RUNE VAN DER MEIJDEN

LOADS OF LIFESTYLES

LOADS OF LIFESTYLES

RUNE VAN DER MEIJDEN

Loads of lifestyles

A latent lifestyle model for interpreting and simulating electrical energy consumption

by

R.P. van der Meijden

in partial fulfillment of the requirements for the degree of

Master of Science
in Applied Mathematics,
specialisation Probability & Statistics,

at the Delft University of Technology,
faculty of EEMCS,

to be defended publicly on Friday July 14, 2017 at 14:00.

Thesis committee:

Supervisor TU Delft	Dr. ir. F. H. van der Meulen	TU Delft – Statistics
Supervisor Stedin I	Ir. J. Pellis,	Stedin N.V. – Strategy & Innovation
Supervisor Stedin II	Dr. ir. Y. Koc,	Stedin N.V. – Risk Management
Full professor	Prof. dr. ir. G. Jongbloed,	TU Delft – Statistics
External professor	Dr. ir. M. B. van Gijzen,	TU Delft - Numerical Analysis

This thesis is confidential and cannot be made public until July 14, 2017.

An electronic version of this thesis is available at `www.repository.tudelft.nl`.

Cover design by Joeri Simonian
Printed by Druk. Tan Heck

Abstract

The rapid and uncertain penetration of distributed energy resources is changing the way people are consuming electricity. This development increases the risk of instability and congestion in the grid, posing great challenges to system operators in the near future. In order to optimally allocate resources and plan grid enhancement, distribution system operators need new tools to monitor the local energy transition and assess its future impact. Smart meter data is an important resource in gaining this insight, but unlocking the full potential requires new analytical methodologies. Recent research focused on identifying typical daily energy consumption patterns – load shapes – but a clear interpretation of these results is lacking. This research proposes a latent lifestyle model, which models energy consumption as a mixture of different lifestyles, each of which is defined by a distribution over load shapes. This extra layer of abstraction proves to be an effective way of identifying lifestyle-like patterns, that allow for a clear interpretation. Consumers can then be grouped based on their mixture of inferred lifestyles, serving as an input for grid simulation. Such a simulation showed that the all-electric homes cause the aggregated load to nearly triple compared to conventional consumers, due to their increased peak demand and simultaneity.

Acknowledgements

The past eight years have truly been an adventure. From being a fanatic race cyclist, dreaming that I would one day – perhaps today – ride the Tour de France, via racing solar powered Nuna7 through the Australian desert, to where I am right now: sitting behind my desk, only 4 hours left until I hand in my master thesis and finish my student life.

Many things have changed since I started my bachelor Applied Mathematics in September 2009: I moved out of my parent's old farmhouse into a student's shed; I lost some hair and gained some weight; some friendships and dreams faded, and were replaced by many new. But one constant factor remained: mathematics. Whether it was on my racing bike, trying to find patterns in the race-numbers on my competitors backs that would predict my victory, to the back seat of the Kia Carnival, trying to keep my head cool, surrounded with five touch-screen monitors and a retired astronaut screaming for new updates on the battery status of Nuna; never was the math far away.

However, all good things come to an end, and so is my time as an applied mathematics student. I would like to take some time and space to thank all the people that contributed to the inspiring, fun and life-changing process that I have been through.

First and foremost, I would like to thank the daily supervisors of this thesis – Frank, Jan and Yakup – for their great guidance throughout this journey. The excellent blend of theory and practice, mathematics and the grid, was the perfect mix for this project. Your kindness, patience and critical glance have greatly contributed to the work that is presented. Most of all, I would like thank you for the pleasant and inspiring time, and the freedom to work at my own pace and on my own ideas. I would also like to thank my two other committee members – Geurt and Martin – for their time to read through this extensive document, and attending my presentation.

Three months of this research were conducted at the Bits and Watts-lab at Stanford University. Here I had the honour to meet, discuss with, and learn from some of the smartest minds I have ever encountered. I would like thank all the people that made this journey possible and a great adventure. First of all, Professor Ram Rajagopal, my host professor, with whom I had many great discussions. Your ideas, motivating critiques and inspiring lectures have really shaped this thesis. Next, I would like to thank Heidi von Korff for our collaboration and the warm welcome that I received. I would also like to thank the rest of Professor Rajagopal's research group: June Flora, Sam Borgerson, Sid Patel, Michaelangelo Tabone, Mark Chen, Chin-Woo Tan, Yang Yu and Camille. Thank your very much, either for your feedback, the coffee, or just the great conversations.

The Rainbow Mansion at Rainbow Drive, Cupertino, was my home away from home. It felt like a dream, those 3 months that I had the pleasure to be among your great community. I am still jealous at myself 10 months ago, entering the big villa looking out over the bay area. Rainbow is more than just the typical hub of smart silicon valley people: it is a warm family from the first second you enter. The nicest place with the nicest people: Justyna, Roberto, Vane, Jeremy, Parnian, Dan, Bob, Uma, Eirik, Gavan, Aurora, I miss you all so much. Special gratitude to Alex for the hours you spent correcting my grammar. Witold, my man, thanks for being the greatest guide and friend.

I would also like to thank Arnoud, Bas, Daniël, David, Henry, Lisette, Marjolein, Michiel, and all other people at Stedin that I interacted with – either discussing content or talking about something totally irrelevant to my research. You made the numerous hours at Blaak a very educational and fun time. Special thanks to Gerja for all the coffee and conversations about cycling, music, career choices, VR and sucrology[1]. Mark, thanks a lot for the large amounts of coffee and the feedback on my writings. Van der Meijden - Klein Entink Consultancy will be reality soon.

Then I would like to thank all of my friends for being the awesome people that your are. Not enough place and time – still 3 hours left – is available to express my gratitude towards you. The adventures I experienced would not have been worthwhile without any of you. Throughout the past years, you have been a crucial ingredient of my happiness. Especially the mathematicians, Huize Bilspan, Nuna7, the Silly's, and Denktank'14 – you made my student life as awesome as possible. With regards to this thesis, special thanks go out to Jorrit, Kristiaan, Jasper, Romke and Joeri. I have asked many hours of your time, to check my writing, design my cover and come up with a thesis title that I didn't use in the end. I hope to be able to pay you back in the near future, either in kind or in beer.

Lastly, but most important; Beppy, Louis, Vidar and Siri, my parents and siblings. I would be nowhere without you and the safe haven of Onder De Zeven Linden. Whether for encouraging words, a good meal, a quiet place to study, or just some family warmth, no place is as good as home. I am infinitely lucky to have had you this near.

This adventure comes to an end, but new ones lie ahead. Thank you all for being part of this, I hope to see you even more on the other side!

Rune van der Meijden,
Delft, 4 July 2017

[1] for more on sucrology: see the end of this thesis

Contents

Acronyms

ACM Autoriteit Consument en Markt

AK-means adaptive K-means

BIC Bayesian information criterion

DSO distribution system operator

EM expectation maximisation

EV electric vehicle

HP heat pump

HV high-voltage

HVAC heating, ventilation and air-conditioning

LDA latent Dirichlet allocation

LV low-voltage

MV medium-voltage

NET new energy technology

NILM non-intrusive load monitoring

PCA principal component analysis

pLSI probabilistic latent semantic indexing

PV solar photovoltaic power system

SMD smart meter data

t-SNE t-distributed stochastic neighbour embedding

TSO transmission system operator

VEM variational expectation maximisation

Symbols

C Corpus, $C = \{\mathbf{w}_m\}_{m=1}^{M}$, the total collection of observations that LDA uses as input for inference.

D_{KL} Kullback-Leibler divergence.

D Dictionary of load shapes. $D = \{\boldsymbol{\mu}_v\}_{v=1}^{V}$.

$\boldsymbol{\alpha}$ Hyperparameter of the Dirichlet prior for the lifestyle distributions. $\boldsymbol{\alpha} \in \mathbb{R}^K$.

$\boldsymbol{\eta}$ The parameter vector $\boldsymbol{\eta} \in \mathbb{R}^V$ of the categorical distribution of load shapes per topic. $\mathrm{p}(W = wt \mid \tau, \boldsymbol{\eta}) = \eta_\tau$

$\boldsymbol{\mu}$ Cluster centres from adaptive K-means algorithm.

$\boldsymbol{\tau}$ Series of latent lifestyles. $\boldsymbol{\tau} = (\tau_1, ..., \tau_N)$

$\boldsymbol{\theta}$ The parameter vector of the lifestyle-per-document distribution. $\mathrm{p}(\tau = t \mid \boldsymbol{\theta}) = \theta_t$.

$\mathbf{P}$ Time series of energy consumption data. $P_t^{(m)}$ is the energy consumption of consumer m at time t.

$\mathbf{w}$ Load shape collection. $\mathbf{w} = (w_1, ..., w_N)$

$\mathbf{x}$ Vector of normalised hourly observations per day. $\mathbf{y}_i = \frac{\mathbf{x}_i}{||\mathbf{x}_i||_1}$

$\mathbf{y}$ Vector of hourly observations per day. $\mathbf{x}_{m,d} \in \mathbb{R}^{24}$ is consumption for home m at day d

τ Latent lifestyle or topic, the latent stochastic variable in LDA

c Index of cluster centre in dictionary

w Load shape, an index in dictionary D.

Chapter 1

Introduction

"President Trump pulls the United States out of the Paris Climate Agreement" was the headline in every major news media on the morning of June 2^{nd} of 2017. What was anticipated, but remained hard to imagine, became reality. The leader of the second largest CO_2-emitting country walked away from his responsibility to the earth and its future inhabitants. However pessimistic this message, the worldwide transition towards a low-carbon energy system is happening and it is happening fast. A report of the World Economic Forum concluded that renewable resources are already the cheapest option for new capacity in 30 countries around the world [1]. In the past year, record low prices for wind energy have been set in the Netherlands and Denmark [2], while prices for solar energy projects broke records in India, Chile and the United Arab Emirates [3]. This development is not only happening fast, but also much faster than expected. The International Energy Agency, like most other institutes, has consistently underestimated the global potential for solar energy in their WEO projections, as is shown in Figure 1.1. In 2010, they did not expect the global capacity to exceed 300 GW by 2030, only to observe that 305 GW was already installed 6 years later [4].

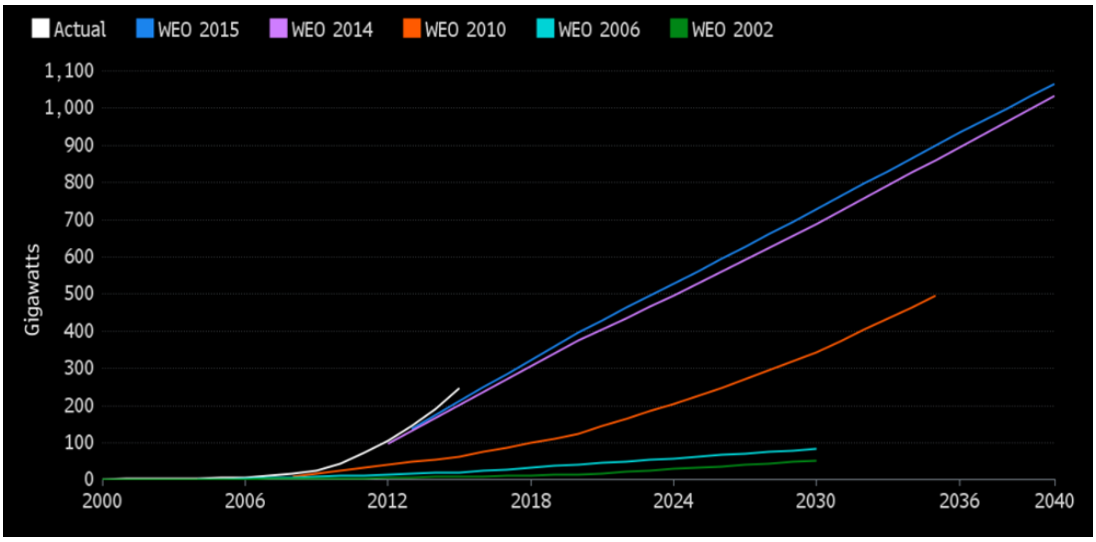


Figure 1.1: Projections of global installed PV capacity versus actual installed solar PV capacity by the International Energy Agency[5]. Forecasts have consistently underestimated the potential of PV.

A similar rapid and uncertain transition can be observed on a local scale. Incentives in the economic and policy sphere do cause a fast, but geographically varying, penetration of technologies like rooftop-PV, electric vehicles and electric heat pumps. For example, local subsidies caused a fast penetration of solar panels in parts of Groningen, the Netherlands, leading to grid instabilities in the summer of 2016 [6].

This global and local energy transition are both quickly changing the way energy is being produced and consumed. This impacts the electric grid and the way it is being operated. Therefore, grid operators require new tools for monitoring this transition and simulating its effects. Increased data availability from smart meters is an important resource in achieving this. This thesis develops a new method to gather insights from this data about the status and impact of the transition. The work in this thesis was done in cooperation with Stedin, a Dutch distribution system operator, and Stanford University.

In this chapter, the context and objectives of this research are discussed. Section 1.1 introduces the distribution system operator Stedin. In Section 1.2, the electric grid is shortly described, after which an overview of the relevant trends and challenges within the grid follows. Additionally, the smart grid is introduced as a valuable asset in dealing with these challenges. This section ends by stating the role of applied mathematics in unlocking this value. In Section 1.3, the research questions and goals of this thesis are stated. Also, an overview of the scope of this research and the used methods is given here. The chapter concludes with an overview of the structure of the report.

1.1 Company profile: Stedin

Stedin is one of the main distribution system operators (DSOs) in the Netherlands, responsible for managing the infrastructure and transportation of energy – electricity and gas – in most of the Dutch Randstad-region[1]. The company serves around 2 million residential, commercial and industrial customers by transporting 20 TWh of electricity on its 45,000 km of low- and medium voltage electricity lines, and 4.6 billion m^3 gas through its 24,000 km of gas infrastructure. Stedin has around 3900 employees and an annual revenue of €1.2 billion [7].

Stedin, like all DSOs, is a natural monopolist serving the vital societal function of delivering energy. Therefore, they are strictly regulated by the authority Autoriteit Consument en Markt (ACM). Stedin is in its geographical domain responsible for:

1. operating and maintaining the low- and medium voltage grid and gas infrastructure;
2. providing a grid connection to every new home, company or other building;
3. minimizing power cuts in low and medium voltage circuits;
4. keeping system costs as low as possible.

ACM annually rewards the best performing DSOs – in terms of outage and current quality – with the permission to increase their rates [8], creating a stimulus for performance-enhancing investments. Moreover, because of their vital function to the grid, DSOs are key stakeholders in the transition towards a low-carbon energy future. This is reflected in the company mission of Stedin: *”Renewable energy for everyone”* [7].

1.2 Research context

This section sketches the contextual landscape in which this research is conducted. The focus of this thesis is on electrical energy rather than energy from gas, thus a brief overview of the electrical grid is given. Relevant trends are mentioned in Section 1.2.2, after which the major challenges that these developments pose to the grid are discussed. In Section 1.2.4 the Smart Grid is introduced as an important

[1] Including three of the four biggest cities of the Netherlands: Rotterdam, The Hague and Utrecht.

concept in dealing with these challenges. Lastly, the role of applied mathematics within this landscape is discussed, linking to the significance of this thesis.

1.2.1 Overview of the electricity grid

The *electricity grid* is the infrastructure that facilitates the transportation of electrical energy from generators to end-consumers. It consists of a network of smaller sub-grids that are interconnected over large geographic areas and interact with each other [9]. The grid can be subdivided into two operation levels:

1. The *transmission grid* transports generated electric power over long distances to large consumers and distribution regions. Transportation is done over over high-voltage (HV) lines – typically 50-300 kV – to minimise electrical losses.
2. The *distribution grid* consists of the medium-voltage (MV) and low-voltage (LV) networks, responsible for delivering the electricity to individual households or other small to medium size consumers. The voltage in these networks ranges between 400 V (LV) and 50 kV (MV).

The transformation of higher to lower voltages – and vice versa – is done at *sub-stations*. Typically, 40-100 homes are connected to a single LV/MV sub-station. A visual representation of the different components in each level of the electrical grid is given in Figure 1.2.

The electricity grid must be operated in such a way that voltage and frequency levels are within specified limits, and that the physical capacity of all components is not exceeded. Voltage levels react to sudden injections and withdrawals of energy in the network. The frequency (50 Hz in Europe) changes when supply and demand are out of balance. Deviations of frequency and voltage in the system can both lead to failing machines and devices. Reaching the maximum capacity in a part of the grid is called *congestion* and can lead to melting and thus permanently damaging components. Avoiding this damage means that a part of the grid has to be temporarily cut off power, resulting in so-called black-outs.

The transmission system operator (TSO) is responsible for operating the transmission grid, which in the Netherlands is being done by TenneT. The distribution grid is operated by several geographically separated DSOs, of which Stedin is one.

1.2.2 The local energy transition

The worldwide ambition to reduce greenhouse gas emissions is a key driver for innovations and other developments in the energy sector. Residential and commercial buildings account for more than 40% of all primary energy usage and contribute to more than 36% of CO_2-emissions in both the EU and the US [11]. Making buildings and their users more energy efficient is thus an important ingredient in reducing CO_2-emissions. Therefore, the Dutch government has articulated the ambition to make all buildings climate neutral by 2050 [12]. Next to efficiency measures, the main ingredients of achieving these goals are the widespread installation of solar panels[2] and the replacement of heating systems based on natural gas by cleaner resources of heat. Another important development that impacts the grid is the increase of the number of electrical vehicles (EVs). These trends are discussed briefly in this section.

Rapid growth of rooftop-PV penetration

Subsidies and dropping prices of solar photovoltaic power system (PV)-systems are driving a rapid increase of the number of solar panels on the roofs of homes and other buildings. The total installed PV capacity in the Netherlands increased from 146 MW to 2,040 MW between 2011 and 2016, an average annual growth of 69% [13]. During the same period of time, the registered installed capacity of PV in Stedin area increased from 15 MW to 218 MW, an annual growth of 71%, see Figure 1.3. The majority

[2] Another word for ‘solar panel’ is ’photovoltaic power system’, or ’PV-system’. ’Solar panel’ and ’PV’ are used interchangeably throughout this thesis.

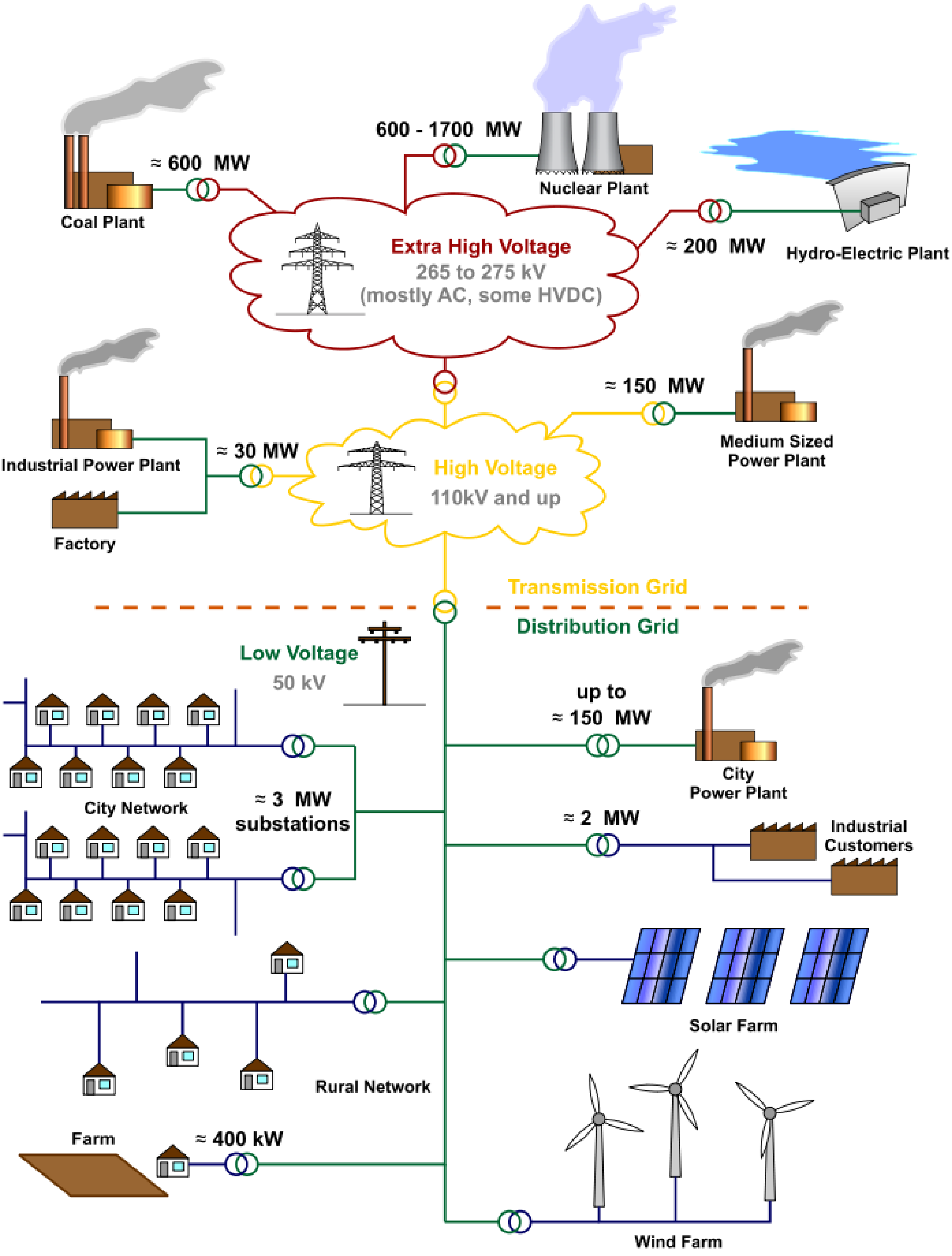


Figure 1.2: A typical electric power grid. From [10].

(85%) of this added capacity is from small and medium size installations on the rooftops of homes and other buildings [14, 13].

Despite the recent growth, the total energy production from PV is still a tiny part of the total energy mix: 0.2% in 2014 [15]. However, the target for the Netherlands is to increase the share of all renewable energy in the total mix from 5.5% in 2014 to 14% in 2020. Because of this and an expected further decline of costs of PV-systems, a growth towards 5 MW installed capacity is projected for 2020 [16].

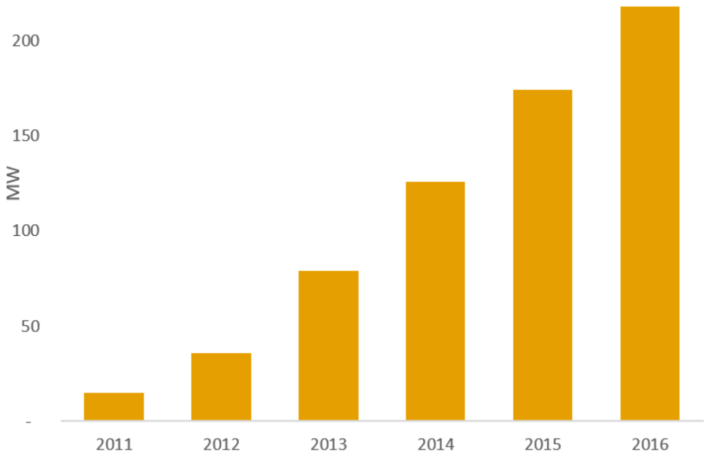


Figure 1.3: Total registered installed PV capacity in Stedin area. In 5 years, the capacity grew from 15 to 218 MW, an annual growth of 71%. Based on figures from [14].

