generation with Wasserstein Deep Convolutional Generative Adversarial Networks. *Energy*, 257:124694, 2022. ISSN 0360-5442. doi:https://doi.org/10.1016/j.energy.2022.124694. URL https://www.sciencedirect.com/science/article/pii/S0360544222015973.

[207] Zhiqiang Chen, Jianbin Li, Long Cheng, and Xiufeng Liu. Federated-WDCGAN: A federated smart meter data sharing framework for privacy preservation. *Applied Energy*, 334:120711, 2023. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2023.120711. URL https://www.sciencedirect.com/science/article/pii/S0306261923000752.

[208] Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series Generative Adversarial Networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/c9efe5f26cd17ba6216bbe2a7d26d490-Paper.pdf.

[209] Gaby Baasch, Guillaume Rousseau, and Ralph Evins. A Conditional Generative adversarial Network for energy use in multiple buildings using scarce data. *Energy and AI*, 5:100087, 2021. ISSN 2666-5468. doi:https://doi.org/10.1016/j.egyai.2021.100087. URL https://www.sciencedirect.com/science/article/pii/S2666546821000410.

[210] Zinan Lin, Alankar Jain, Chen Wang, Giulia Fanti, and Vyas Sekar. Using GANs for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions. In *Proceedings of the ACM Internet Measurement Conference*, IMC '20, page 464–483, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450381383. doi:10.1145/3419394.3423643. URL https://doi.org/10.1145/3419394.3423643.

[211] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using

implicit variational learning. *Advances in neural information processing systems*, 30, 2017.

[212] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs Created Equal? A Large-scale Study. *Advances in neural information processing systems*, 31, 2018.

[213] Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of Generative Adversarial Networks. *arXiv preprint arXiv:1806.07755*, 2018.

[214] Gretel Labs, Inc. Gretel Synthetics (v. 0.20.0), 2023. URL https://github.com/gretelai/gretel-synthetics.

[215] Rory V. Jones, Alba Fuertes, and Kevin J. Lomas. The socio-economic, dwelling and appliance related factors affecting electricity consumption in domestic buildings. *Renewable and Sustainable Energy Reviews*, 43:901–917, 2015. ISSN 1364-0321. doi:https://doi.org/10.1016/j.rser.2014.11.084. URL https://www.sciencedirect.com/science/article/pii/S1364032114010235.

[216] Zhifeng Guo, Kaile Zhou, Chi Zhang, Xinhui Lu, Wen Chen, and Shanlin Yang. Residential electricity consumption behavior: Influencing factors, related theories and intervention strategies. *Renewable and Sustainable Energy Reviews*, 81:399–412, 2018. ISSN 1364-0321. doi:https://doi.org/10.1016/j.rser.2017.07.046. URL https://www.sciencedirect.com/science/article/pii/S1364032117311164.

[217] Paul Westermann, Chirag Deb, Arno Schlueter, and Ralph Evins. Unsupervised learning of energy signatures to identify the heating system and building type using smart meter data. *Applied Energy*, 264:114715, 2020. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2020.114715. URL https://www.sciencedirect.com/science/article/pii/S0306261920302270.

[218] Statistics Flanders. Heating method of households. https://www.vlaanderen.be/statistiek-vlaanderen/energie/verwarmingswijze-woning, 2020.

[219] Johannes Weniger, Tjarko Tjaden, and Volker Quaschning. Sizing of Residential PV Battery Systems. *Energy Procedia*, 46:78–87, 2014. ISSN 1876-6102. doi:https://doi.org/10.1016/j.egypro.2014.01.160. URL https://www.sciencedirect.com/science/article/pii/S1876610214001763. 8th International Renewable Energy Storage Conference and Exhibition (IRES 2013).

[220] Sylvain Quoilin, Konstantinos Kavvadias, Arnaud Mercier, Irene Pappone, and Andreas Zucker. Quantifying self-consumption linked to solar home battery systems: Statistical analysis and economic assessment. *Applied Energy*, 182:58–67, 2016. ISSN 0306-2619. doi:https://doi.org/10.1016/j.apenergy.2016.08.077. URL https://www.sciencedirect.com/science/article/pii/S0306261916311643.

[221] G. W. Chang, S. Y. Chu, and H. L. Wang. An Improved Backward/Forward Sweep Load Flow Algorithm for Radial Distribution Systems. *IEEE Transactions on Power Systems*, 22(2):882–884, 2007. doi:10.1109/TPWRS.2007.894848.

[222] JA Michline Rupa and S Ganesh. Power flow analysis for radial distribution system using backward/forward sweep method. *International Journal of Electrical, Computer, Electronics and Communication Engineering*, 8(10):1540–1544, 2014.

[223] Roger C. Dugan and Thomas E. McDermott. An open source platform for collaborating on smart grid research. In *2011 IEEE Power and Energy Society General Meeting*, pages 1–7, 2011. doi:10.1109/PES.2011.6039829.

[224] Ali Hariri, Alvi Newaz, and Omar Faruque. Open-source python-OpenDSS interface for hybrid simulation of PV impact studies. *IET Generation, Transmission and Distribution*, 11(12):3125 –3133, 2017. doi:10.1049/iet-gtd.2016.1572.

[225] Rémy Cleenwerck, Hakim Azaioud, Robbert Claeys, Thierry Coosemans, Jos Knockaert, and Jan Desmet. An approach to the impedance modelling of low-voltage cables in digital twins. *Electric Power Systems Research*, 210:108075, 2022. ISSN 0378-7796. doi:https://doi.org/10.1016/j.epsr.2022.108075. URL https://www.sciencedirect.com/science/article/pii/S0378779622002991.