The heat transition

Of the 8 million buildings in the Netherlands, currently 95% are heated by systems that uses natural gas, such as a boiler or a gas stove. However, in a future with a climate neutral built environment, there is no place for natural gas [17]. Therefore, a heat transition is expected to take place.

As concluded in [17], the economically optimal replacement technology for natural gas heating is different for various building types and geographic locations. Technologies like electrical heat pumps (HPs), geothermal heating, district heating and the usage of bio-gas are options that are considered, each with their own pros and cons. The potential of various kinds of heat pumps is specifically of interest for this research, since they are the only heat source with an electric connection to the LV grid.

As visualised in Figure 1.4, the number of electric heat pumps increased from 29,000 to 368,000 during the past 10 years, an average annual growth of 29% [18]. This is around 5% of all buildings in the Netherlands. According to [19], the total potential for heat pumps is between 500,000 and 650,000 installed systems by 2030, based on a growth in both new and existing buildings.

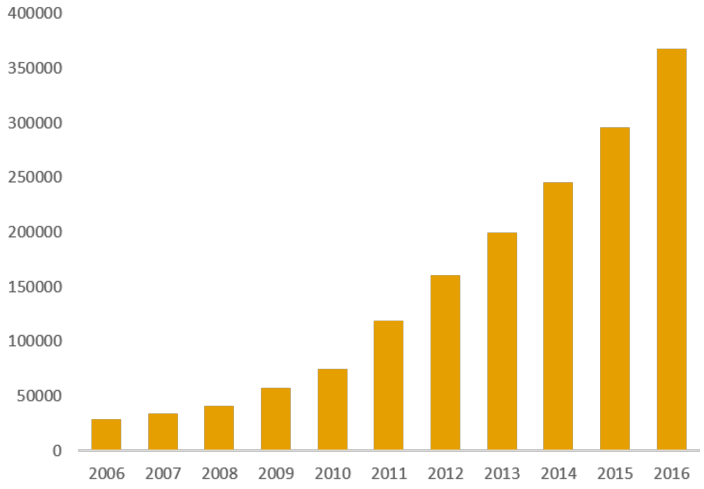


Figure 1.4: Growth of the number of residential HPs in the Netherlands. At the end of 2016, around 5% of all homes had a heat pump, and this number is growing by 20-30% annually. From [18].

The dawn of the electric mobility era

Another transition that is taking place is the one from fossil fuelled to electrical powered cars and other vehicles. The penetration of these electric vehicles (EVs) is driven by the falling cost of storage, economies of scale, and the demand for more environmentally friendly modes of transportation. Figure 1.5 shows the growth in the Netherlands of battery-powered electric vehicles (BEVs), plug-in hybrid electric vehicles (PHEVs) and extended range electric vehicles (E-REVs). From 2012 to 2016, the total number of full and partial electrically powered cars, buses and motorbikes increased from 7,400 to 115,200, almost doubling every year [20]. Future targets for the total number of electrical vehicles in the Netherlands are 200.000 by 2020 and 1 million by 2025 [21].

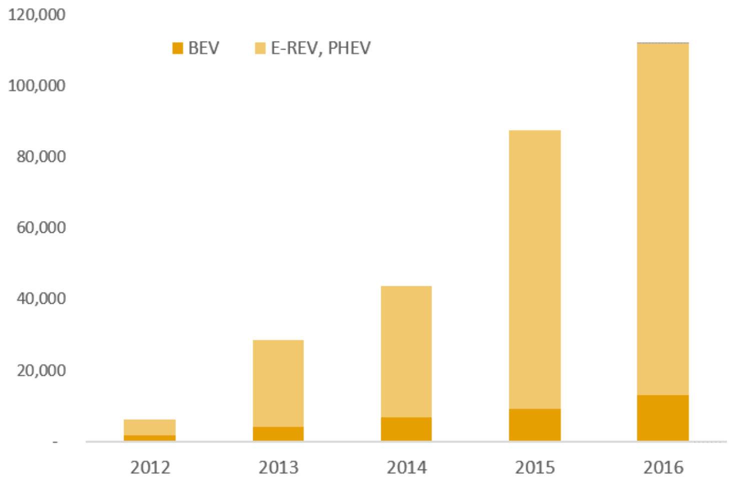


Figure 1.5: Growth of different types of electric vehicles in the Netherlands from 2012 to 2016. The average annual growth of these vehicles has been 99% [20].

In order to charge all these electric vehicles, the presence of sufficient charging infrastructure is essential. Between 2012 and 2016, the number of public charging stations grew from 3,600 to 26,100 (64% annual growth) and the number of public fast charging stations grew from 63 to 612 (75% annual growth). The total number of private charging stations was estimated to grow from 5,000 to 72,000 (95% annual growth) [20].

1.2.3 Challenges for the DSO

Each of the three previously mentioned components of the local energy transition – PV, HP and EV – are expected to have high impact on the grid and how this is operated in the future. All three of them can become a burden or an asset, depending on where they are located and how they are operated. In this section, the challenges that these technologies pose to the Distribution System Operator is discussed.

Prosumers and the Duck Curve

The adoption of PV-systems made it possible for energy consumers to generate and use their own electricity. As a consequence, consumers that were traditionally only extracting electricity from the grid, are now also exporting it when generation exceeds consumption. Agents that are both producing and consuming energy are called *prosumers* [22]. Distributed generation with PV reduces the local electricity demand around mid-day. The electrification of transport and heating will increase the peak consumption at the early evening, when people arrive at home and start charging their electric vehicles, heating their homes with heat pumps, and start cooking on their electric stoves. This combination of decreasing mid-day consumption and increasing evening peaks results in a shift of the aggregated demand towards the so-called *duck curve* [23]. This is shown in Figure 1.6.

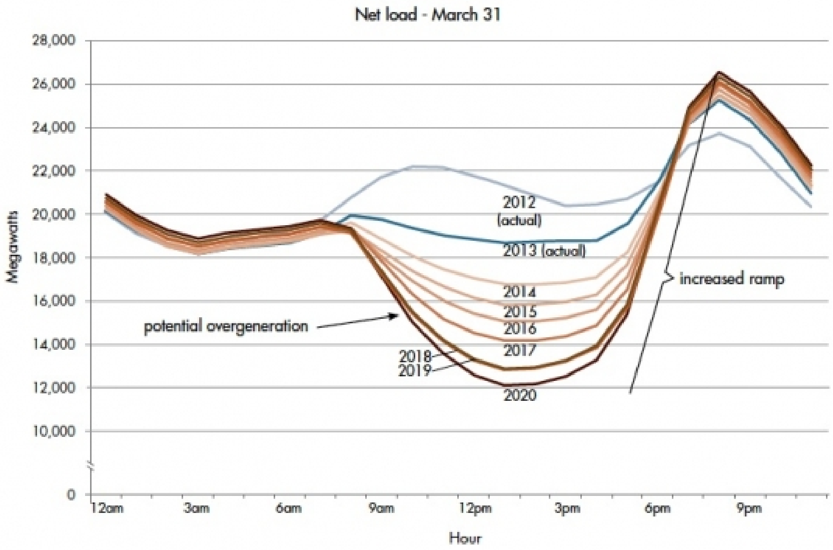


Figure 1.6: Effect of the energy transition on the daily aggregated load. Technologies like PV, HP and EV change the shape from the traditional 'camel'-shaped curve, with a morning and evening peak, to the 'duck'-shaped curve, with a dip mid-day and a steep ramp towards the evening. This figure is based on the Californian situation, but is generally applicable [23].

Congestion and and voltage problems

Due to the previously mentioned trends, the electricity demand in the built environment will likely change in terms of magnitude, variability and predictability. Higher peaks in generation power due to weather-dependent resources, and higher peaks in demand due to the electrification of heating and transportation, might cause congestion problems in the LV and MV grid [24]. An example of this is given in Figure 1.7, in which the aggregated load of several homes, heat pumps and electric vehicles is simulated, exceeding the transformer's capacity.

Distributed generation will require the grid to be able to process a two-way electricity flow, potentially increasing the voltage in the grid and causing stability issues [25, 26, 27]. Germany, having higher PV penetration rates than the Netherlands, has been facing multiple grid frequency and congestion issues during recent years [28].

A fast and uncertain transition

As mentioned in the previous sections, this transition is taking place at a high, uncertain and geographically dissimilar pace. This, combined with the operational risks of congestion and instabilities, poses a challenge for the system operators: expanding grid capacity is expensive and should be avoided if possible, but waiting too long can cause operational dysfunction in the future. In order to keep the grid reliable and the costs low with increasing penetration rates of these new energy technologies (NETs), the system operators are facing the following main challenges:

- monitoring the pace and effects of the local energy transition;
- simulating the effect of this transition in the existing grid;
- understanding what the consumption in new grids will look like;
- Identifying where and when grid expansion is inevitable;
- finding new ways to avoid grid expansion if possible.

In order to achieve this, smart meter data and smart grids offer help.

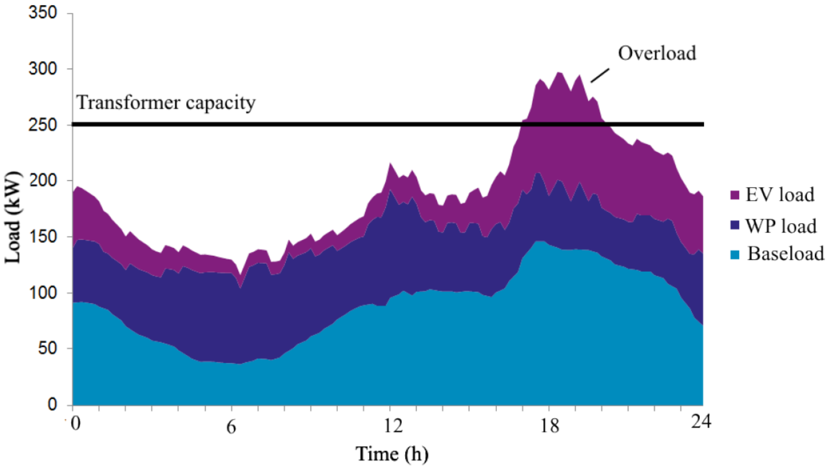


Figure 1.7: An example aggregated load profile of a single day. The aggregated demand of the baseload, heatpumps and electric vehicles cause an overload of the transformer. From [24], based on simulations.

1.2.4 Smart Grid concepts as a solution to future grid challenges

Two trends that are becoming important assets for system operators in dealing with the challenges that were mentioned above are:

1. the fast growing availability of energy consumption data;
2. the increased connectedness of appliances.

These developments enable the *smart grid* to become reality. This smart grid is defined by the U.S. department of Energy as

> *"A fully automated power delivery network that monitors and controls every customer and node, ensuring a two-way flow of electricity and information between the power plant* [or distributed resources] *and the appliance, and all points in between. Its distributed intelligence, coupled with broadband communications and automated control systems, enables real-time market transactions and seamless interfaces among people, buildings, industrial plants, generation facilities, and the electric network"* [29].

Thus, the smart grid enables system operators to gain more insight in, and control over, the network that they are managing. In this section, some benefits from this development are briefly discussed and related to the challenges discussed above.

Leveraging Smart Meter Data

The smart meter is a key technology that enables generating insight in how people, buildings and companies are consuming energy. It is a digital energy metering device that records customer consumption frequently – usually once per 15-60 minutes – and provides for daily or more frequent transmittal of measurements over a communication network such as GPRS [30]. This customer consumption data is referred to as smart meter data (SMD). The smart meter replaces the analogue meters that are read out manually, typically once a year.

Figure 1.8: A smart meter as used by Stedin.

The Dutch government has articulated the target to equip 80% of all active electricity connections with a smart meter by 2020 [31]. This national roll-out started in 2015 and has currently led to a penetration rate of over 30% – 3 out of 8 million households – nationwide by end 2016. In the Stedin area, smart meter penetration has grown at a comparable rate as can be seen in Figure 1.9.

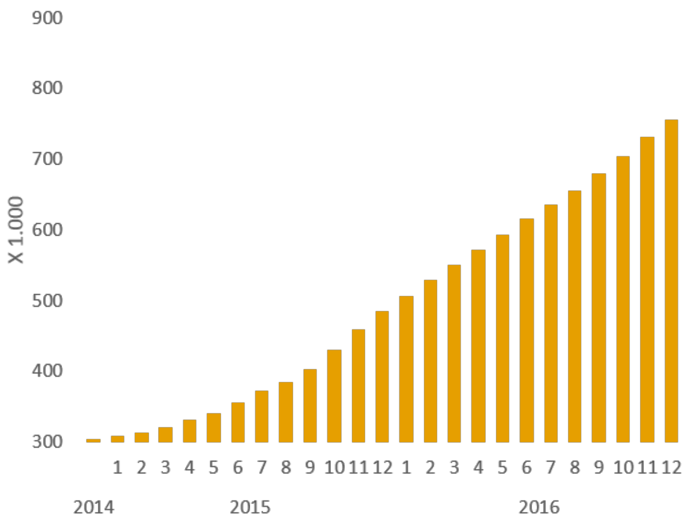


Figure 1.9: Growth of the number of smart meters in Stedin territory. By the end of 2016, 750,000 of Stedin's 2.1 million connections were equipped with a smart meter. Source: Stedin, adapted by Henri Bontenbal

The availability of this high-frequent and near real-time data enables various services, like automatic billing, dynamic tariffing, and providing customers feedback about their energy consumption. For the DSO specifically, smart metering enables outage detection, equipment diagnosis and the ability to gain insight in the status and effect of the energy transition [32].

Demand Response and Flexibility

An important feature of the smart grid is the connectedness of appliances, such as heat pumps and EV charging infrastructure. This two-way connection enables to control the load of these appliances remotely. This is especially useful if the load is flexible, i.e. the perceived utility from an appliance is only indirectly related to its actual operation. People are not concerned about whether their heating is turned on or off, as long as the temperature is within certain comfortability limits, for example $20-22^{o}$C. Therefore, an electric heat pump has a more flexible load than lighting, for which the perceived utility – the illumination of a space – is directly related to the light being switched on or off. The *flexibility* of a load can help balancing the supply and demand in terms of capacity and volume, across time and location, in order to address voltage and congestion problems. Next to heat pumps, other loads that increase the flexibility of the grid are the charging of EVs and batteries. Such loads can react to automated signals from utilities, DSOs or third parties that want to utilise this flexibility, a process called *demand response* (DR) [33]. The principle of shifting flexible load is visualised in Figure 1.10.

1.2.5 The role of Applied Mathematics

Obtaining a better understanding of how people, buildings and companies are consuming energy, how different generating entities are producing electricity, and how this will change as an effect of new technologies and policies, is important for grid planning and investment strategies. The increased volume and variety of available data resources has great potential, but in order to lead to actionable and statistically sound insights, applied mathematics and advanced statistics are essential [35]. Examples of applications that rely on mathematical and statistical principles are:

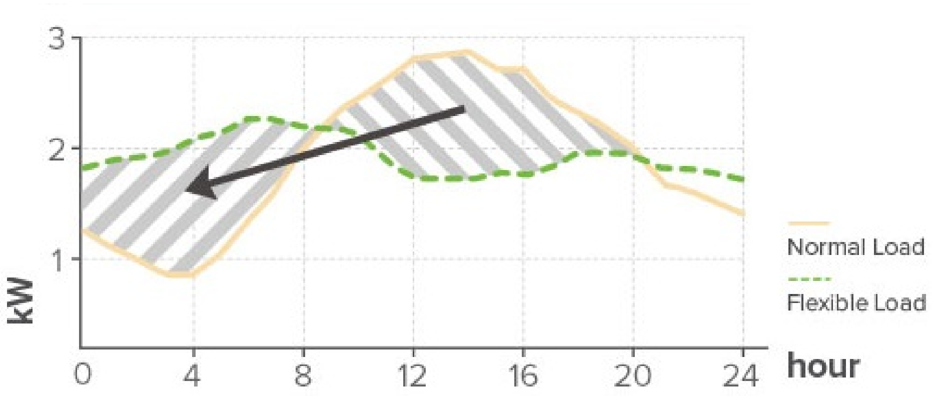


Figure 1.10: Demand response for flexible load, from [34].

- The clustering of energy consumption patterns to summarise the large amounts of highly variable smart meter data. These clustered patterns are called *load shapes* and are daily distributions of energy for one consumer, or an aggregated set of consumers. Load shapes can be used to analyse (peak) demand or as an input for grid simulations [36].
- Disaggregation of the total load into single appliances and technologies in order to monitor the energy transition or to identify flexibility and DR opportunities [37].
- Customer segmentation based on load patterns to assess grid impact of different consumer types [38].
- Simulating the effect of different energy-related developments on the grid for risk analysis [24].

Other grid solutions that largely involve mathematics such as predictive maintenance, energy theft detection, optimisation of storage allocation, and demand response and flexibility optimisation, are beyond the scope of this thesis.

1.3 Thesis Objectives

Smart meter data is becoming a valuable resource to gain insight in how people are consuming electricity. These insights can be used to monitor the local energy transition and simulate its effects on the LV and MV grids. In this thesis, a new method is developed that enables effective customer segmentation and interpretation of the results, leading to a basis for future grid simulation. In this section, the research questions, scope, methodology and contributions of this work, is briefly discussed.

Research Questions

The main Research Question of this thesis is:

Can load shapes be used as a basis for simulating the effect of the local energy transition on the aggregated peak demand and simultaneity?

This question is answered by analysing smart meter data from various sources. Supporting questions to answer this main research question are:

1. What is the current state-of-the-art in the ansalysis and visualization of Smart Meter Data? *Analysis of SMD is a young but active field, and use-cases from practice and literature can serve as a starting point for this research.*

   (a) What software tools are available for energy consumption analysis and visualization?
   (b) What methodologies are most effective in analysing energy consumption patterns?

2. What are the biggest differences in energy consumption statistics between conventional and all-electric communities?
   An explorative analysis of the two available datasets is needed to get a first understanding of how these communities consume energy.
   (a) What are differences in total and peak consumption on an individual consumption level?
   (b) How do the aggregated peak and simultaneity develop for increasing aggregation sizes?
3. What load shapes do summarize the variability in daily consumption patterns in the dataset?
   Daily consumption patterns can vary strongly between different days and homes. To gain insight in these patterns, a set of summarizing load shapes is required.
   (a) How many load shapes are needed to sufficiently describe all daily consumption patterns in the dataset?
   (b) What are the most occurring load shapes in the datasets?
   (c) What is the behaviour of the residuals of individual and aggregated load shapes?
4. Can load shapes serve as a basis for lifestyle based customer segmentation?
   In order to better understand why and how energy is consumed by an individual, a model that infers behaviour based on load shapes is needed.
   (a) What lifestyles can be identified by analysing series of load shapes?
   (b) How do distributions over lifestyles look like for individual consumers?
   (c) How can consumers be segmented based on their estimated distribution of lifestyles?
   (d) Is the developed model successful in identifying PV presence and commercial occupancy from the data?
5. What is the effect of varying the consumer segment mix within a fixed number of homes on the aggregated peak load and simultaneity?
   Now in order to give answer to the main question, simulations are needed to compute the expectation and variance of these statistics.

These questions will act as a basis for the chapters in this report.

Research scope

This research is focused on developments in the LV/MV distribution grid. Therefore, only residential and small to medium commercial buildings are considered, and aggregation levels do not exceed 100 buildings. Based on data availability, a time resolution of 1 hour is chosen and a selection of technologies is made. This scoping can be found in Table 1.1.

Datasets

Two datasets are used to perform the proposed analysis:

- **NL14conv**, a large dataset of conventional homes and small businesses in the Netherlands. Occupation, location and building types are unknown and mixed.
- **NL15ae**, a small dataset of a pilot-project on all-electric homes in the Netherlands. The homes have electric cooking devices and heat pumps. Some have solar PV and/or batteries.

The names of the datasets are anonymised for privacy purposes.

Table 1.1: Thesis scoping

	Inside scope	**Outside scope**
Type of buildings	Residential and small commercial buildings	Large businesses, industrial consumers, governmental buildings
Energy demand	Electricity	Gas
Data	Hourly consumption data	Weather and other context data More frequent data
Technologies	All-electric homes with a HP, some of which have rooftop-PV and a battery	EVs, CHP, etc.
Aggregation level	1 - 100 buildings (LV/MV)	> 100 buildings (MV/HV) single technology profiles

Methodology

Various mathematical methods are used during this analysis:

1. *Adaptive K-means clustering* is used to cluster daily loads into a ‘dictionary’ of daily load shapes.
2. *Latent Dirichlet allocation (LDA)* is used as a generative probabilistic model to make inference on latent lifestyles. Fitting of this model is done with *variational expectation maximisation (VEM)*.
3. *Hierarchical clustering* is used to make a consumer segmentation based on lifestyle distributions.
4. *t-Distributed stochastic network embedding (t-SNE)* is used as a dimensionality reduction method for visualization and improved performance.
5. *Stochastic simulation* is used to simulate the effect of different mixes of consumer segments on the aggregated peak load and simultaneity.

Each of these methods are explained in the relevant chapters. All implementation for analysis and visualization is done in open-source software package **R**.

Contributions

The main contribution of this research is the development of a new model for residential energy demand that:

- models energy consumption as a mixture of lifestyles;
- enables a lifestyle based segmentation of consumers;
- is a framework for load shape based simulation of future energy demand.

Several use cases of this model are worked out for Stedin.

1.4 Outline of Report

Chapter 2 provides an overview of the state-of-the-art in energy consumption analysis, based on literature and experience from a 3-months stay at Stanford University. The chapter concludes with a positioning of this work within this domain. In Chapter 3, the used data sets are described and explored, and a first comparison of energy consumption of conventional and all-electric homes is made. Chapter 4 treats the clustering of daily load shapes with adaptive K-means in Chapter. Latent Dirichlet allocation is introduced as a model to infer lifestyles from series of daily load shapes in Chapter 5, and fitted to the data. In Chapter 6, customers are clustered based on these inferred lifestyles. In Chapter 7, a model validation is performed and two use cases for Stedin are worked out. The last chapter answers the research questions and discusses the strengths and weaknesses of the developed model. This chapter ends with recommendations for future applications and model improvements.

Chapter 2

State of the art and research positioning

In Chapter 1 was concluded that the energy transition is posing many challenges to the electric grid and the parties that operate it. The growing availability of smart meter data is an important resource in dealing with these challenges, since it enables grid operators to gain insight in how people are currently consuming electricity, and to monitor how this is changing over time. However, unlocking the value of this data is not straightforward due to its quantity, quality and dimensionality. Therefore, new tools for data analysis and visualisation are required. Various of such tools for analysing grid data have been developed during recent years. Also, new methods in applied statistics and machine learning are developed to gain advanced insights from these new data sources. However, the field of smart meter data analysis is relatively young, and the demand for new tools and insights keeps growing as the energy transition and growth of daily availability continues.

This chapter reviews the state of the art of smart meter data utilisation for grid analysis. In Section 2.1 some available software tools for smart meter data visualisation are evaluated. Two open source tools that are developed at Stanford University are investigated in more depth. Section 2.2 reviews the literature on three mathematical methodologies that are used for energy consumption analysis: load shape clustering, load disaggregation and consumer segmentation. These first two sections outline the landscape of smart meter data analysis. The positioning of this research within this landscape is given in Section 2.3.

2.1 Software for analysis and visualisation of energy consumption data

The expansion of the amount and diversity of available energy consumption data leads to a strong need for tools to summarise, analyse and visualise these datasets in a fast and clear way. Visualising high volumes of complex and multidimensional data can be a challenging process. However, it is an important ingredient in both the start and the end of the data analysis process. In the beginning of an analysis, visualisation helps forming new hypotheses about the available data. Statistical methodologies can then be deployed to test these hypotheses. When the outcomes of these analyses are unclear, visualisation can help with the interpretation. After the analysis is done, the conclusions often need to be communicated to others that are more distant to the data. Visualisation is essential in bringing across the key findings and solidifying the conclusions.

Multiple software tools for data processing, analysis and visualisation have been developed or are under development. These tools can be either *open source* or *closed source*. Open source means that the code is free available and can be adapted for own purposes by the user. Closed source means that the software is developed by a software company that asks for a fee to use it. Several software packages for grid analysis are being sold by large technical companies like GE and Siemens. Other closed source solutions are (1) *DPG.sim*[1], developed by a spin-off from ETH Zürich, (2) *Utilytics*, developed by Enoro[2], and (3) *Apadtix.Grid* from Sensewaves[3]. Next to these closed source packages, several open source tools are becoming available. The advantage of open source is that these tools are freely available and adaptable for own purpose. The remainder of this section evaluates two open source tools that are currently being developed at Stanford University: VISDOM for demand analysis of individual consumers, and VADER for grid analysis.

2.1.1 VISDOM

VISDOM[4] is a tool developed at Stanford University to quickly analyse large amounts of data from smart meters and other sources, in to generate hypotheses for future analysis. As shown in Figure 2.1, VISDOM consists of four main parts:

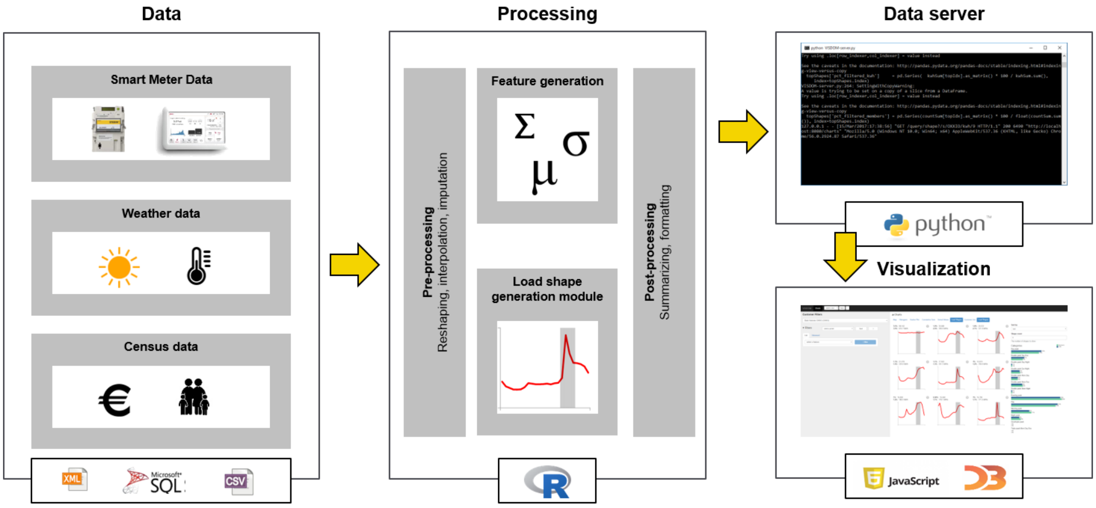


Figure 2.1: A graphical representation of all components of VISDOM.

1. The `DataSource`-module makes the connection to the various data sources that need to be incorporated. This module accepts different sources – smart meter, weather, socio-economic – and formats – .csv, .xml, or SQL database – of data and maps them to the data structures that are used in the computational model of VISDOM. The `DataSource`-object defines not only how to link to the raw data, but also imposes a set of functions that can be used to load, select, manipulate and combine parts of the data.
2. The `feature generation`-module is the main computational part of the software, in which a wide variety of numerical features can be computed. The user of the software can choose to compute the pre-defined default features, such as *yearly energy consumption*, *variance of daily consumption* or the *cooling sensitivity*, but also has the freedom to define his/her own features of interest for tailor-made analysis Examples of these features are given in Table 2.1. Next to this, the `load shape generation`-module can be deployed to identify the frequent occurring consumption patterns in the dataset and to encode the encode consumption data in terms of these

[1] DPG.sim = Distributed Prosumer and Grid Simulation. From `www.adaptricity.com/`
[2] `www.utilytics.com/usecases.pdf`
[3] `www.sensewaves.io/adaptix-grid/`
[4] VISDOM = Visualisation and Insight System for Demand Operations and Management.

load shapes. The underlying methodology and algorithm of this load shape clustering is *adaptive K-means*, developed in [39]. Section 2.2.1 and Chapter 4 explain this in more detail.

3. The `data server` links the computation and visualisation modules, by storing the calculated features for quick access by the visualisation module.
4. The last module is the `visualisation`-module, in which the features and load shapes can be visualised fast and clearly. Numerous filters can be applied to zoom in on specific subsets of the data.

Three main capabilities of VISDOM are introduced in the remainder of this section.

Feature analysis

VISDOM offers flexibility for its user to implement the computation of any statistic of interest. Some examples of these are shown in Table 2.1. After generating this, potentially extensive, set of features, the visualisation tool can be used to quickly scan through a large number of plots. The potential relationships and patterns that these plots reveal can then be formulated into hypotheses. Now the analyst goes back to the processed data to do hypothesis testing on the generated features. This iterative process of feature generation, data visualisation, hypothesis generation, and hypothesis testing can be repeated multiple times. Figure 2.2 shows various generated insights with VISDOM.

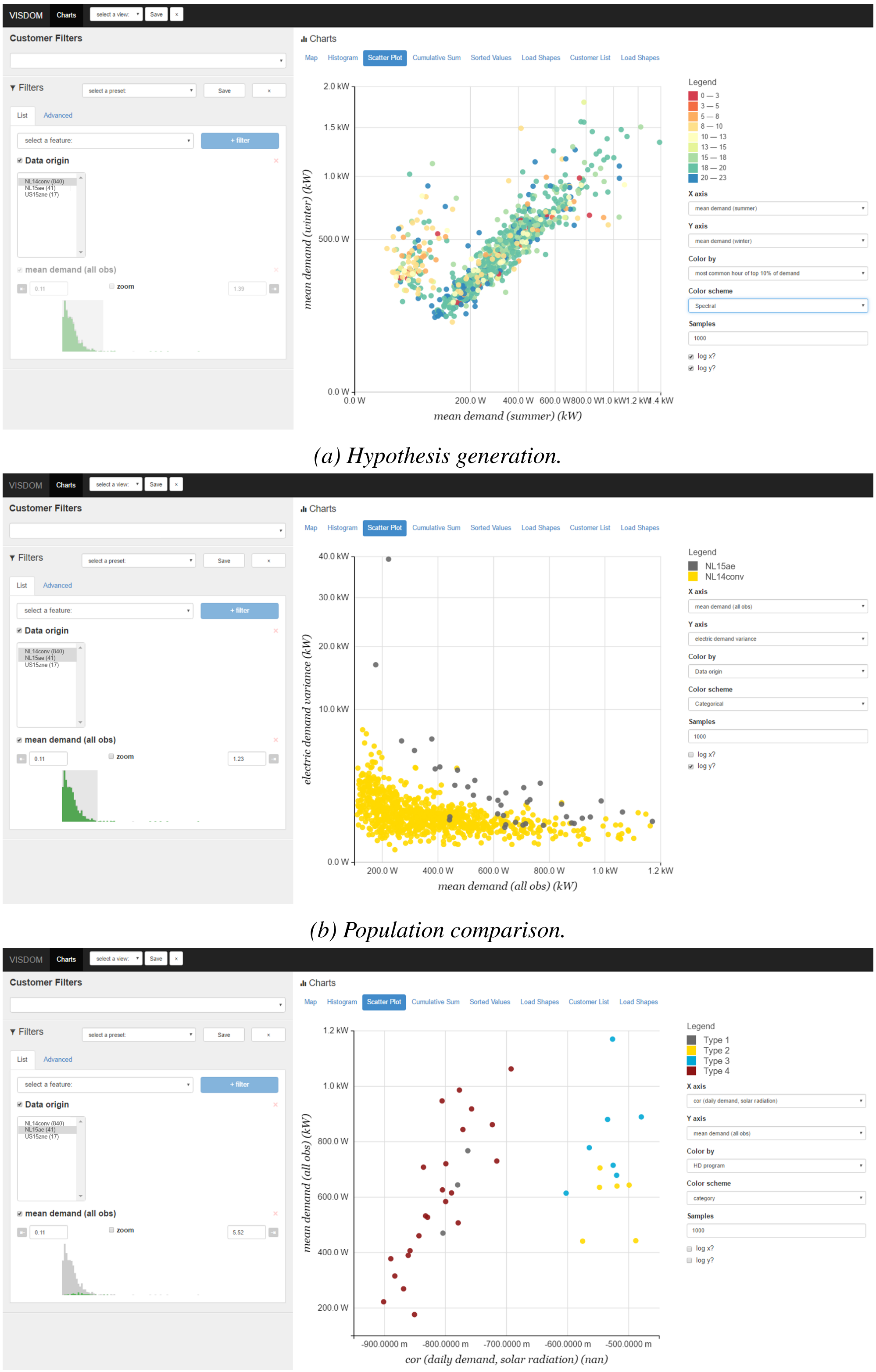
Table 2.1: Different feature types and examples of VISDOM.

Feature type	**Feature examples**
Basic statistics	The average and variance of daily energy demand, total yearly energy consumption, yearly maximum load, and average daily maximum load.
Temporal statistics	The average consumption per hour-of-day, day-of-week, week-of-year, month, weekdays, weekenddays, or seasons.
Extremal statistics	The number of peaks, duration of peaks, hour of highest average demand.
Weather statistics	The correlation with outside temperature, cooling degree days, daily solar radiation, hours of daylight.
Grid impact	Correlation with the aggregated demand at the feeder level or regional/national demand.
HVAC[5]	Estimated parameters by advanced models for AC & electric heating detection, total cooling/heating energy consumption, setpoint estimation, etc.

Spatial analysis

Another way to visualise the generated features with the VISDOM visualisation module is the map-tool shown in Figure 2.3. This tool visualises features of interest over a (subset of) geographical parameters in a map. In this way, energy consumption patterns can be compared on a zip-code or more granular resolution. For example, Figure 2.3 shows the geographical spread of estimated cooling energy over California. These kind of figures can help identifying different drivers of local grid challenges, or to identify which areas need to be targeted to achieve the maximum impact of a demand response program.

[5] HVAC = Heating, Ventilation, and Air Conditionig



☑ Data origin
NL14conv (840)
NL15ae (41)
US15zne (17)

☑ mean demand (all obs): 0.11 – 1.39

X axis: mean demand (summer)
Y axis: mean demand (winter)
Color by: most common hour of top 10% of demand
Color scheme: Spectral
Samples: 1000
☑ log x?
☑ log y?

(a) *Hypothesis generation.*

☑ Data origin
NL14conv (840)
NL15ae (41)
US15zne (17)

☑ mean demand (all obs): 0.11 – 1.23

X axis: mean demand (all obs)
Y axis: electric demand variance
Color by: Data origin
Color scheme: Categorical
Samples: 1000
□ log x?
☑ log y?

(b) *Population comparison.*

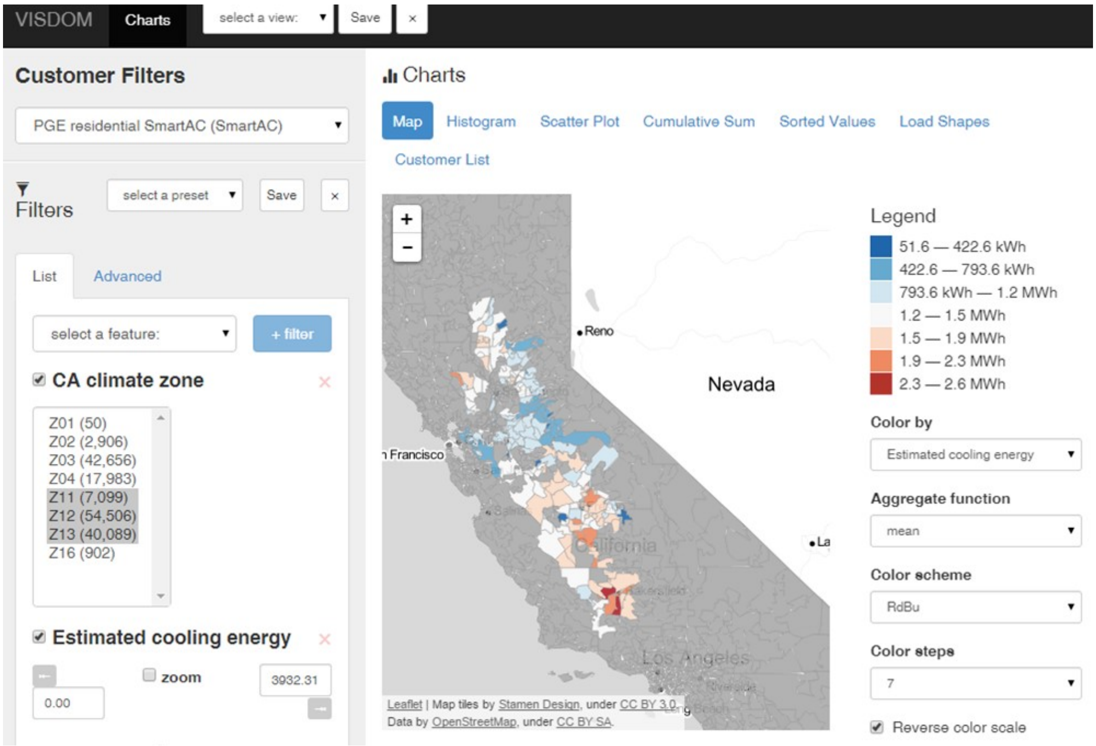
☑ Data origin
NL14conv (840)
NL15ae (41)
US15zne (17)

☑ mean demand (all obs): 0.11 – 5.52

X axis: cor (daily demand, solar radiation)
Y axis: mean demand (all obs)
Color by: HD program
Color scheme: category
Samples: 1000
□ log x?
□ log y?

(c) *Program effect analysis.*

Figure 2.2: Three insights from the feature visualisation capability of VISDOM. The visualisation in (a) leads to the hypothesis that residential and commercial buildings can be separated by the ratio between summer and winter consumption. Visualisation (b) shows that all-electric homes have higher relative variability compared to conventional homes. In (c) it is shown that two out of four battery+solar strategies result in significant less correlation between daily load and solar radiation.



PGE residential SmartAC (SmartAC)

☑ CA climate zone

Z01 (50)
Z02 (2,906)
Z03 (42,656)
Z04 (17,983)
Z11 (7,099)
Z12 (54,506)
Z13 (40,089)
Z16 (902)

☑ Estimated cooling energy

□ zoom

0.00

3932.31

Color by: Estimated cooling energy

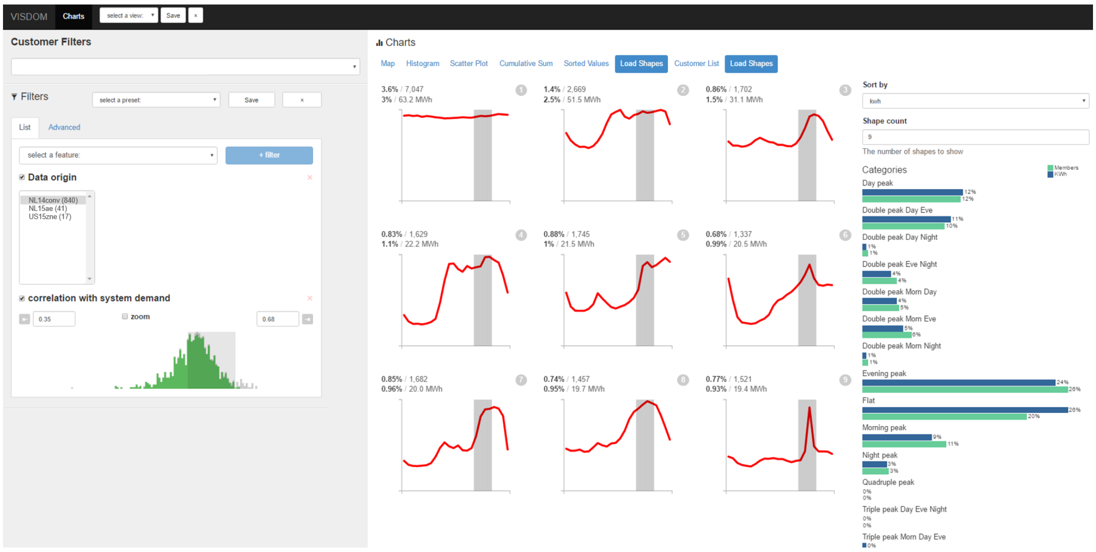
Aggregate function: mean

Color scheme: RdBu

Color steps: 7

☑ Reverse color scale

Figure 2.3: Visualisation with VISDOM of the average estimated amount of energy yearly used for cooling per region in California. Strong fluctuations among regions can be observed. Retrieved from [40].



☑ Data origin

NL14conv (840)
NL15ae (41)
US15zne (17)

☑ correlation with system demand

0.35 □ zoom 0.68

3.6% / 7,047
3% / 63.2 MWh

1.4% / 2,669
2.5% / 51.5 MWh

0.86% / 1,702
1.5% / 31.1 MWh

0.83% / 1,629
1.1% / 22.2 MWh

0.88% / 1,745
1% / 21.5 MWh

0.68% / 1,337
0.99% / 20.5 MWh

0.85% / 1,682
0.96% / 20.0 MWh

0.74% / 1,457
0.95% / 19.7 MWh

0.77% / 1,521
0.93% / 19.4 MWh

Sort by: kwh

Shape count: 9

The number of shapes to show

Figure 2.4: Visualisation with VISDOM of the most common load shapes. The most common used load shapes (3.6% of the time) is the flat load shape. Other often recurring load shapes have an evening peak or a morning and an evening peak.

Load shape analysis

The third functionality is the *load shape*-module. This module displays the most common daily consumption patterns, so called *load shapes* in the data. A more exact definition of a load shape is worked out in Chapter 4. This module helps answering the question how different electricity consumers use their electricity over time. Figure 2.4 visualises such analysis, showing the 9 most common load shapes of the NL14conv-dataset.

2.1.2 VADER

To get more insights in the effect high penetration of distributed generation (PV) and demand side developments (EV, DR, HP) on the grid, Stanford University and the SLAC National Accelerator Laboratory developed the software package VADER[6]. Unlike VISDOM, which is focused on demand analysis, VADER looks at the distribution system itself. This data analysis platform enables integration of various large data streams in the distribution grid for real-time monitoring, analysis and control of distributed energy resources [41]. Both analysis of the current situation and impact analysis of different future scenarios are incorporated. Because of this integrated grid approach, VADER is potentially a valuable tool for Distributed System Operators[7].

2.2 Energy consumption analysis

The introduction of smart meter data has enabled various advanced analytical methodologies that give valuable insights for parties that act on the grid. *Load shape clustering, load disaggregation* and *consumer segmentation* are three of the most used approaches for consumer load analysis. A literature review of each of these three concepts will follow in this section.

2.2.1 Load shape clustering

For grid planning and load forecasting, utilities are currently using standardised consumption profiles, based on average consumption patterns. Such a standard load shape is shown in Figure 2.5. However, these load shapes often do not capture the variability among consumers on a geographical scale and over time. With the expected increased penetration of EVs, PVs and HPs, this variability will even increase. Therefore, an approach that zooms in on these individual and technological differences is required. Smart meter data is an important resource for acquiring such a granular insight.

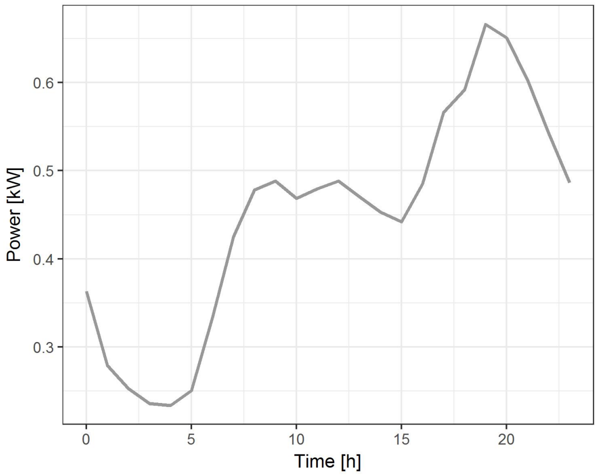


Figure 2.5: A typical load shape, based on standard profiles from NEDU [42].

Load shape clustering is the task of identifying consumption patterns that summarize how electricity is consumed over time, typically a day. These daily load shapes are a useful input for load forecasting and grid impact analysis [43]. In conventional grid analysis standard profiles are used, but these profiles do usually not capture the natural variety in energy consumption. Various clustering algorithms have been proposed in literature to construct sets of these load shapes.

[6] VADER = Visualisation and Analytics of Distributed Energy Resources

[7] For more information, visit `www.sites.slac.stanford.edu/vader/`.

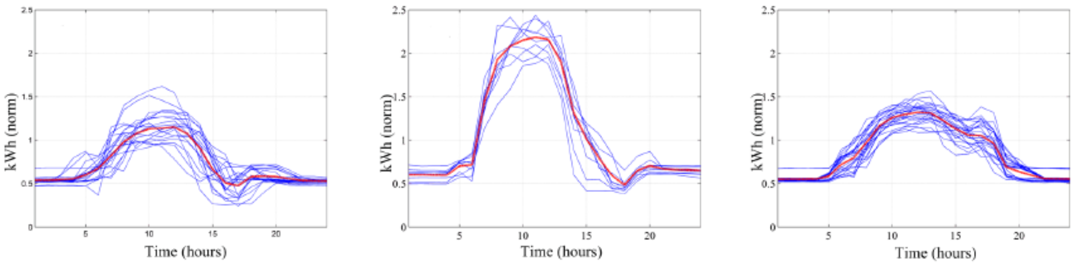


Figure 2.6: Examples of results from load shape clustering. Three inferred cluster centres are visualised in red, with the corresponding observations in blue. The used methodology is hierarchical clustering with Euclidean distance [44].

Literature review

Many clustering approaches are available for a wide variety of applications. In this section the approaches that have been proposed for energy demand clustering are discussed. Three characteristics of each method are discussed, namely:

1. The **representation** of the load data that is used for clustering. This representation can be the raw energy consumption data or some calculated numerical features, like the mean and variance of consumption. Another option is to transform the raw data to some lower dimensional space with procedures like principal component analysis (PCA) or canonical correlation analysis (CCA).
2. The **similarity** measure or dissimilarity measure used in the algorithm to define what objects are *close* or *distant* to each other.
3. The **algorithm** used for grouping together the observations.

In [45] two algorithms, the self-organising maps (SOM) and the follow-the-leader algorithm are used for load shape clustering. [46] uses SOM for dimension reduction and K-means for clustering of load shapes. A decision tree is used for the clustering of consumers based on these load shapes. SOM is used to filter, classify and extract load patterns and monitor the evolution of consumption patterns over time in [47]. In [38] various clustering algorithms (hierarchical clustering, K-means, fuzzy K-means, SOM) and three dimensionality reduction methods (PCA, CCA and Sammon maps) are compared. Their purpose is to classify different consumer types for segmented tariffing. They find that follow-the-leader and hierarchical clustering with average distance linkage emerge as the most effective algorithms. Also, Sammon maps have provided slightly similar clustering validity for different data size reductions, so indicating a robust behaviour, whereas the CCA has exhibited the best performance but is less robust. Their main conclusion is that a total of 15-20 load shapes could fit the energy supplier's needs. In [48] K-means, adaptive vector quantisation (AVQ), fuzzy K-means and several hierarchical agglomerative clustering algorithms are considered for load shape identification and consumer segmentation. K-means and hierarchical agglomerative methods presented the best results. [36] uses hierarchical clustering, SOM, and K-means for load shape segmentation. In [44] different similarity measures are compared: the Euclidean distance, Mahalanobis distance, Pearson correlation and Dynamic time-warping (DTW). They conclude that Euclidean distance obtains the most balanced solutions, while DTW can give improvements in certain applications. In [49] DTW is used as a distance measure combined with the hierarchical clustering methodology. This results in a relatively smaller number of classes and higher clustering efficiency measure, comparing to the Euclidean norm and K-means. In [39], the adaptive K-means algorithm is combined with hierarchical clustering to construct a rich dictionary of load shapes from normalised daily loads.

2.2.2 Load disaggregation

In order to better understand the energy consumption of a building, it is often of interest to identify what appliances are installed and when they are being used. This can be done by analysing detailed energy consumption measurements, an approach called appliance load monitoring (ALM). Two major approaches to ALM are known. The first is Intrusive Load Monitoring, which uses multiple sub-meters to get the required insight. However, because of its "intrusiveness" this is not an option for most utilities and other parties. Therefore non-intrusive load monitoring (NILM) is often used as an alternative, because it only uses one single meter per building and disaggregates this signal into the consumption of single appliances. A visual example of NILM is given in Figure 2.7. An often used application of these results is giving feedback to households about their energy consumption and suggesting ways to be more energy efficient [50, 51].

Literature review

A comprehensive overview of existing NILM approaches has been given recently in [52] and [50]. An important conclusion in these articles was that for effective disaggregation, the optimal time resolution is in the order of 1 second - 1 minute. The currently used time-resolution of 15 - 60 minutes for smart meter data is not of sufficient granularity for most NILM applications. What can be achieved with these low-resolution meters is a segregation into general categories like 'base load' and 'variable load', or estimations of energy consumed by heating, ventilation and air-conditioning (HVAC) or produced by PV [50, 53, 54, 37, 55].

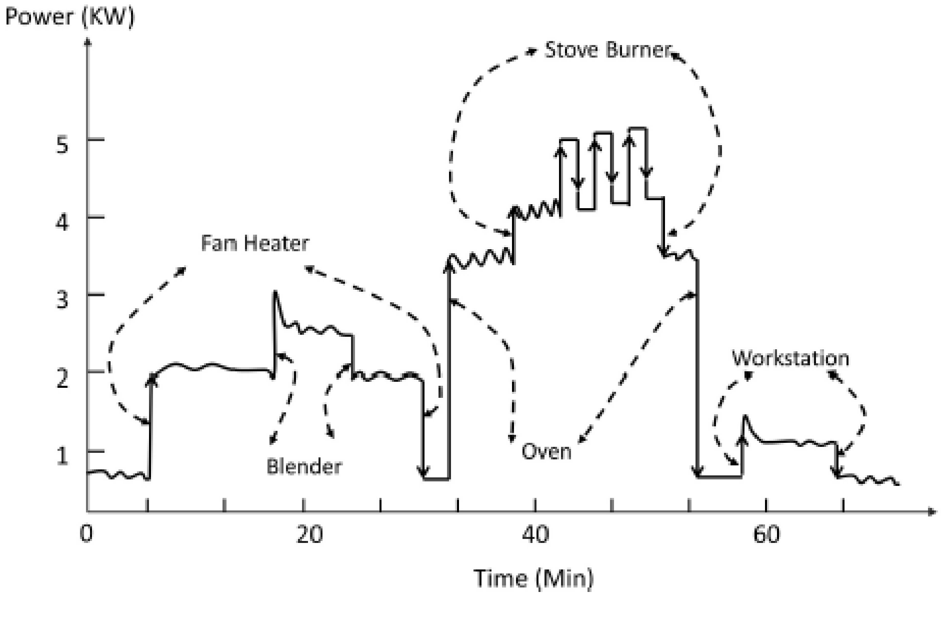


Figure 2.7: Visualisation of load-disaggregation with non-intrusive load monitoring (NILM) [52].

2.2.3 Consumer segmentation

Consumer segmentation is used in many industries to subdivide the consumer base into groups that have similar consumption characteristics. In an energy consumption context, consumer segmentation can be used to distinct between different consumer types, for example residential versus commercial consumers, or families versus students. This enables classifying grid impact scores –high, medium or low impact – or identifying lifestyles of consumers. Segmentation can be performed on raw load data or load shape clusters, but also on other available context information, like socio-economic data. The latter is called psychographic segmentation.

Literature review

There is broad literature on the segmentation of energy consumers. Segmentation can have various purposes. A widespread application of energy consumer segmentation is improving tariff structures [45, 46, 48, 56]. A segmentation is used by [57] for aggregate demand forecasting. The results of [58] were used by an energy utility for increasing the effectiveness of their communication about energy conservation. [59] and [60] use consumer segmentation for the assessment of DR potential.

Psychographic segmentation aims to classify consumers based on data about how people feel, think and act. The data used is typically a survey with questions about attitude and behaviour. [58] uses survey data about attitude towards electricity conservations and segments 6 types of consumers based on their progressiveness. A similar segmentation is performed by [61], but in addition to focusing on how people's attitude towards conservation of energy, they focus on how likely people will purchase energy-saving appliances and goods. [62] zooms in on what and how often appliances are being used by each consumer. One of their main findings is that people in cities consume less energy because they use less AC and laundry machines.

With the introduction of smart meter data, survey data can now be replaced or enriched by actual energy consumption data. Various approaches of this type have been suggested in literature. [63] clusters 471 commercial consumers based on load shapes to identify inefficient billing practices. [64] clusters 660 consumers into clusters using iterative self-organising data analysis, with as input the yearly load profile and several weather dependency parameters obtained through a regression analysis. [46] identifies groups of consumers with similar energy consumption behaviour by using self-organising maps, and suggests a decision tree to assign new unclassified consumers to a cluster. [59] uses the distribution of load shapes - the output of [39] - as a basis for consumer clustering. A K-means algorithm, with cosine distance measure, is used to cluster the load shape frequency vectors per consumer. The outcome is used for assessing the potential of energy program targeting. [36] have a similar application of their consumer segmentation approach, but clusters based on peak timing rather then load shape series. They also use a modified K-means algorithm for the segmentation. [60] uses a hidden Markov model to model energy consumption of households, which are then clustered by using spectral clustering, a graph-theoretic segmentation technique. A similar approach is followed by [65], who first performs a symbolic aggregate approximation and then applies a time-based Markov model to model the dynamic of the electricity consumption. The resulting state transition matrices are clustered by fast search and find of density peaks, and the distance between any two consumption patterns is measured by the Kullback-Leibler divergence. Another approach by [66] uses the power demand distribution of a home and the two-sample Kolmogorov-Smirnoff statistic as a similarity measure and the K-Medoids algorithm for clustering. [67] uses both socio-demographic factors and load data to segment into different consumer classes, so-called lifestyles. Their main findings are that three factors (household size, net income and employment status) have the biggest effect on energy consumption and load shape behaviour. [68] designs a generic framework for consumer segmentation tasks for combined smart meter and survey data. They use various feature representations (mean, median, standard deviation, IQR) and different temporal contexts and use a K-means algorithm to show some results of this framework.

2.3 Positioning and novelty of this research

As was shown in this chapter, smart meter data analysis is a relative new and active field in which many different approaches have been suggested in the past decade. consumer segmentation is often used for tariff diversification or grid impact analysis, and a variety of methods for this has been proposed. Many of these use load shapes as a feature base for segmentation. These load shapes can at their turn be generated in a wide variety of ways. However, this type of consumer segmentation assumes that consumers can be classified as belonging to one type. This is a simplifying assumption, since homes and their occupants can often not be put into one box. To capture this behavioural variety in a more realistic way, this thesis proposes a topic mixture model, which assumes the existence of some unobserved lifestyles. These lifestyles live in between the load shapes space and the consumer classes. By doing this, consumers are no longer characterised as belonging to just one consumer type, but by a mixture of lifestyles. This allows for a more specific and interpretable classification of consumers that is useful for a broad variety of applications. Furthermore, this model generalises to a generative model that can act as a basis for aggregated load and grid simulation in the future.

Chapter 3

Data description and exploration

Chapter 1 presented the context and goals of this research. and Chapter 2 elaborated on the state-of-the-art in energy consumption analysis and the positioning of this thesis. This chapter introduces and explores the datasets that are used in this research. Section 3.1 introduces the two smart meter datasets: one containing consumption of conventional homes and businesses, and one containing consumption of all-electric homes. In Section 3.2 the quality of these datasets is assessed, and the pre-processing of the data is explained. An explorative analysis follows in Section 3.3. In order to get an understanding of how energy is consumed in each of the two datasets, four characteristics are evaluated. Section 3.3.1 compares some basic statistics about daily and annual consumption. Then, in Section 3.3.2, the hourly load curves will be compared with the standard profiles that are currently used for analysis. Section 3.3.3 explores some statistics about daily 1-hour peak loads on both single home and aggregation level. Lastly, Section 3.3.4 introduces the concept of simultaneity and compares how the two datasets differ in simultaneity and aggregated peak load. The chapter concludes with the key insights and an overview of the differences between the two datasets.

3.1 Used datasets

Two main datasets are used within this research: one containing a large mix of different buildings and one containing all-electric homes. The datasets are encoded for privacy purposes.

- **NL14conv** is a large dataset of energy consumption data from a Dutch electricity provider. The customers are not geographically specified, can be both commercial or residential consumers, and potentially occupy a large variety of building types. However, no specific information about this is provided. The time series have a 1 hour resolution and the dataset contains 19,991 non-empty files with data gathered in 2014.
- **NL15ae** is a smaller dataset, containing all-electric high efficiency homes from a pilot project with 42 newly built homes in the Netherlands. The homes can have PV systems and/or battery storage and have different demand response strategies. This dataset contains observations for the year 2015.

More information about these datasets is given in Table 3.1.

Table 3.1: Description of the datasets used.

	NL14conv	**NL15ae**
Type of buildings	Mix of conventional residential and commercial buildings.	New residential buildings, high efficiency, all electric. Some have PV and/or batteries.
Resolution	1 hour	15 minutes
Year	2014	2015
# homes (of which > 95% coverage)	19.991 (5,152)	42 (42)
# consumption-days (complete for > 95% coverage)	6.740.499 (1.790.642)	14.802 (14.721)
Avg. missing values	32.2%	3.6%
Size of dataset	551MB	18MB
Context information	-	Type of technology, type of battery/DR proposition.

The NL15ae-dataset is originating from a pilot study in which 42 all-electric homes participated. The participants could choose out of four propositions that were tested in the study, some of which included solar PV and/or batteries. The details of these proposition are presented in Table 3.2.

Table 3.2: All-electric propositions

Proposition	PV	Battery	Description	Number of homes
A	Yes	Yes	Generated electricity stored in local battery	24
B	Yes	No	Generated electricity and smart appliances	3
C	No	Yes	Battery charged during night time and discharged during the day	8
D	No	No	Only measurement data is provided	7

3.2 Data quality

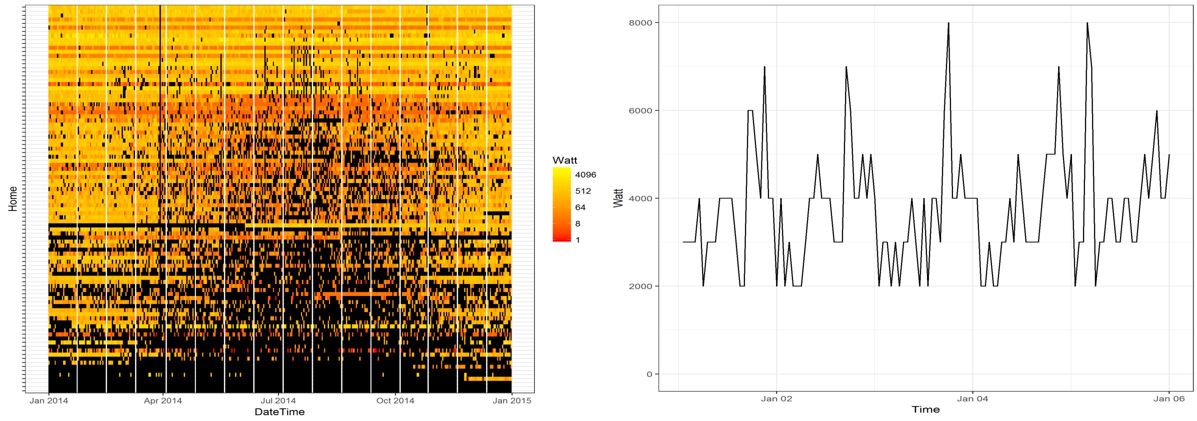
Before giving a deeper insight on what this data looks like, first the quality of the data needs to be assessed.

3.2.1 Missing values and anomalies

Two main problems have been identified looking at the NL14conv-dataset.

1. **Missing values.** number of missing values is relatively high. The 19.991 buildings in the dataset have on average 32.2% missing values in the year 2014, and only 25.9% (5.173) of these buildings have less then 5% missing values. See Figure 3.1a for a visual representation of this.
2. **1000W-anomaly.** Some buildings have unexplainable energy consumption patterns that only expose consumption levels that are a multiple of 1000W, like the one shown in Figure 3.1b.

The next section explains how these data issues are dealt with.



(a) Heatmap of consumption of 100 homes.

(b) Suspicious consumption pattern.

Figure 3.1: Overview of data issues in the NL14conv-dataset. A heat map of daily consumption of 100 homes is shown in (a). Black values are missing value and homes are sorted from most (bottom) to least (top) missing values. In (b) the anomalous behaviour of one of the buildings is presented, showing observations that are all a multiple of 1000.

3.2.2 Data preparation

There are several approaches to deal with these data quality issues. The simplest one is to delete all data that does not meet a required quality level. This is also the one that is used in this thesis. The motivation for not using data imputation is that for the development and validation of this method, the quantity of data is not a limiting factor, even after throwing away a majority of the dataset. An important assumption that is made here, is that this does not introduce a selection bias. In other words, the errors and missing values in the data are not related to the type of consumer. Now the data is selected as follows:

1. Select customers with at most 5% missing values,
2. Select customers that have at most 2% of their values a non-zero multiple of 1000W.

This results in a subset of 5,152 homes. This selection process is shown in Figure 3.2. From the NL15ae-dataset, all 42 homes will be used and no significant anomalies were detected. If in a future application inference must be made about consumers that have data of limited quality, several imputation methods can be used. The author proposes a multiple-imputation method, using a statistical model to generate several copies of the dataset, with missing values replaced by generated values from the statistical model. Analysis can then be performed on each of these datasets, after which averaging can be used to come to one main conclusion. An R-package that implements such an approach is `AMELIA`, which uses expectation maximisation and bootstrapping for multiple imputation [69].

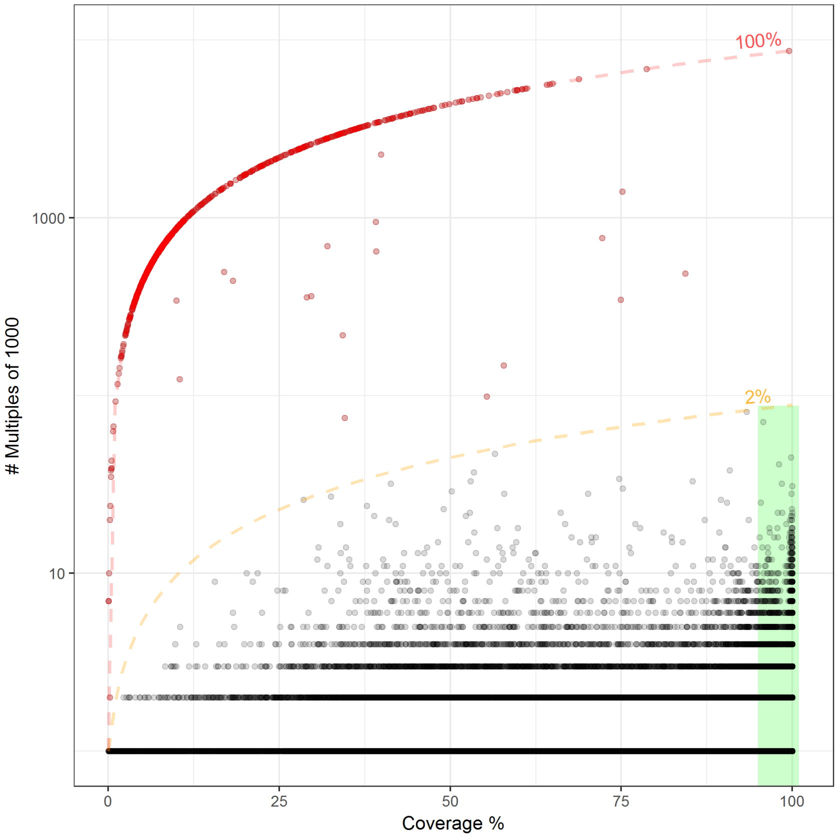


Figure 3.2: Selected buildings in the NL14conv-dataset. The red dots are buildings with the anomalous behaviour shown in Figure 3.1b. The green square shows the selected data, with maximum 5% missing values and 2% equal to a multiple of 1000W.

3.3 Explorative data analysis

Looking at the raw data, a large variety of load patterns and consumption levels can be observed. One of these is plotted in Figure 3.3. This building exhibits energy consumption that seems like it is coming from a dwelling, with a small peak in the morning and a large peak in the evening. Various other load profiles are shown in Figure 3.6, showing the variety of consumption pattern in the dataset. This variability suggests that this dataset is not containing only homes, but possibly also non-residential consumers like shops and offices.

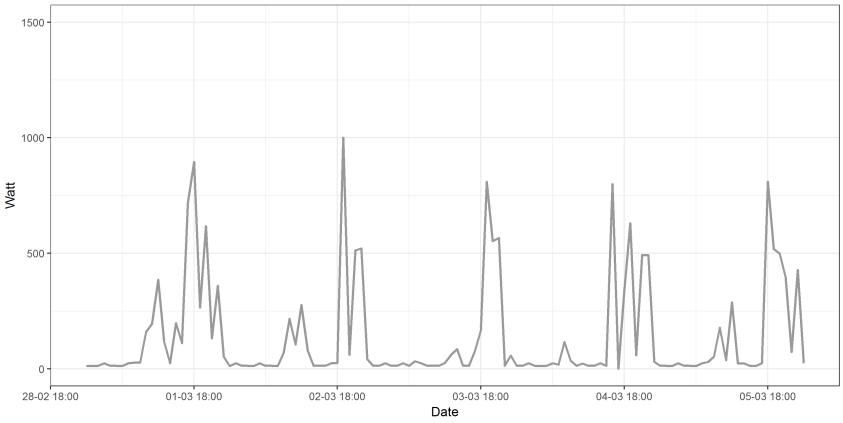


Figure 3.3: Energy consumption over several days of a single consumer. The small morning and large evening peaks suggest this is a residential consumer.

The remainder of this section takes a closer look at what type of consumers are in the datasets and how this relates to the customer base of Stedin. First, the distribution of annual and daily energy consumption is compared with a dataset of energy consumption from Stedin customers. Then the hourly load data is explored and compared with the standard load shapes currently used. Next, the distribution of individual and aggregated peak loads is investigated. The last section introduces the concept of simultaneity and evaluates this for both datasets.

3.3.1 Total energy consumption per consumer

As mentioned, the NL14conv-dataset consists of a mixture of different types of buildings and consumers, not restricted to the Stedin domain. The NL15ae-datset consists of high efficient homes from a pilot project. Thus, both datasets might not be representative for the average Stedin client. Therefore, the distributions of annual aggregate energy consumption of each of these datasets is assessed.

Annual energy consumption

For reference, the two datasets are compared with a dataset containing total energy consumption of 10,658 E1B-connections[1] in the Stedin area in 2016. As shown in Table 3.3, the annual consumption of NL14conv-buildings is on average 29% lower then for E1B-buildings. Figure 3.4 shows that the NL14conv-dataset has many small consumers, while the Stedin-E1B has a ‘fat tail’: the dataset has a small number of extremely large consumers compared to the mean. The hypothesis is that the NL14conv-dataset contains many residential consumers with small consumption – perhaps apartments – while the E1B-dataset contains more medium size commercial consumers, like offices and shops. The NL15ae-homes consume on average 22% more on a yearly basis compared to the E1B consumers. This is explained by the fact that the generated electricity by solar panels does not compensate for the extra consumed electricity from heating and cooking.

Table 3.3: Energy consumption statistics per dataset. Energy in kWh.

Dataset	n	μ	σ	min	q_{25}	median	q_{75}	max
NL14conv	$4,735$	$3,463$	$2,484$	843	$1,916$	$2,942$	$4,257$	$48,402$
NL15ae	41	$5,910$	$1,559$	$2,529$	$4,085$	$5,606$	$7,330$	$13,173$
Stedin-E1B	$10,658$	$4,835$	$4,835$	2	$2,626$	$3,936$	$5,711$	$187,008$

In Figure 3.4 the empirical density functions of annual energy consumption of each dataset are visualised. The distribution of E1B is skewed, with the bulk of the data consisting of small consumers and a small number of very large consumers. This *fat tail* is also observed in Table 3.3, with the largest consumer consuming 39 times the average. The NL14conv and NL15ae datasets also look skewed, but not as much as the E1B data, suggesting that the consumers within each of these 2 datasets are more similar.

A non-parametric test is used to test whether these distributions are significantly different. The used test here is the Wilcoxon-Mann-Whitney test [70]. The p-values in Table 3.4 reject the null-hypotheses that any two of the three datasets originate from the same distribution, which corresponds with Figure 3.4.

Daily energy consumption

In Figure 3.5, the relative consumption per day is plotted as a fraction of the average daily consumption. The median and quantiles of the NL14conv-dataset are plotted and compared with the predictions of

[1] E1B is a NEDU-code for consumers with a 3x25A-connection, both residential and commercial customers using night tariff. See [42].

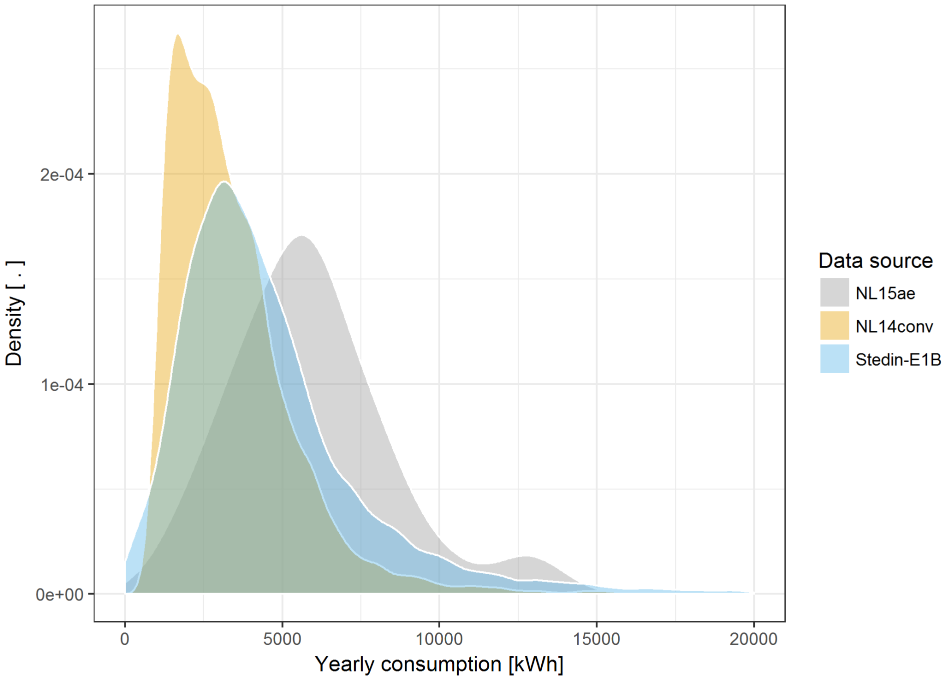


Figure 3.4: Distribution of annual energy consumption. The NL14conv-dataset has a significant lower annual consumption, while the consumers from the NL15ae-dataset consume significant more then the reference data. P-values for the Wilcoxon-test are given in Table 3.4.

Table 3.4: P-values of the Wilcoxon-Mann-Whitney test for similarity between each two datasets.

	NL14conv	NL15ae	STEDIN
NL14conv	1	$9.2e^{-12}$	$4.8e^{-153}$
NL15ae	$9.2e^{-12}$	1	$8.0e^{-51}$

NEDU [42]. The trend over the whole year of the NL14conv-dataset is similar to the E1B-predictions. However there seems to be a stronger weekly seasonality.

3.3.2 Hourly load profiles

In the previous section was concluded that the NL14conv and NL15ae datasets have a total consumption that is of a significant different – respectively smaller and larger – magnitude than that of the average Stedin consumer. Next to *how much* electricity is being used, it is also of interest look at *when* electricity is being consumed.

Load shapes

First the raw consumption data is plotted for four homes in Figure 3.6. More examples can be found in Appendix A. These plots show a strong variability in shape and magnitude, and support the hypothesis that this dataset contains a multiple of different consumer types.

Average loads

When the hourly energy consumption is scaled by the annual mean energy consumption, it is possible to distinct load shape from magnitude. Now these normalised loads can be compared with the E1B predictions, which are also normalised over a year. This comparison for two full weeks, one in winter

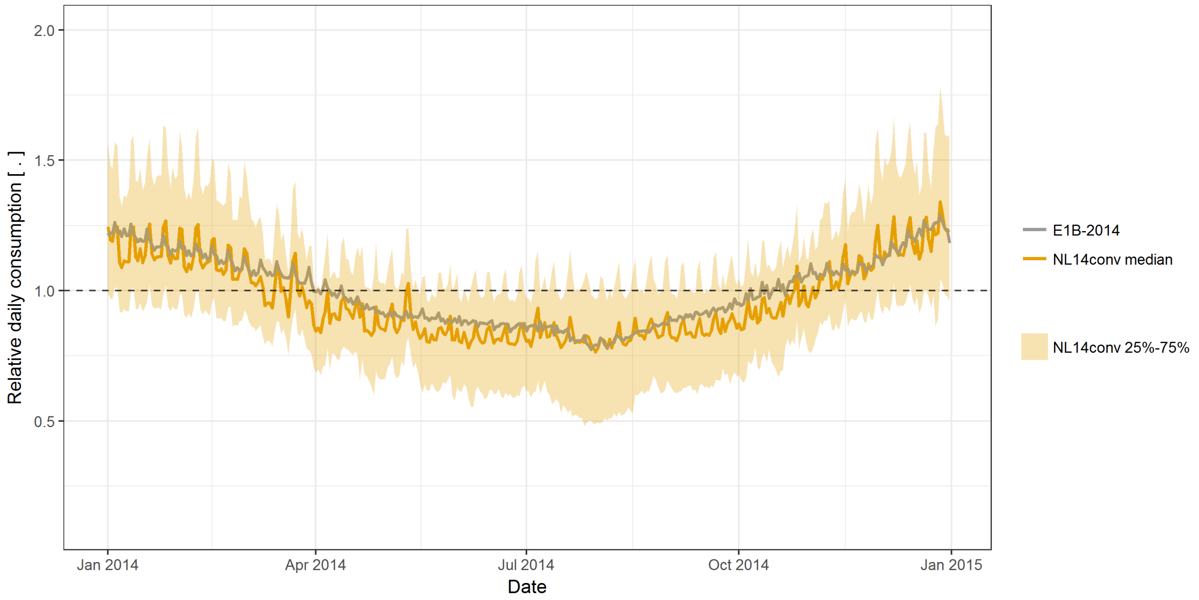


Figure 3.5: Trend line of daily energy consumption relative to annual mean. The trend over the year of the NL14conv-data is similar to the predictions for E1B connections by NEDU [42]. The week-seasonality seems a bit stronger in the observed data, which could imply a different blend of commercial and residential buildings.

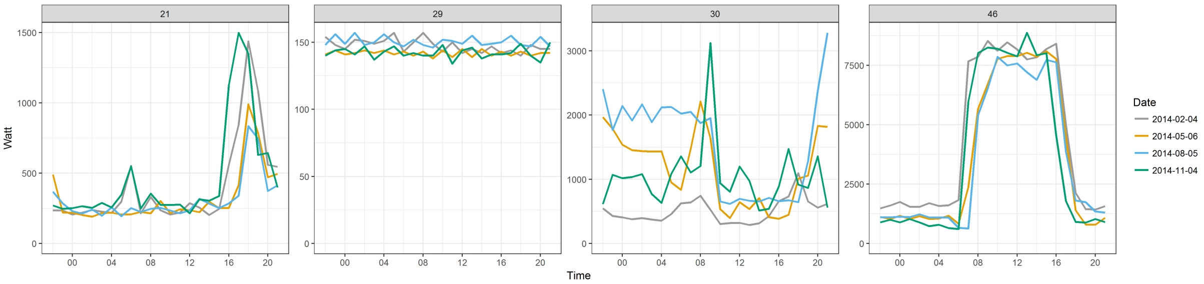


Figure 3.6: Daily loads of four consumers for four random Tuesdays. Consumer 21 has a typical residential load shape, with a small peak in the morning and a larger peak in the evening. Consumer 29 has a constant, slightly oscillating, and relatively low load. Consumer 30 is more chaotic and has a load that is not easily interpretable. The load of consumer 46 is very stable, and likely to come from a commercial building.

and one in summer, is shown in Figure 3.7. During both weeks the E1B-profiles seem to follow the mean profile quite well. For July 27th the NL14conv-mean and E1B-forecast are almost identical.

3.3.3 Peak demand

DSOs like Stedin design the grid to have sufficient capacity at peak times. Therefore it is of interest to look at the distribution of the magnitude and timing of daily and yearly peaks, as well as to what extend the individual peaks are averaged out due to difference in timing, which is quantified in the 'simultaneity'-measure.

Peak magnitude

First the peak magnitude is explored. In Figure 3.8a the distribution of daily peak loads per home is visualised. In general NL15ae-homes have far higher peaks (μ = 2.22 kW) than those in the NL14conv-dataset (μ = 1.16 kW). In Table 3.5 statistics about daily peak loads per home can be found.
Since grids are designed on annual (or even longer term) peaks, it is also of importance to look at the annual peaks per home. This distribution is plotted in Figure 3.8b. Also here it can be seen that the

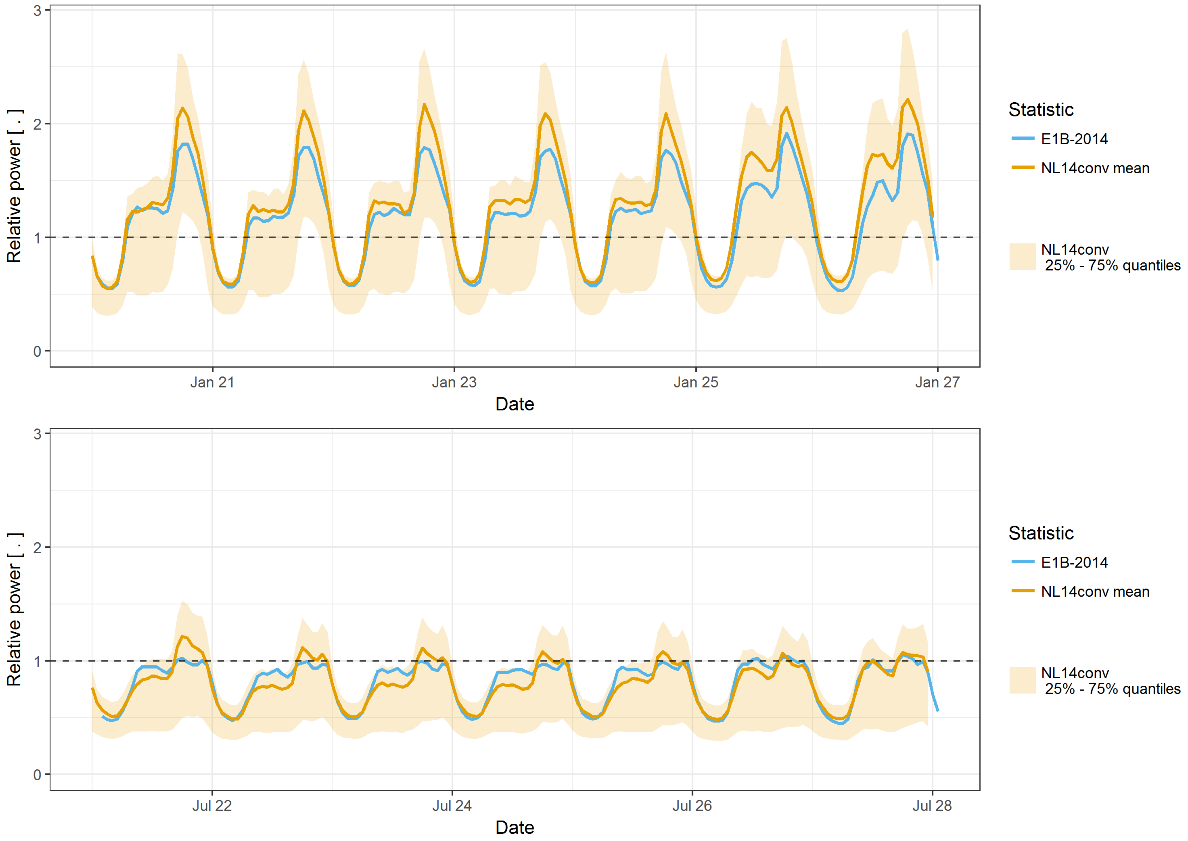
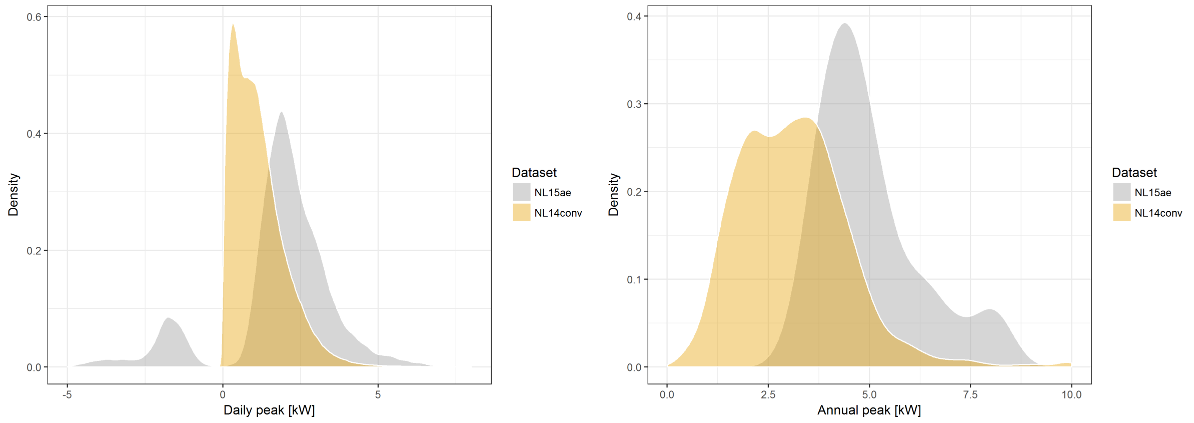


Figure 3.7: Time series of normalised hourly consumption for two weeks in 2014. In the winter week, the predicted consumption is below the mean, whilst in summer it lies slightly above it.



(a) Distribution of daily peak loads per home.

(b) Distribution of annual peak load per home.

Figure 3.8: Distribution of peak loads. The NL15ae-homes have significantly higher peaks than the NL14conv-buildings. NL15ae-homes sometimes have negative peaks, but on an annual basis the maximum consumption is always positive.

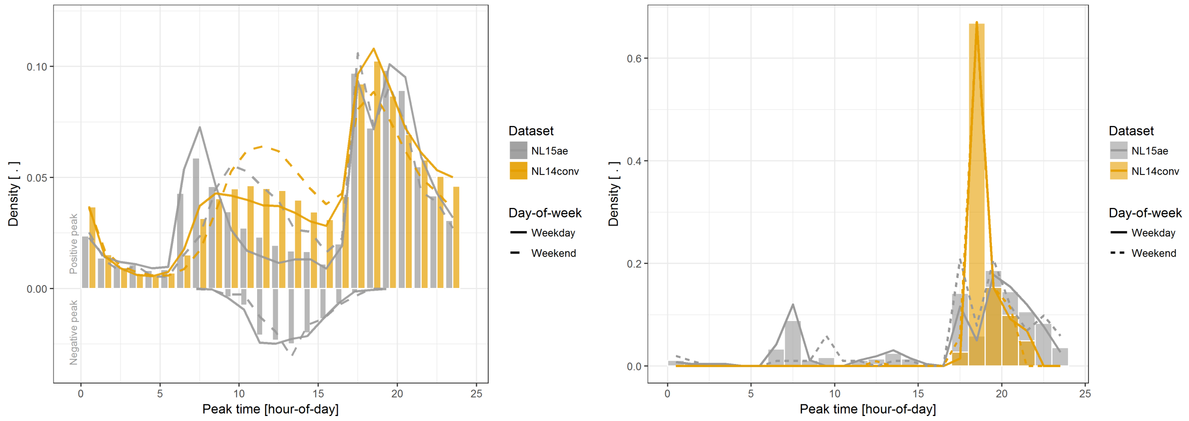
annual peaks are significantly higher for the all-electric homes (μ = 4.99 kW) than the large NL14conv-dataset (μ = 3.12 kW). The peak load distribution is in both cases similar for week and weekends. More statistics can be found in Table 3.5.

Table 3.5: Summary statistics of daily and annual peaks (in kW).

Dataset	Subset	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
NL14conv	Daily peak	0.001	0.486	0.985	1.156	1.591	9.996
	Annual peak	0.015	2.101	3.036	3.120	3.929	9.996
NL15ae	Daily peak	−5.162	1.396	2.009	1.819	2.768	8.271
	Annual peak	3.210	4.054	4.630	4.987	5.484	8.271

Peak timing

Next to the magnitude of the peak, it is valuable to know how these peaks are distributed over time. In Figure 3.9 the empirical distributions over individual and aggregated peak timing are plotted. From Figure 3.9a can be concluded that NL15ae-homes have their daily peaks earlier in the morning (between 6am and 9am) than NL14conv buildings, while the evening peaks are between 5pm and 9pm for both sets. Furthermore, in the weekends morning peaks are ± 2 hours later for both sets. In Figure 3.9b the distribution of peak times from the aggregated loads is shown.



(a) Peak-timing distribution of individual buildings. 'Negative' density indicate a negative load.

(b) Peak-timing distribution of aggregated loads.

Figure 3.9: Distribution of peak timing for individual demand (a), and aggregated demand (b) per dataset. The bars show the overall distribution, whereas the lines show the respective distributions per weekdays and weekends. NL15ae-homes have their peaks more concentrated and earlier in the morning than NL14conv-consumers. In the weekends both datasets have later morning peaks. The aggregate of NL14conv has its daily peak between 6pm and 8pm over 80% of the time.

3.3.4 Simultaneity

A measure that includes both the timing and magnitude of a peak is the *simultaneity*[2] factor [71]. This is an often used metric that quantifies to what extent a group of energy consuming agents – e.g. homes that are connected to the same HV/MV-sub station, or all appliances in a home – have their peak loads synchronised in time. A high simultaneity (close to 1) means that the total peak is close to the sum of the individual peaks, indicating that peaks are occurring simultaneously. A low simultaneity (close to 0) indicates that individual peaks are not happening at the same moment in time, thus that the load is smoothed out. Similar metrics are the *demand diversity* factor and the *coincidence* factor. No agreement about the exact definition of each of these metrics is found in the literature [72]. Therefore the definition of *simultaneity* [71] is used, which is equal to the used definition by Stedin:

[2] In Dutch *gelijktijdigheid*

Definition 1. *The* ***Simultaneity*** *γ of the energy consumption of a set of homes H is the ratio between the peak of the aggregated loads and the sum of the individual peaks of each home over a time interval T. With $\mathbf{P}^{(h)} = (P_1^{(h)}, ..., P_{8760}^{(h)}) \in \mathbb{R}^{8760}$ the annual time series of energy consumption of home h, this simultaneity is defined as:*

$$\gamma(H) = \frac{\max\limits_{t\in T} \overbrace{\sum_{h\in H} P_t^{(h)}}^{\textit{Aggregated load}}}{\underbrace{\sum_{h\in H} \max_{t\in T} P_t^{(h)}}_{\textit{Ind. peak load}}}, \tag{3.1}$$

with $P_t^{(h)}$ the energy consumption of home h at time t. The time interval T can be either a day or a year. For grid design, year-simultaneity is mostly used. With prosumers, congestion could also come from aggregated negative peaks resulting from on-site generation. In this scenario, the simultaneity is given by:

$$\gamma(H) = \begin{cases} \gamma^+(H), & \textit{if } \max S(H) > -\min S(H) \\ \gamma^-(H), & \textit{otherwise,} \end{cases}$$

with $S(H) = \sum_{h\in H} P_t^{(h)}$, $\gamma^+(H)$ is $\gamma(H)$ from Equation 3.1, and $\gamma^-(H)$ equal to $\gamma^+(H)$ with both max*-operators replaced with* min.

To evaluate this concept for each of the two datasets individually, a sampling procedure is constructed. This algorithm samples n homes from a dataset and calculates the simultaneity with Equation 3.1. Varying n and repeating 100 times for each n gives an estimation of the relationship between the sample size and the simultaneity. The pseudo-algorithm for this procedure is given in Algorithm 1. To calculate the simultaneity within a day, this procedure should be repeated for each day and with $T = (1, 2, ...24)$.

Algorithm 1 Sampling procedure for annual simultaneity calculation

```
$H$       := collection of homes to be evaluated
$\{P\}_t^h$   := collection of annual time series of energy consumption per home
$N$       := maximum sample size to evaluate
$n.rep$ := number of sample replications
procedure SIMULSAMPLE($H, N, n.rep$)                    ▷ Sampling simultaneity on annual basis
    $T := (1, 2, ..., 8760)$                                   ▷ All hours in the given year
    for n:=1 to N do                                       ▷ Iterate different sample size s
        for j:=1 to n.rep do                               ▷ Generate multiple replications
            Sample $n$ homes randomly from dataset H                          ▷ $H_n \subset H$
            $\alpha_{n,j} = \max_{t\in T} \sum_{h\in H_n} P_t^{(h)}$
            $\beta_{n,j} = \sum_{h\in H_n} \max_{t\in T} P_t^{(h)}$
            $\gamma_{n,j} = \frac{\alpha_{n,j}}{\beta_{n,j}}$                                         ▷ Simultaneity
        end for
        $\mu_n = \mathrm{mean}(\gamma_{n,\cdot})$
        $\sigma_n = \mathrm{sd}(\gamma_{n,\cdot})$
        ...                                  ▷ Other relevant statistics such as quantiles
    end for
    return $\mu, \sigma, ...$                                          ▷ Output statistics
end procedure
```

The results of this procedure is applied to each of the two datasets separately. Figure 3.10 shows the results for $N = 1, ..., 150$. In both cases the simultaneity is 1 for $N = 1$ (of course) and decreasing for larger N. The NL14conv-buildings show lower simultaneity than the NL15ae-homes. This supports the belief that the latter set of homes has more homogeneous consumption patterns, since they originate from similar home types on the same location. Furthermore, the simultaneity seems to be converging for both datasets for $n > 30$, although not enough data is available for NL15ae to verify this.

The remainder of this section analyses two more aspects of simultaneity: the time-dependency of day-simultaneity and the relationship between peak magnitude and simultaneity.

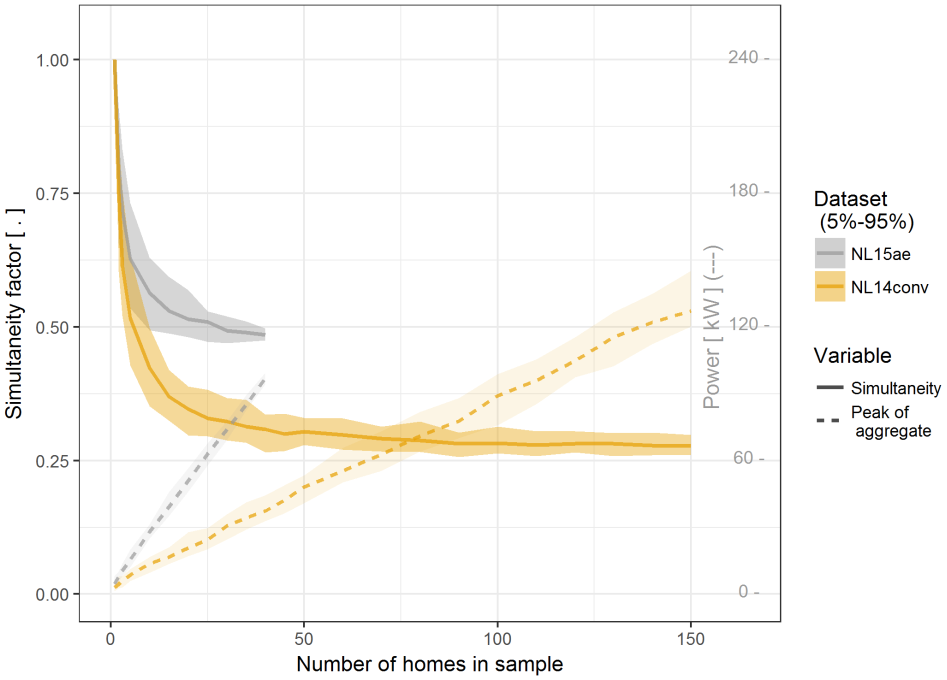


Figure 3.10: Simultaneity and aggregated peak as a function of sample size. The plot shows that the aggregated peaks are smoothing out when sample size s get bigger. For n > 30, the simultaneity converges to 0.3 *for the NL14conv-dataset. NL15ae-homes have higher simultaneity, which is consistent with the hypothesis that these homes are more homogeneous than those in the NL14conv-dataset.*

Time-dependency of day-simultaneity

First, the time-dependency of the day-simultaneity is investigated. To do this, the day-simultaneity is calculated for sample size s of $n = (20, 25, 30, 35, 40)$. These calculated samples are shown in Figure 3.11, together with a local regression line to show the trend over time. The NL14conv-dataset shows a slight periodicity with a little dip in the summer. This amplitude is larger for the NL15ae-homes, which varies from an average simultaneity of ±0.7 during winter to ±0.5 in summer. For the NL15ae-homes, the spread is larger in summer than during winter, with some positive extreme simultaneity values at days when there are large negative aggregated peaks from on-site electricity generation. The higher winter-simultaneity for the NL15ae-homes is likely to be a result of the increased electric heating demand.

Relation between median peak demand and simultaneity

For grid design it is also important to know whether groups of consumers with higher peak demand also have their peaks more simultaneous. Therefore, the second analysis looks at the relationship between the simultaneity and the median individual annual peak demand of all consumers in the sample. The median is used instead of the mean, since the mean is sensitive to extremes in small samples. For the NL15ae-dataset the positive and negative demand are analysed separately to look at the effect of solar panels. Figure 3.12 presents the results of such a simulation. This figure indicates that there is only a non-zero linear relationship between median peak demand and simultaneity for the NL15ae-homes at when the absolute aggregated peak results from solar panels. This suggests that roughly the same simultaneity factor holds for samples with different median peak demand, unless enough local PV capacity is present in the sample to dominate the aggregated peak. From that moment, larger PV capacity yield higher simultaneity. If all homes within a street have solar panels of sufficiently large capacity, a simultaneity of 1 can be expected.

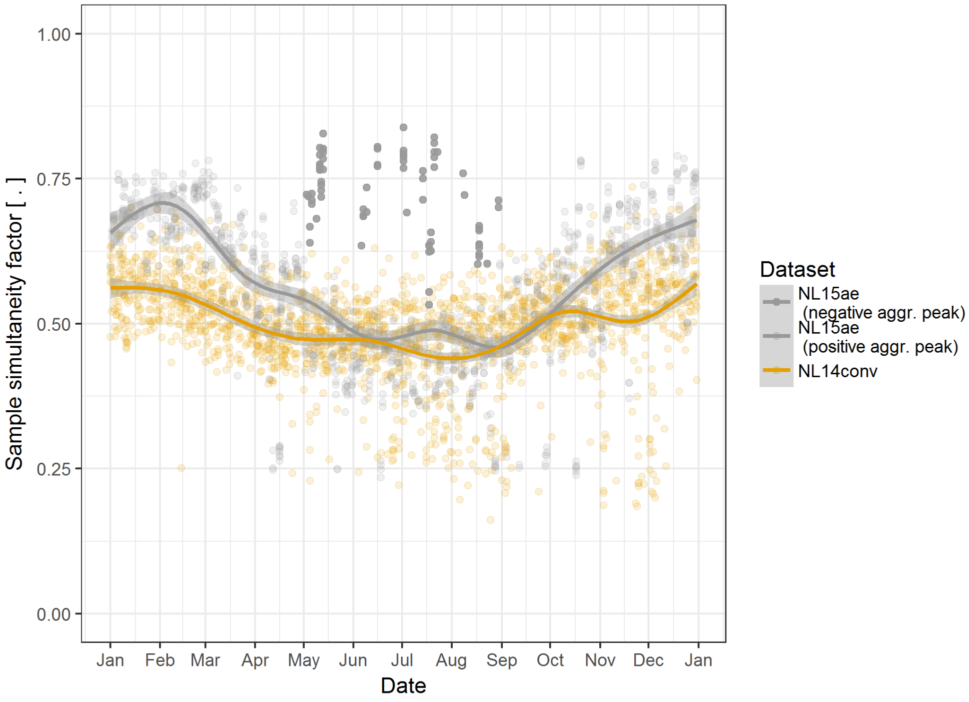


Figure 3.11: Trend of day-simultaneity over a year, generalised additive model (GAM) is used for the trend line. The difference between summer and winter is present in both datasets, but much larger for the NL15ae-buildings. In summer the highest simultaneity is observed at samples with a negative aggregated peak.

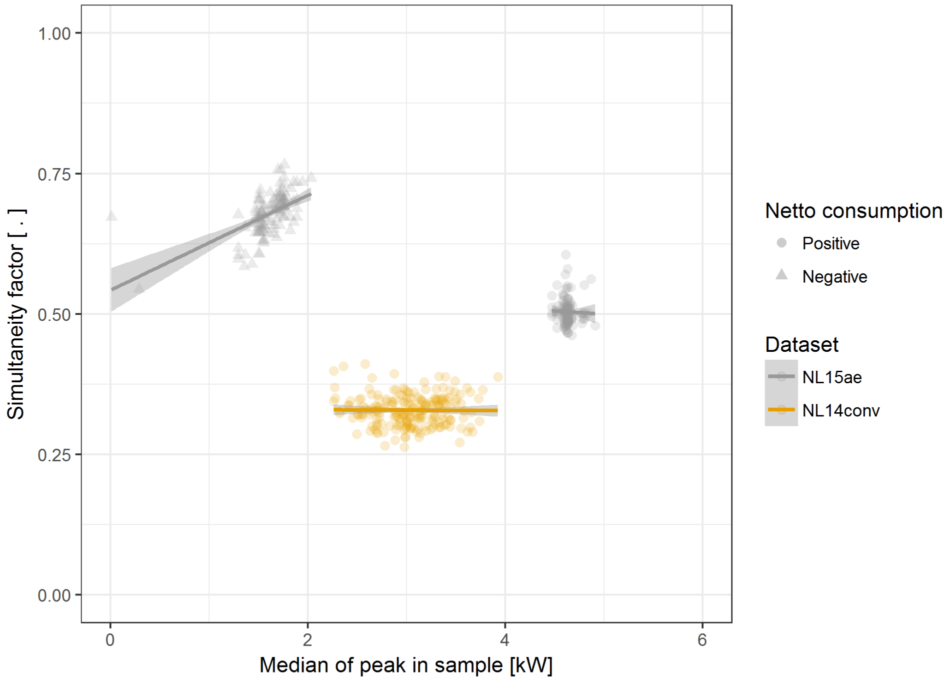


Figure 3.12: Relationship between a sample's simultaneity factor and median peak level. The trend lines are fitted linear models plus a 95% confidence interval. There seems to be a non-zero correlation only for the negative peaks in the NL15ae-dataset, which are the result of solar powered generation. Only sample sizes of $20 \leq n \leq 40$ are used here.

3.4 Conclusions

In this chapter two datasets were introduced, cleaned and explored. The NL14conv-dataset is a large dataset, containing various types of consumers. The NL15ae-dataset contains all-electric homes, of which some have solar panels and/or a battery. Cleaning this datasets results in a collection of 5,152 conventional consumers and 42 all-electric homes. These datasets are compared based on a variety of characteristics. This is summarised in Table 3.6.

Table 3.6: Energy consumption average statistics per dataset.

Dataset	n	Annual consumption $[MWh]$	Daily 1h-peak $[KW]$	Annual 1h-peak $[KW]$	Simultaneity[1] $[\cdot]$	Effective peak[1] $[KW]$
NL14conv	$4,735$	3.46	1.16	3.12	0.31	0.97
NL15ae	41	5.91	1.82	4.99	0.49	2.43

[1]Based on simulations of 40 homes

Compared to the NL14conv-dataset, the all-electric homes have 71% higher average consumption, 60% higher average annual peak demand and 58% higher average simultaneity. The higher peak demand and simultaneity lead to an increase in the effective peak per home of 151%. This means that based on these results, design parameters need to be adjusted for new grids, and existing grids should be actively monitored in order to prevent the aggregated peak from exceeding network capacity. A solution to avoid congestion could be applying demand response.

Another insight from this chapter is that the day-simultaneity is higher during winter than summer for both datasets. This effect is stronger for all-electric homes than conventional buildings, probably due to the extra heat demand in winter. Furthermore, simultaneity does not seem to be correlated with the median peak demand of a sample, unless the presence of PV within the sample is sufficiently large. In that case the simultaneity increases with the added PV capacity.

In this chapter the two datasets were analysed separately. However, the local energy transition occurs gradually, hence it is important to look at the effect of the penetration of technologies within an existing street. In the next chapters a framework will be constructed to simulate the effect of different mixes of consumer types on the aggregated peak demand and simultaneity. In order to do this, the NL14conv-dataset needs to be segmented into different consumer types. Chapter 4 will construct the basic element for this segmentation: the load shape.

Chapter 4

Clustering daily load shapes

In Chapter 3 was concluded that there is a big difference between energy consumption patterns in conventional buildings - like in the NL14conv-dataset - and all-electric homes such as the NL15ae-homes. Not only individual consumption levels and peak behaviours showed statistically significant differences, but also the extent to which these consumption patterns are synchronised over time. This information is of great importance for grid planners, as they estimate the future aggregate peak demand based on individual peak demand and simultaneity. As a result of increased penetration of NETs and changing consumer behaviour, both of these factors are likely to change. If both the individual peak and simultaneity increase by more than 50% this results in a doubling of aggregated peak demand. This could in turn motivate the DSO to intervene in the grid, implement demand response, or adapt design principles. However, the NL14conv-dataset consists of a large variety of consumer types, potentially including villa's, apartments, offices and shops, and thus is not a good representation of a single street. In order to draw a more valid conclusion on the effect of the local energy on sub-station-aggregate level, a way to cluster together consumers that have similar energy consumption patterns is needed. In this and the next chapters, a method to do this clustering is developed and used to simulate the effect of the local energy transition on the aggregated electricity demand.

In this chapter *adaptive K-means clustering* is deployed to create a ‘dictionary’ of normalised daily load shapes that summarises the variability in daily energy consumption. This method is based on the ordinary K-means clustering procedure, but has the advantage to let the number of clusters – i.e. the size of the dictionary – grow with the size of the data. Normalised rather than absolute consumption is considered to reduce the number of clusters, and to focus on timing of consumption rather than magnitude. The outputs of this chapter, the dictionary of daily load shapes and a mapping of each consumer in the dataset to these load shapes, will be used as input for lifestyle identification in Chapter 5, and the aggregated peak load assessment in Chapter 7.

Section 4.1 states the problem for load shape clustering and Section 4.2 introduces and explains the adaptive K-means algorithm. Afterwards, Section 4.3 presents the implementation and model choices. In the subsequent section the results are shown and discussed. The residuals resulting from this clustering are analysed in Section 4.5, both for individual and aggregated load shapes. The chapter concludes with some key insights.

4.1 Problem statement for load shape clustering

Given $\mathbf{P}^{(h)} = (P_1^{(h)}, ..., P_{8760}^{(h)})$, the annual energy consumption time series for home h, the daily time series for day d becomes

$$\mathbf{y}_d^{(h)} = (P_{t_d}^{(h)}, ..., P_{t_d+24}^{(h)}), \tag{4.1}$$

with index $t_d = 24(d-1)+1$, $d \in (1, 2, ..., 365)$ and $\mathbf{y}_d^{(h)} \in \mathbb{R}^{24}$. For the creation of the load shape dictionary the day and consumer are irrelevant, so the data indices are simplified using

$$\mathbf{y}_d^{(h)} \rightarrow \mathbf{y}_i,$$

with $i = 1, ..., 365N$, the number of consumption days in the dataset, assuming all days have data. Note that his is a bijection, therefore the time series and corresponding load shapes can always be re-assigned to the original day and consumer and vice versa.

In order to focus on the timing of consumption rather then the magnitude of consumption, and to simplify the problem, these load shapes are now normalised:

$$\mathbf{x}_i = \frac{\mathbf{y}_i}{|\mathbf{y}_i|} = \frac{\mathbf{y}_i}{\sum_{t=1}^{24} |\mathbf{y}_{i,t}|}. \tag{4.2}$$

The goal of load shape clustering now is to find a set of load shapes :

$$M = \{\boldsymbol{\mu}_j : \boldsymbol{\mu}_j \in \mathbb{R}^{24}, j = 1, ..., k\}$$

, and load shape assignments:

$$C = (c_1, ..., c_{365N}), \text{ with each } c_i \in \{1, ..., k\}$$

, such that every normalised daily load $\mathbf{x}_i$ is assigned to exactly one load shape $\boldsymbol{\mu}_{c_i}$. The problem of load shape clustering is now to find M and C such that:

1. Observations $\mathbf{x_u}$ and $\mathbf{x_v}$ assigned to the same load shape $\boldsymbol{\mu}_j$, $c_u = c_v = j$, are 'close' to each other,
2. Observations $\mathbf{x_u}$ and $\mathbf{x_v}$ assigned to different load shapes $\boldsymbol{\mu}_{c_u}$ and $\boldsymbol{\mu}_{c_v}$, $c_u \neq c_v$, are 'far' from each other.

4.2 Adaptive K-means clustering

An often used method for clustering of unstructured, high-dimensional data, is the iterative clustering algorithm *K-means* [73]. This algorithm is popular because of its simplicity and effectiveness. One of the disadvantages of K-means is that a pre-set number of clusters k needs to be chosen, and the final obtained clustering heavily depends on this k. Another disadvantage, one of great practical concern, is that K-means is inflexible in allowing new observations to alter the previous clustering result in a natural way. adaptive K-means (AK-means) is a method that works around this, by starting with a low number of initial clusters k_0 and iteratively increasing this amount until each cluster satisfies a compactness criterion. In this section, both algorithms are introduced. Their advantages and disadvantages are briefly explained and methods for model selection are discussed.

4.2.1 Ordinary K-means and its limitations

K-means is one of the most commonly used clustering methods, and was proposed 50 years ago [73]. Its aim is to partition n observations into k clusters, with each observations belonging to the cluster with the closest mean, while minimizing the sum of the squared errors within each cluster. However, searching through each possible k-partitioning is an NP-hard problem [74]. For example with $k = 2$ there are $2^{n-1} - 1$ possible partitions, which grows exponentially in the number of observations. Therefore, K-means uses an iterative algorithm to approximate the solution.

The K-means algorithm

Now a clustering of dataset X is considered, $X = \{x_1, ..., x_n\}, x_i \in \mathbb{R}^m$. The proposed algorithm follows the next steps, visualised in figure 4.1:

1. Randomly choose k initial centroids: $M_0 = \{\mu_1, ..., \mu_k\}$, with $\mu_i \in \mathbb{R}^m$. Set $p = 0$. (Figure 4.1a).

2. Assign each observation to its closest centroid, resulting in cluster assignments (Figure 4.1b):

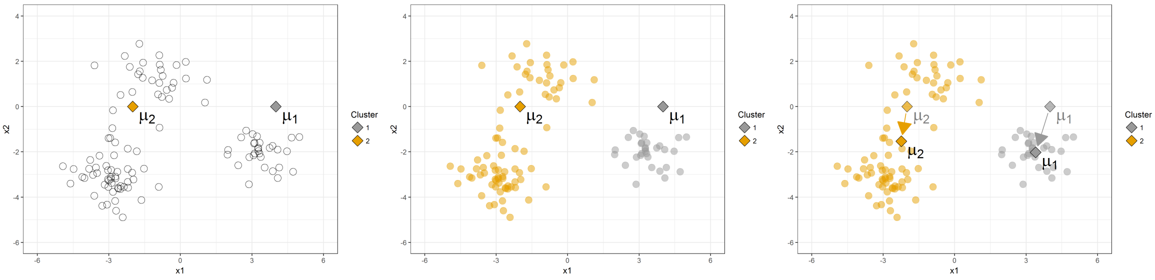
$$C_p = \left\{ c_1, ..., c_n : c_i = \underset{l \in \{1,...,k\}}{\operatorname{argmin}} \|x_i - \mu_l\|_2, \mu_l \in M_{p-1} \right\}, \tag{4.3}$$

   thus minimising the euclidean distance.

3. Calculate the mean vector of each cluster and use these as the new centroids (Figure 4.1c):

$$M_{p+1} = \left\{ \mu_1, ..., \mu_k : \mu_j = \frac{1}{|X_j|} \sum_{x \in X_j} x, \text{ with } X_j = \{x_i : c_i = j\} \right\}. \tag{4.4}$$

   Set $p = p + 1$.
4. Repeat steps 2 and 3 until convergence, i.e. no observations change their cluster membership in step 2.
5. Repeat steps 1-3 with different initialisations to avoid local optima. Choose the clustering that optimises a chosen cluster validity metric.



(a) *Initializing the clustering.* (b) *Assign x_i's to closest centroid.* (c) *Recalculate the cluster means.*

Figure 4.1: Example of K-means clustering. An initial clustering is shown in (a). The algorithm repeats step 2 (b) and 3 (c) until convergence.

This leads to the algorithm worked out in Algorithm 2, which is repeated for various initial choices of M_0 to avoid local optima.

Assumptions

The underlying assumptions of this algorithm are [75]:

1. There are k different clusters, i.e. each observation originates from one of k different m-dimensional distributions.
2. The minimisation of the euclidean distance in Equation 4.3 implies that clusters are spherical, meaning they have equal variance in each dimension.
3. Clusters are of approximately equal size.

Algorithm 2 Ordinary k-means clustering algorithm

```
Require: Data $X = \{x_1, x_2, ..., x_n\}, x_i \in \mathbb{R}^m$
  Choose $k$                                                        ▷ Number of clusters
  procedure kMeans($X, k$)
    Initialise $\mu_1, \mu_2, ..., \mu_k \in \mathbb{R}^m$ randomly     ▷ Cluster centroid initialisation
    repeat
      for $i := 1$to $n$ do
        $c_i := \underset{j}{\text{argmin}} ||x_i - \mu_j||$            ▷ Update cluster assignment
      end for
      for $j := 1$to $k$ do
        $\mu_j := \frac{1}{||\{i:c_i=j\}||} \sum_{i:c_i=j} x_i$         ▷ Update cluster means
      end for
    until convergence                        ▷ Converged if cluster assignments don't change
    return $C = \{c_i\}_{i=1}^{n}, M = \{\mu_j\}_{j=1}^{k}$
  end procedure
```

4. For a chosen k, the K-means algorithm tries to minimise the *sum-of-squared-errors* (SSE), given by:

$$SSE = \sum_{i=1}^{n} ||x_i - \mu_{c_i}||_2^2, \tag{4.5}$$

thus implying that minimising this quantity is what the modeller is interested in.

In load shape clustering, assumption 4 is valid, since a replacement of the normalised loads is required that approximates the original load well. Equation (4.5) is a measure of how well the original data is replaced by the assigned cluster means. Assumption 1 and 3 are not valid but dealt with when the adaptive K-means clustering is introduced. Assumption 2 is probably not valid although this is not checked within this research. The consequence of this assumption should be investigated in future work.

Model selection

An important part in applying this algorithm is the choice of the number of clusters k. Various methodologies have been suggested and evaluated in literature, like in [76]. The most common used methods are:

- Visual inspection, in which the modeller chooses the number of clusters based on what 'looks' reasonable. This approach is only feasible if data is 2 or, in some cases, 3-dimensional, and is subjective and potentially ambiguous. For example, for the clustering in Figure 4.1 both $k = 2$ and $k = 3$ can be reasonable suggestions.
- Heuristics, like the *elbow method* [77] and *kernel method* [78], or the rule-of-thumb to use the square-root of the number of observations: $k \approx \sqrt{n/2}$.
- Information-criterion approaches that use a measure like the Bayesian Information Criterion (BIC) [79] or the Akaike Information Criterion (AIC) [80] to choose the optimal k. For this approach, the likelihood needs to be computed, which requires the assumption of an underlying distribution. Usually the normal distribution is used here.
- Minimizing or maximizing cluster validity indices like the Silhouette-index [81], the Dunn-index [82] or the Davies-Bouldin Index [83] . These indices quantify the success of a clustering, based on the variance of observations within a cluster (which should be low) and the distance between different clusters (which should be high).

Advantages and limitations of K-means

K-means clustering is a frequently chosen method for clustering, especially when large, high-dimensional and sparse datasets are considered [84]. Benefits are its easy implementation and relative practical efficiency [85]. Although the theoretical worst-case running time is $2^{\Omega(n)}$ already in two dimensions [86], performance is much higher in practice: the number of iterations until the clustering stabilises is often linear in n [87]. However, there are some practical disadvantages of K-means:

- Each initialisation is converging to a local optimum, potentially resulting in counterintuïtive clusterings. This issue is partly addressed by using multiple initialisations of M_0 and choosing the solution that is most ”optimal” in some chosen sense. However, for high dimensional solutions, many initialisations need to be considered, and no guarantee on optimality can be given.
- The algorithm is sensitive to outliers and extremes, which have a big impact of the mean within a cluster.
- The choice of k is in most cases not trivial and does often not have a clear interpretation or motivation the way other tuning parameters might have.
- Selection methods that are presented in the previous section help out in choosing this k, but this has the disadvantage of making the algorithm less flexible: if new observations are to be included in the analysis, the whole procedure needs to be redone, without the guarantee to have a solution that is ”close” to the previous one.

Several suggested variations on K-means deal with some of these issues by learning the “optimal” k while running the algorithm. The G-means clustering method splits a cluster if a statistical test rejects the hypothesis that the observations within the respective cluster are normally distributed [88]. The tuning parameters in this model are the statistical significance level α and the variance σ. When this assumption of normality is not valid, or other cluster splitting criteria are preferred, the Adaptive K-means clustering algorithm offers a solution.

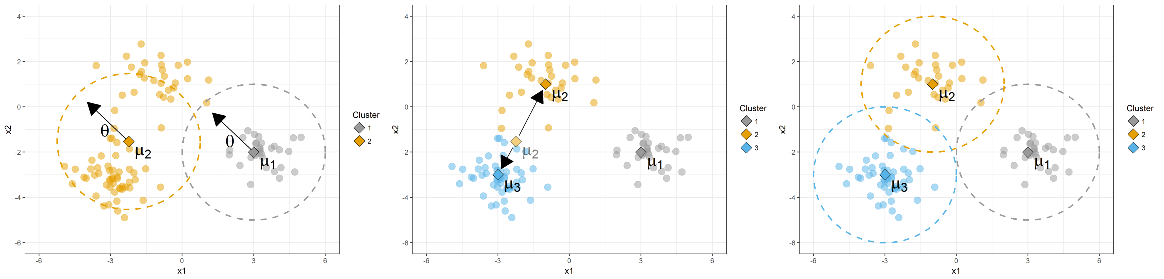
4.2.2 The Adaptive K-means algorithm

Adaptive K-means is a variation on ordinary K-means that replaces the choice of the number of clusters k by the choice of an upper bound θ on the compactness of the identified clusters. In many applications – e.g. load shape clustering – this is a choice that is easier to substantiate and interpret. Adaptive K-means was first introduced by [89], and applied to the clustering of load shapes in [39].

The main concept of this clustering method is to first start with an ordinary K-means clustering, with $k = k_0$ relatively small. Then for each identified cluster is evaluated if it does meet the compactness criterion:

$$d(X_j) \leq \theta, \tag{4.6}$$

with a compactness measure $d(X_j) = f(\mu_j, X_j)$, where X_j is the subset of observations assigned to cluster j: $X_j = \{x_i \in X : c_i = j\}$, and μ_j the cluster mean. Each cluster that does not meet this criterion is then split into two new clusters, by applying K-means with $k = 2$. The evaluation of compactness and the splitting of violating clusters is iterated until all clusters meet Equation 4.6. This concept is visualised in Figure 4.2.



(a) Result with k_0-means, $d_2(X_2) > \theta$ *(b) Run 2-means on X_2* *(c) Now all clusters comply*

Figure 4.2: Example of Adaptive K-means clustering. (a) shows the result from ordinary K-means clustering with $k_0 = 2$ and M_0 the result from Figure 4.1. It shows that $d_2(X_2) > \theta$, with $d_2(X_j) = \max_{x_i \in X_j} \|x_i - \mu_j\|_2^2$. Thus cluster 2 is split by applying 2-means clustering in (b). Now the compactness criterium is met for all three resulting clusters (c).

A pseudo-algorithm is presented in Algorithm 3. The algorithm works as follows:

1. Choose initial number of clusters k_0, distance measure $d(\cdot)$ and threshold θ.
2. Run K-means clustering with $k = k_0$, resulting in a set of cluster means $M_0 = \{\mu_1, ...\mu_{k_0} : \mu_j \in \mathbb{R}^m\}$ and cluster assignments $C_0 = \{c_1, ..., c_N : c_i \in (1, 2, ..., k_0)\}$ according to Equations 4.3 and 4.4.
3. Now for $j = 1$ to k, check for every X_j if it violates the $d(\cdot)$-threshold:

$$d(X_j) \leq \theta \tag{4.7}$$

   which results in the set of violating clusters V:

$$V = \{j : d(X_j) > \theta\}. \tag{4.8}$$

4. Now every cluster $j \in V$ is split with a 2-means clustering, each resulting in a set of cluster means M_j^* and cluster assignments C_j^*. The updated clustering then becomes:

$$C_p = \left(C_{p-1} \backslash \bigcup_{X_j : j \in V} c_i \right) \cup \bigcup_{j \in V} C_j^*,$$

   with cluster means

$$M_p = \left(M_{p-1} \backslash \bigcup_{j \in V} \mu_j \right) \cup \bigcup_{j \in V} M_j^*,$$

   which leads to Update k_{l+1} and set $l = l + 1$.
5. Repeat 2-4 until no points violate equation 4.7 and $V = \emptyset$.

Algorithm 3 Adaptive k-means clustering algorithm.

```
Require: Data $X = \{x_1, x_2, ..., x_n\}, x_i \in \mathbb{R}^m$ and distance measure $d(\cdot)$
  Choose $k_0$                                        ▷ Initial number of clusters
  Choose $\theta$                                     ▷ Threshold
  procedure AKMEANS($X, k_0$)
      $k := k_0$
      $C, M \leftarrow$ kMeans($X, k$)                 ▷ Run ordinary k-means
      repeat
          $k' := k$
          for $j := 1$to$k'$ do
              $X_j = \{x_i : c_i = j\}$                 ▷ All observations assigned to cluster $j$
              if $d(X_j) > \theta$ then                 ▷ Check if compactness criteria is met
                  $C^*, M^* \leftarrow$ kMeans($X_j, 2$) ▷ Split violating cluster with 2-means
                  $M = M_{-j} \cup M^*$                 ▷ Replace centroid $\mu_j$ with two new centroids $\mu_1^*$ and $\mu_2^*$
                  $C = C_{-j} \cup C^*$                 ▷ Assign to new clusters
                  $k := k + 1$
              end if
          end for
      until convergence                               ▷ if $k = k'$
      return C,M
  end procedure
```

Note that there are three main practical benefits of this model:

- The number of clusters is learned while running the algorithm
- It allows to pose an upper bound on how far away observations within one cluster can be from each other. In other words, a limit on the error can be given as input to the clustering
- New data that are added to a previous fit of the model do not require to rerun the algorithm from scratch, but rather fit in naturally: they are assigned to the closest cluster mean, after which compactness is checked and violating clusters are split

4.2.3 Model selection

The two important model choices that need to be made are the compactness measure $d(\cdot)$ and the threshold θ. Obvious choices for the compactness measure are the maxima of the conventional norms:

$$
\begin{aligned}
d_1(X_j) &= \max_{x_i \in X_j} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|_1 = \max_{x_i \in X_j} \sum_t |\mathbf{x}_i - \boldsymbol{\mu}_j|_{(t)}, && (4.9)\\
d_2(X_j) &= \max_{x_i \in X_j} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|_2 = \max_{x_i \in X_j} \sqrt{\sum_t \left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)^2_{(t)}}, && (4.10)\\
d_\infty(X_j) &= \max_{x_i \in X_j} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|_\infty = \max_{x_i \in X_j} \max_t |\mathbf{x}_i - \boldsymbol{\mu}_j|_{(t)}. && (4.11)
\end{aligned}
$$

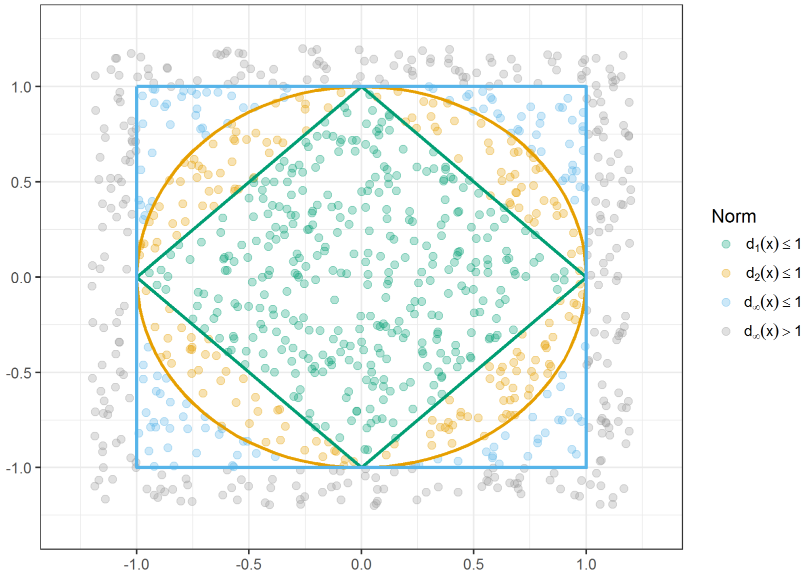


Figure 4.3: Examples of the three norms mentioned in Equations 4.9 - 4.11 in 2 dimensions.

In Figure 4.3, these three norms are visualised for 2-dimensional loads. Suggested in [39] is to use a measure that scales with the variability of a cluster mean:

$$
d(X_j) = \max_{x_i \in X_j} \frac{\|\mathbf{x}_i - \boldsymbol{\mu}_j\|_2^2}{\|\boldsymbol{\mu}_j\|_2^2}. \qquad (4.12)
$$

The actual choice of this measure can depend on the preference of the modeller, or can be motivated from the perspective of the application, as will be shown in section 4.3.2. The choice of θ depends on the used measure, and can be motivated in several ways.

4.3 Implementation of AK-means on energy consumption data

Now the implementation of this method on energy consumption data is considered. As mentioned in Section 4.1, the energy consumption time series $\mathbf{P}^{(h)}$ are chopped into vectors of normalised daily consumption $\mathbf{x}_i$, which are now clustered with the Adaptive K-means algorithm. This is being done to construct a ‘dictionary’ of load shapes that summarises consumption patterns while still capturing the variability.

4.3.1 Pre-processing

First the data is pre-processed. This is done in four steps:

1. **Split** annual consumption time series $\{\mathbf{P}^{(m)}\}_{m=1}^{M}$ into daily loads $\mathbf{y}_i \in \mathbb{R}^{24}$:

$$(P_1^{(m)}, ..., P_{8760}^{(m)}) \rightarrow \{\mathbf{y}_1^{(m)}, ..., \mathbf{y}_{365}^{(m)}\} \rightarrow \{\mathbf{y}_i, ..., \mathbf{y}_{i+364}\}. \tag{4.13}$$

2. **Select** only vectors that have no empty fields for load shape generation:

$$I = \{i : \mathbf{y}_{i,t} \neq \text{ NAN for all } t = 1, ..., 24\}. \tag{4.14}$$

3. **Normalise** like in Equation 4.2:

$$X = \left\{ \mathbf{x}_i = \frac{\mathbf{y}_i}{\sum_{t=1}^{24} |\mathbf{y}_{i,t}|} : i \in I \right\}. \tag{4.15}$$

4. **Sample** n normalised load vectors from X to determine the load shape dictionary with:

$$X \supseteq \tilde{X} = \{\mathbf{x}_i \in X : i \in \tilde{I}\}. \tag{4.16}$$

with $\tilde{I}$ sampled from I.

Step 4, the sampling of data, is advised when computing time or memory is a limiting factor, or when a trade-off needs to be made between the size of the dictionary and the accuracy of the clustering.

4.3.2 Model choices

The two important model choices that need to be made are:

1. the compactness measure $d(\cdot)$;
2. the threshold θ used to split clusters X_i for which $d(X_i) > \theta$.

Furthermore, since the datasets used are too big to apply the AK-means algorithm efficiently, a sample size n need to be chosen.

Compactness measure $d(\cdot)$

As mentioned in Section 4.2.3, various measures are proposed in literature. The choice of this measure can be different for different applications. An upper bound on the total absolute error resulting from using the cluster mean can be given by using the 1-norm, $d_1(\cdot)$, which is preferred in case total energy consumption over time is predicted with these load shapes. If the modeller wants to optimally capture the overall variability, the Euclidean norm $d_2(\cdot)$, which bounds the root mean squared error (RMSE), would be the preferred option. Since the focus of this thesis is on the peak consumption, having an upper bound on the error at peak time is desirable, leading to the choice of the infinity norm $d_\infty(\cdot)$.

A derivation of the upper bound on the aggregated error at peak time is shown in equations 4.18 - 4.22. For some sample of homes I, the aggregated load at time t is given by $(\sum_{i\in I} \mathbf{y}_i)_t$. This is maximised at peak time $t^* = \text{argmax}_t(\sum_{i\in I} \mathbf{y}_i)_t$. If now the inferred load shapes M are used instead of the raw data, the approximation of the daily consumption becomes $\mathbf{y}_i = \tilde{E}_i \mathbf{x}_i \approx \tilde{E}_i \boldsymbol{\mu}_{c_i}$, with

$$\tilde{E}_i = ||\mathbf{y}_i||_1, \tag{4.17}$$

the total absolute consumption on this day.

Now the aggregated error at peak time t^* is bounded by:

$$
\begin{aligned}
\left|\left(\sum_i \mathbf{y}_i\right)_t - \left(\sum_i \tilde{E}_i \boldsymbol{\mu}_i\right)_t\right|_{(t=t^*)} &\leq \max_t \left|\left(\sum_{i\in I} \mathbf{y}_i\right)_t - \left(\sum_i \tilde{E}_i \boldsymbol{\mu}_i\right)_t\right| && (4.18)\\
&= \left\|\sum_i \mathbf{y}_i - \sum_i \tilde{E}_i \boldsymbol{\mu}_{c_i}\right\|_\infty && (4.19)\\
&= \left\|\sum_i \tilde{E}_i (\mathbf{x}_i - \boldsymbol{\mu}_{c_i})\right\|_\infty && (4.20)\\
&\leq \sum_i \tilde{E}_i \|(\mathbf{x}_i - \boldsymbol{\mu}_{c_i})\|_\infty && (4.21)\\
&\leq \theta \sum_i \tilde{E}_i && (4.22)
\end{aligned}
$$

which means that the the error in the aggregated peak is bounded by a percentage of the total absolute consumption in the sample of that day.

Threshold θ

Now the threshold θ needs to be chosen. This parameter balances the trade-off between having a dictionary that is rich enough to approximate the actual loads within some desired limits, while remaining small enough to act as a summarisation that captures the similarities between different daily load profiles. To explore the relationship between θ, sample size n and the number of load shapes in the dictionary k, the algorithm is run for various sample sizes and values of θ. Results of this analysis are shown in Table 4.1 and Figure 4.4.

Table 4.1: Output number of clusters with AK-means and ∞-norm compactness measure.

	θ					
n_{obs}	0.01	0.02	0.05	0.10	0.15	0.20
1,000	972	835	379	120	60	39
10,000	9,511	7,131	2,328	622	237	133
100,000			14,946	3,332	1,245	699

These results suggest that the size of the load shape dictionary increases exponentially when θ decreases, and more-or-less linear in the sample size. This indicates that decreasing the upper bound of the error in Equation 4.22, or increasing the number of samples, both result in a much larger dictionary.

In further analysis, the dictionary is constructed based on a sample of 10,000 daily normalised loads. An upper bound on the maximum error per load shape assignment of $\theta = 0.1$ is used.

Remark. An important note here is that the choice to uniformly sample 10,000 daily loads was made relatively in the beginning of this research. The motivation for this was to reduce computation time – the distance matrix between points and cluster means takes a long time to compute – and the assumption that a trade-off between the accuracy of the load shapes and the variance in the LDA-model – which will be introduced in Chapter 5 – needed to be made. However, no time to further investigate this assumption was left. This is recommended for future research. Chapter 7 shows that a validation of the model fails for the NL15ae-homes without PV. This is assumed to be a consequence of the fact that only 15 homes ($\approx$ 0.3% of all data, thus ±30 daily loads in the sample) in the total dataset belong to this group. This is likely to have resulted in an under-representation of this group in the dictionary. Section 8.3 presents some recommendations to deal with this in the future.

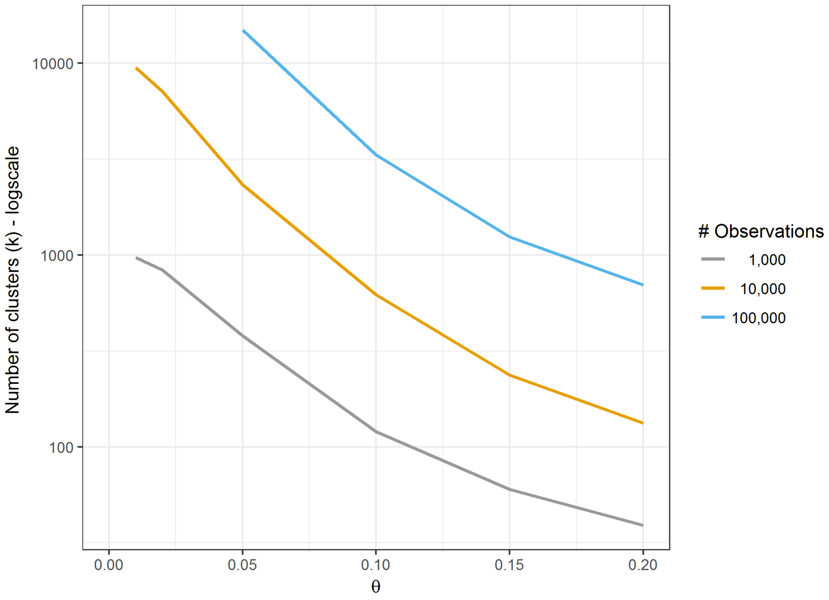


Figure 4.4: Number of identified clusters k as a result of performing adaptive K-means clustering on various sample sizes with different values for θ. The number of clusters seems to increase exponentially for smaller values for θ and linear in the sample size.

4.4 Results

Now 10,000 normalised loads are sampled randomly from the dataset and fed into the adaptive K-means algorithm with $\theta = 0.1$ and $d(\cdot) = d_\infty(\cdot)$. This results in a dictionary of 633 load shapes, the top and bottom 20 of which can be found in Appendix B. After this, all 1,805,363 normalised load shapes in the original dataset are assigned to the nearest element from the dictionary. The three load shapes with the highest cumulative consumption (= number of occurrences times the average consumption of actual loads in the cluster) are plotted in Figure 4.5. The flat load shape (#10) is the most frequent load shape with more than 100,000 daily loads assigned to this cluster. In Table 4.2, the statistics of the 5 most and least occurring load shapes are shown.

Table 4.2: Statistics of the 5 most and 5 least occurring load shapes.

order	cluster	n	μ (kWh)	σ (kWh)	rel. n (%)	rel. cum. consumption (%)
1	10	$102,031$	5.5	11.0	5.7	3.5
2	7	$28,992$	8.6	11.8	1.6	1.5
3	8	$18,996$	10.7	11.0	1.1	1.3
4	43	$19,136$	9.1	8.0	1.1	1.1
5	47	$19,602$	8.6	9.1	1.1	1.0
...	...	...	...	...	...	...
659	378	31	1.7	2.9	0.002	0.000
660	175	30	-4.3	2.7	0.002	-0.001
661	205	25	6.8	3.1	0.001	0.001
662	423	16	2.6	3.7	0.001	0.000
663	483	10	2.8	2.0	0.001	0.000
Total	all	$1,805,363$	9,0	6.0	100	100

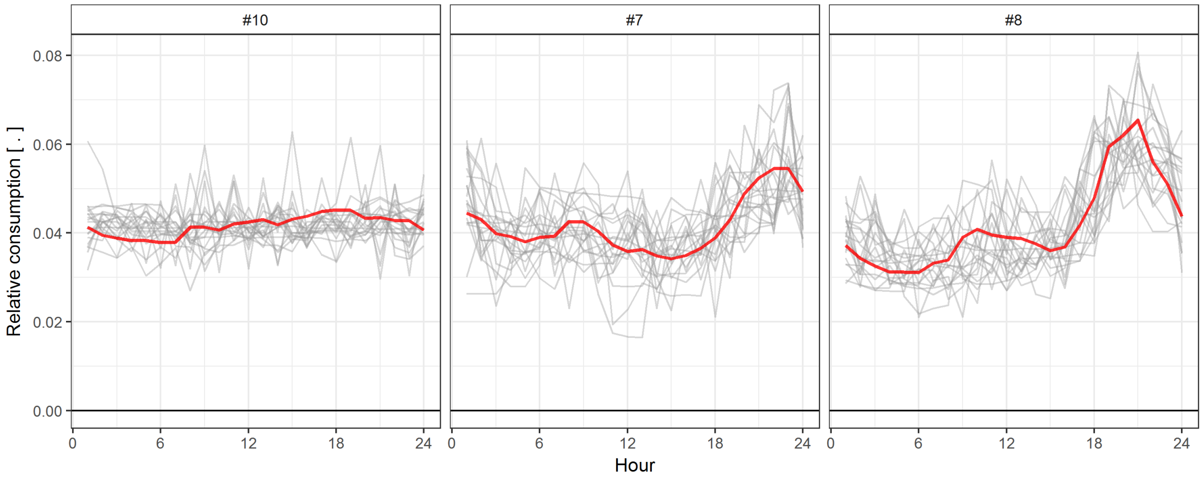
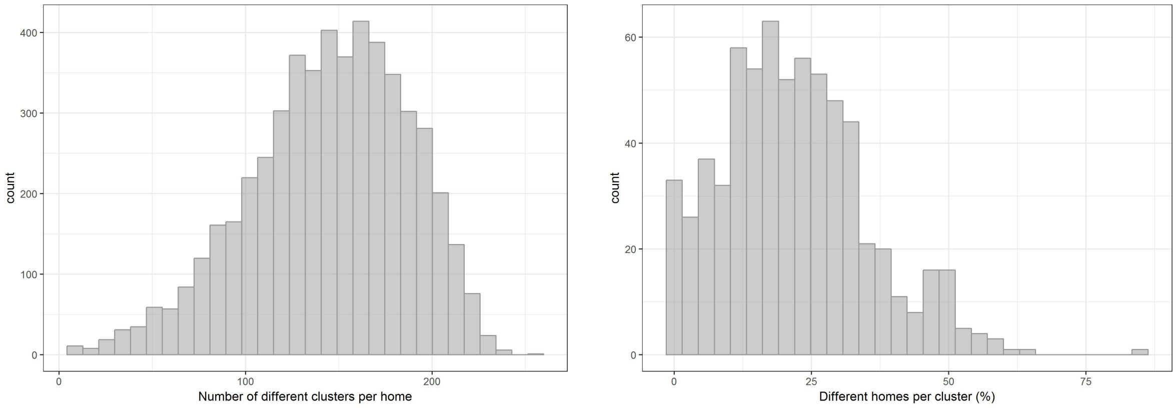


Figure 4.5: Three frequently occurring load shapes, resulting from Adaptive K-means. Each of the grey lines is one of the observations assigned to this load shape. The red line is the mean of the cluster.

From Figure 4.6a and Table 4.3, it follows that 50% of the consumers have between 117 and 176 different load shapes per year. In other words: for every home, every occurring load shape is occurring on average 2-3 times per year. Furthermore, load shapes are typically observed in 10-30% of consumers, with only load shape #10, the flat ‘absent/holiday’ load shape in Figure 4.5, being observed at least once in 85% of annual time series.



(a) Number of different clusters per home

(b) % of all homes that are assigned to a cluster

Figure 4.6: Distributions of (a) number of different homes per loadshape, and (b) % of total homes that observe a load shape. The load shape that is observed in 85% of the annual time series is load shape 10, the flat load shape from Figure 4.5.

Table 4.3: Quantiles of number of different load shapes per home

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Load shapes per home (#)	5.0	117.0	148.0	144.8	176.0	252.0
Homes per load shape (%)	0.08	12.2	20.2	21.8	29.4	84.9

4.5 Residual analysis

Now the residuals of the clustering are considered. The residual is defined as the error vector of an observation $\mathbf{x}_i$ and the cluster $\boldsymbol{\mu}_{c_i}$to which this observation is assigned:

$$\boldsymbol{\epsilon}_i = \mathbf{x}_i - \boldsymbol{\mu}_{c_i}. \tag{4.23}$$

4.5.1 Distribution of residuals

The first thing to note is that, since each cluster mean is by definition the mean of all loads assigned to it, the approximation of a normalised load by its assigned load shape should be unbiased, i.e.:

$$\mathbb{E}\left(\boldsymbol{\epsilon}_i\right) = \mathbf{0}. \tag{4.24}$$

Since the load shape clustering is based on a sample of the observations, it needs to be evaluated if this still holds in practice. The average error per hour of the day is given in Table 4.4 and plotted in Figure 4.7, as well as the standard deviation (± 2 times the standard deviation in the plot). The average error is close – compared to the standard deviation – to zero for every hour, motivating that this approximation is indeed unbiased.

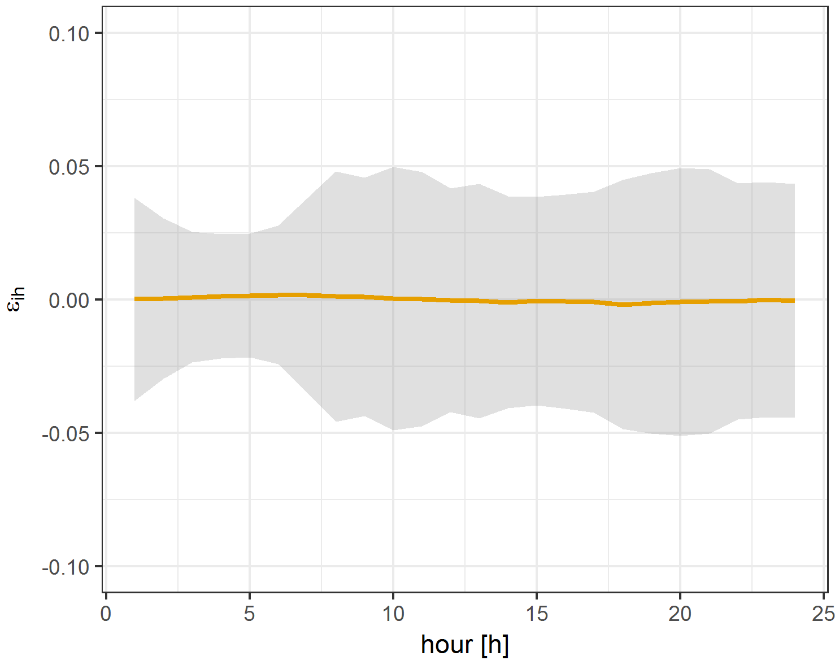


Figure 4.7: Plot of $\boldsymbol{\mu}_{\boldsymbol{\epsilon}} \pm 2\boldsymbol{\sigma}_{\boldsymbol{\epsilon}}$. The nearly flat line close to zero - compared to the confidence interval - suggests that this approximation is unbiased. The standard error σ_{ϵ} is largest during morning and evening peak times and smallest during the night.

When looked carefully at Figure 4.7, a subtle wave-like trend over time can be observed. This is confirmed in Table 4.4, which shows positive, though small, average errors for the night and morning hours, and negative average errors for most hours after 12AM. Compared to the variance in the error this trend is insignificant. However, these errors increase when series are aggregated, as is encountered in Section 4.5.3.

Table 4.4: Average and standard deviation of $\boldsymbol{\epsilon}_i$ (x10^3) per hour.

Hour	1	2	3	4	5	6	7	8	9	10	11	12
μ	0.13	0.34	0.92	1.15	1.50	1.65	1.53	1.08	1.10	0.18	0.19	−0.16
σ	18.97	15.17	12.34	11.68	11.65	12.81	17.96	23.74	22.79	24.35	23.36	21.14
	13	14	15	16	17	18	19	20	21	22	23	24
μ	−0.52	−1.00	−0.63	−0.60	−0.95	−1.81	−1.34	−0.92	−0.73	−0.67	0.03	−0.43
σ	22.32	19.96	19.66	20.23	20.83	23.24	24.48	25.23	24.85	22.23	22.25	21.85

4.5.2 Error magnitude

An upper bound $\theta = 0.1$ and compactness measure $\mathrm{d}_\infty(\mu_j) = \max_{(i:c_i=j)} ||\epsilon||_\infty$ were chosen in Section 4.3.2. Since the clustering is performed on a sample of merely 0.6% of all the data, it is now important to assess to what extent this criteria is met after all observations are assigned to their clusters. The $\infty-$norm is computed for each error vector and plotted in Figure 4.8. A summary of this data can be found in Table 4.5.

Table 4.5: Statistics of $\infty-$norm of residuals.

Statistic	Mean	St. Dev.	Min	Max
$\lVert\boldsymbol{\epsilon}\rVert_\infty$	0.048	0.039	0.002	0.696

What is immediately clear from these results is that $||\boldsymbol{\epsilon}_i||_\infty > 0.1$ for quite some of the observations i, thus violating the compactness criterion: 7% of the assigned actual observations have a maximum error that is higher than upper bound θ. Thus taking a sample of 0.6% of the data results in a clustering that represents 93% of the data within the specified limits. Although the sampling resulted in some violating cluster assignments this effect was not very large. However, larger sample sizes could and should be considered in future research.

4.5.3 Aggregation of residuals

Next to serving as input for lifestyle modelling, one of the potential purposes of a clustering of load shapes is getting better predictions of aggregated (peak) loads. Therefore, an analysis of the behaviour of aggregated errors is done.

Aggregated error over time

A sampling procedure is followed to explore the distribution of aggregated errors. For several sample sizes n, ranging from 1 to 200, a sample of n error vectors $\{\tilde{\boldsymbol{\epsilon}}_j\}$, with $\tilde{\boldsymbol{\epsilon}}_j$ the error-vector rescaled to the actual consumption:

$$\tilde{\boldsymbol{\epsilon}}_j = \boldsymbol{\epsilon}_j ||\mathbf{y}_j||_1, \tag{4.25}$$

is drawn from the computed set of error vectors and summed. This is repeated 1000 times for each n. For $n = 40$ and $n = 100$ the mean and $\pm 2\sigma$-intervals are plotted in Figure 4.9. This shows a fluctuating average error, which is positive in the morning and negative in the afternoon. The amplitude of this error increases for larger samples, which was expected. For $n = 40$, the average aggregated error peaks around 6AM at approximately +300 Watt, or 7.5 Watt per home. Compared with the effective peak of 0.97 kW of NL14conv-consumers (see Table 3.6) this error corresponds to less than 1%. Thus, the aggregated

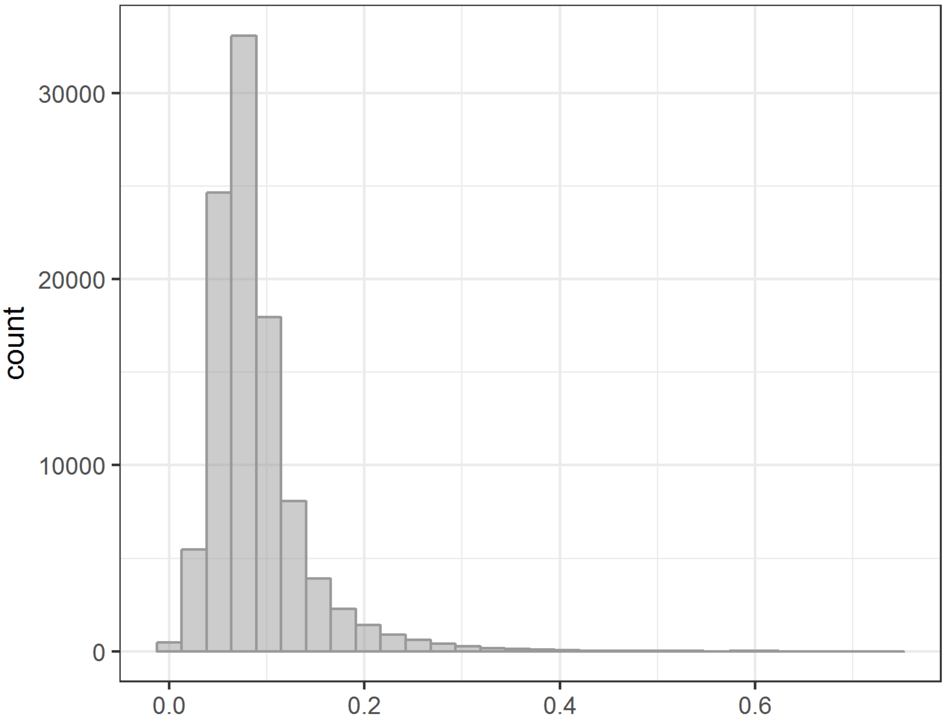


Figure 4.8: Histogram of $\|\epsilon\|_\infty$*, the maximum error between a normalised daily load and its assigned cluster mean. The* ∞*-norm exceeds the previously chosen upper bound* $\theta = 0.1$ *in 7% of the observations.*

load is approximated reasonably well. Important to note here is that the errors only result from an error in the *shape* of the load, since load *magnitude* is assumed to be known. In practice this daily consumption also needs to be predicted, resulting in an error that is probably of greater magnitude. However, this is not within the scope of this thesis.

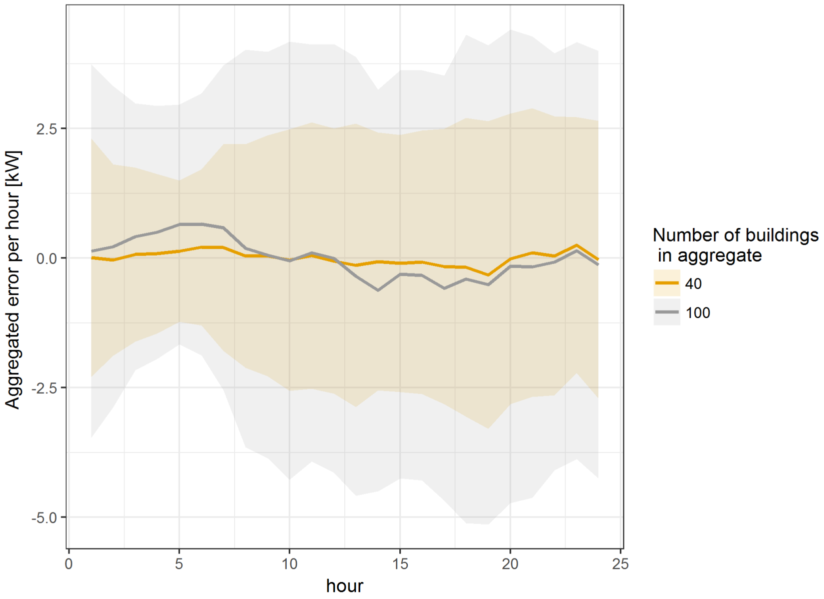
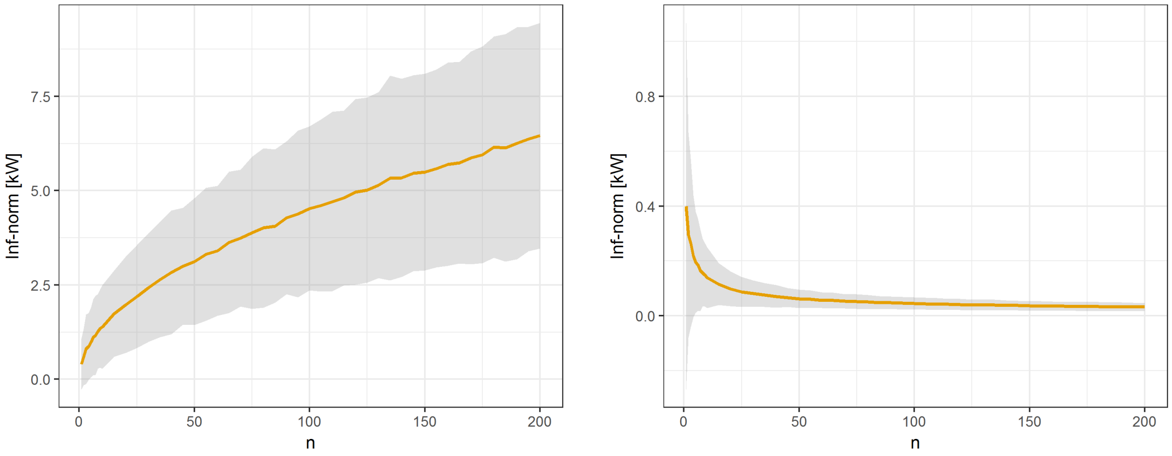


Figure 4.9: Distribution of error of aggregated load shapes per hour. The wave-like shape in the error is apparent, and the amplitude is increasing for larger sample sizes.

Maximum absolute aggregated error

To see how the distribution of the maximum absolute aggregated error of a sample depends on the sample size, $\|\sum_{j=1}^{n} \tilde{\epsilon}_j\|_\infty = \max_h |\sum_{j=1}^{n} \tilde{\epsilon}_j|$ is plotted in Figure 4.10 for different n. The maximum aggregated error (4.10a) increases to an average of $6.3kW$ for 200 homes. The average error per home (4.10b) is relatively high for small n but decreases towards zero for increasing n due to the *Law of Large Numbers*: the mean of a sample converges to the expected value.



(a) Maximum aggregated error

(b) Average error per home

Figure 4.10: $\|\sum_j \tilde{\epsilon}_j\|_\infty$ as a function of sample sizes. The aggregated error keeps increasing (a) but the average error per consumer stabilises (b).

Table 4.6: Mean and standard error of infinite norm of aggregated and mean residuals in kW.

$n =$	10	20	30	40	50	70	100	150	200
$\mu\left(\|\sum_j \tilde{\epsilon}_j\|_\infty\right)$	1.380	1.976	2.422	2.800	3.188	3.717	4.452	5.482	6.331
$\sigma\left(\|\sum_j \tilde{\epsilon}_j\|_\infty\right)$	0.488	0.623	0.686	0.790	0.861	0.926	1.052	1.317	1.487
$\mu\left(\|\frac{1}{n}\sum_j \tilde{\epsilon}_j\|_\infty\right)$	0.138	0.099	0.081	0.070	0.064	0.053	0.045	0.037	0.032
$\sigma\left(\|\frac{1}{n}\sum_j \tilde{\epsilon}_j\|_\infty\right)$	0.049	0.031	0.023	0.020	0.017	0.013	0.011	0.009	0.007

4.6 Conclusions

In this chapter, the clustering procedure Adaptive K-means is used as a method to efficiently and effectively summarise the variability in (normalised) daily loads. This choice was motivated by the fact that adaptive K-means combines the attractive properties of K-means - effective for high-dimensional datasets and easy to implement - with the possibility to pose an upper bound on the error resulting from approximating observation $\mathbf{x}_i$ with its nearest cluster centre. This also implies that the number of clusters k does not need to be chosen prior to the clustering, as is the case in K-means clustering. An important practical benefit from this is that the algorithm does not need to be rerun from scratch when new data is observed, but can do the clustering sequentially, starting with an old result and incorporating new observations without changing the old clustering too much.

Applying adaptive K-means clustering, with an upper bound on the maximum error of $\theta = 0.1$, on a sub-sample of 10,000 daily loads, leads to a set of 633 load shape clusters. The largest cluster represents

5,7% of the data, while several small clusters have very low presence. After this, all 1,8 million loads are assigned to their closest cluster mean. Although clustering is performed on 0.6% of all observations, the resulted clusters still meet the specified compactness criterion for 93% of all the observations. A subtle but insignificant overestimation of the morning peak can be observed. When the errors are rescaled to actual consumption and aggregated, the order of magnitude of this overestimation is around 1% of the daily peak, indicating that these load shapes approximate the actual load very well. An important remark is that the sub-sampling resulted in an under-representation of the all-electric homes in the training set. This can influence the results, which is likely to be the case in Chapter 7. In future research, a large sample size should be taken or the samples should be stratified.

In the next chapter a method is introduced that identifies lifestyle-like patterns based on the inferred load shapes.

Chapter 5

Latent lifestyles in energy consumption data

In the previous chapter, Adaptive K-means was deployed to derive a list of daily load shapes from a collection of energy consumption time series. This dictionary of 633 load shapes summarizes the 1.8 million consumption days in the dataset. A year of energy consumption data for one building can now be represented by a sequence of 365 elements from this dictionary, instead of 8760 hourly observations. The purpose of this chapter is now to find patterns in these sequences of load shapes that can be related to lifestyles, in order to enable an interpretable clustering of consumers.

The methodology that is used to identify these lifestyle patterns is latent Dirichlet allocation (LDA), a model widely used in text mining and other fields that look for structure in discrete data. In this generative probabilistic model, energy consumption is assumed to be originating from a mix of unobserved lifestyles per consumer, where each lifestyle is characterized by a distribution over the load shapes from the previous chapter. An illustration of this concept is showed in in Figure 5.5.

This chapter starts with a short motivation of the proposed model. In Section 5.2 an introduction to latent Dirichlet allocation (LDA) follows. For convenience, terminology and notation of text modelling is used initially. LDA is introduced introduced step-by-step, starting at a basic generative model for text analysis, increasing in complexity in several stages. This section concludes with a full statistical model for LDA. In Section 5.3 inference and parameter estimation in LDA with variational expectation maximisation (VEM) is treated. Section 5.4 explains the analogy between text and energy consumption modelling, after which the underlying assumptions of LDA is explained in the context of energy consumption. The implementation and model selection are discussed briefly in the next section. In Section 5.6, the results of this analysis are shown, including the relevant estimated posterior distributions and some characteristics of the inferred lifestyles. The key conclusions of this chapter are stated in the last section.

5.1 Motivation

Earlier attempts to classify energy consumption focused on identifying different appliances rather than *when* and *why* people are consuming energy [60]. Such segregation of technologies based on smart meter data requires more granular data than is now available in most situations [50]. Other approaches, such as the one in [59], assume that consumers belong to a single consumer cluster. In reality, people usually do not have one single lifestyle: colleagues that also live in the same neighbourhood could have almost identical consumption patterns during weekdays, but the first might sleep in most Saturday mornings while the other brings his kids to football at 9AM. One's individual 'lifestyle' can thus better be represented as a mixture of different lifestyles that vary per day-of-the-week and over time. Modelling energy consumption as a mix of these lifestyles can potentially capture this concept, leading to more realistic and interpretable consumer clusters. Therefore, latent Dirichlet allocation (LDA) is proposed to infer lifestyles from energy consumption data and applied to the transformed smart meter data from the previous chapter. The benefits of this model are:

1. The lower dimensional representation of consumers improves the clustering of consumers.
2. The concept of a lifestyle enables a more interpretable characterisation of consumers than conventional consumption based clustering approaches.
3. The generative model that is proposed implies a framework for load shape based simulation of aggregated demand.

The last point is not directly used in this thesis, since actual data is available for simulation, but various model expansions for load simulation are suggested in Section 8.3.

5.2 Latent Dirichlet Allocation for text modelling

LDA originates from text analysis [90], but is applied in a variety of fields in which finding patterns in collections of discrete data is of interest, such as bioinformatics [91] and image classification [92]. For convenience, the terminology of text analysis will be used in this section to explain the model, and in the next section to explain how to make inference in LDA. Therefore, the terms *documents*, *words* and *topics* will be used rather than *annual energy consumption*, *load shapes* and *lifestyles*. In Section 5.4, the analogies between these terms will be further clarified.

5.2.1 Intuitive interpretation

Consider a library in which there are 10,000 scientific books and articles. The librarian wants to label each of these documents based on the multiple topics they treat, and put this in a digital database in order to help researchers find relevant literature. An article about evolutionary biology might treat the topics *genetics*, *evolution* and *data analysis*, while a book on computational neurology is about *neurology*, *data science* and *stochastic simulation*. The librarian has no time to read through all of these books and extract the themes by himself. However, he has access to digital copies of each of the documents. Latent Dirichlet allocation helps him to identify these unobserved topics, i.e. clusters of words that often occur together throughout documents, and labels each document with one or more of these topics.

The generative probabilistic model LDA assumes that this library is generated in approximately the following way: 1) each book is characterised by a distribution over some topics, 2) each topic is characterised by a distribution over all non-trivial words[1] in the dictionary, 3) each non-trivial word in the book is sequentially generated by 3a) first drawing a topic from the topic-per-book distribution, and then, 3b) drawing a word from the words-per-book distribution. This is repeated until each document in the library is finished. This generative process is visualized for one document in Figure 5.1. Several statistical methods can be used to make inference about these underlying distributions based on the observed data, i.e. the books and articles in the library.

[1] trivial words are words that contain no actual information, like articles and conjunctions

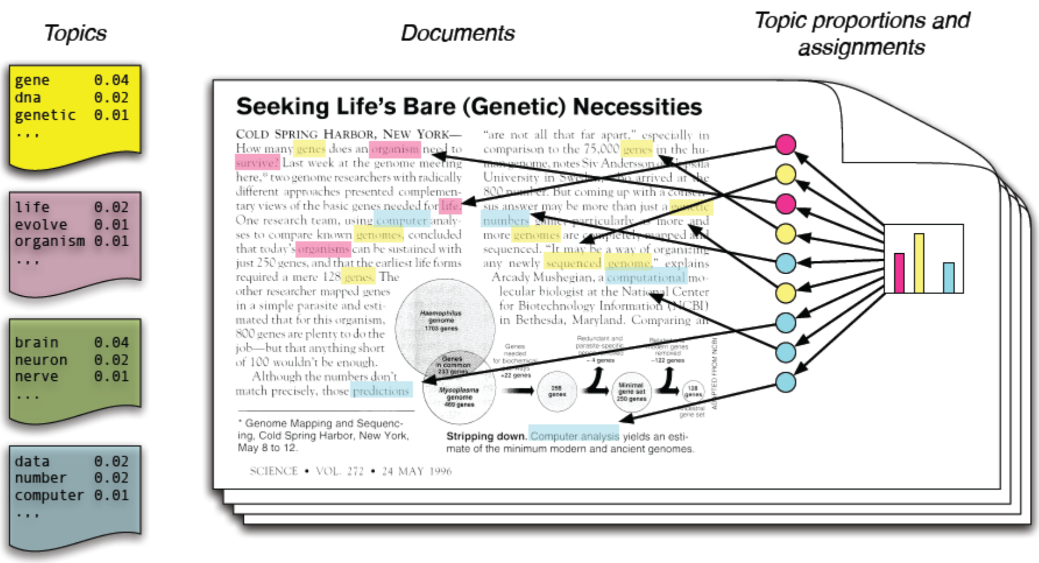


Figure 5.1: The intuition behind LDA. A document is assumed to have a distribution over topics (histogram on the right), and each topic has a distribution over words (colored blocks on the left). The words in the document are generated by first drawing a topic from the topic-per-document distribution, and then drawing a word from the word-per-topic distribution. The topics and topic assignments in this figure are illustrative —- they are not fitted on real data [93].

5.2.2 Notation and terminology

Because the most common application of LDA is text analysis, most literature speaks of *words* as the observed discrete data, *documents* as an observed collection of words, and *topics* as the latent variable to make inference about.
More formally, the following terms are defined:

- The *dictionary* D is the collection of all words that can occur in a document. The size of this dictionary is V.
- A *word* w is the basic unit of discrete data in this framework and refers to the index of an element in the *dictionary* D.
- A *document* $\mathbf{w}_m$ is a series of N words denoted by $\mathbf{w}_m = (w_{m,1}, w_{m,2}, ..., w_{m,N})$. The number of words, N, can vary per document. However, since the final model will consists of a full year of energy consumption data, N is considered to be constant for all documents.
- A *corpus* C is a collection of M documents denoted by $C = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$. In the figurative example of the previous section, the corpus is the library.
- A *topic* τ is the unobserved stochastic variable in this model that needs to be inferred from the data. A more rigorous definition follows in the next sections.

5.2.3 Generative probabilistic models and latent topics

LDA is a generative probabilistic model, which means that it assumes the observed data to be generated by a process consisting of several connected stochastic variables with some hidden parameters. Thus every $w_{m,n}$, the n-th word of the m-th document, is a realisation of a stochastic variable $W_{m,n}$. Such models are represented by a hierarchical Bayesian framework of one or multiple connected (conditional) probability distribution functions. Often these models have one or multiple latent[2] stochastic variables that are included in the Bayesian framework. In LDA, this latent stochastic variable is τ, the *topic* that inference needs to be made about.

[2] latent = hidden or unobserved

Several of such generative probabilistic models are suggested in literature. LDA can be seen as a relatively complex member of this family. In the rest of this section, several of these models will be introduced briefly as a way of explaining how LDA is built up. The relevant distributions can be found in Appendix C.

Unigram model

The most simple generative model that is the unigram. In this model, words are drawn independently from a single V-dimensional categorical distribution that is identical for each document in the corpus. Mathematically, this means that for $1 \leq m \leq M, 1 \leq n \leq N$:

$$W_{m,n} \mid \boldsymbol{\eta} \overset{\text{iid}}{\sim} \text{Cat}(\boldsymbol{\eta}), \tag{5.1}$$

for every observed $w_{m,n} \in \mathbf{w}_m$, every $\mathbf{w}_m \in C$ and $\boldsymbol{\eta} \in \{\mathbf{x} \in \mathbb{R}^V_{\geq 0} : \sum \mathbf{x} = 1\}$, an unknown V-dimensional probability vector that is equal for every book in the corpus. The graphical representation of this generative model is given in Figure 5.2a. Inference about the parameter vector $\boldsymbol{\eta}$ can be made with maximum likelihood estimation (MLE) or a Bayesian procedure, using a Dirichlet distribution as conjugate prior to the Categorical distribution (see Appendix C).

Mixture of unigrams

The assumption that all documents have the same identical word-distribution is not very realistic. Hence the concept of a topic is introduced. In the *mixture of unigrams* [94] it is assumed that a document $\mathbf{w}_m$ is generated by first choosing out of K topics, $\tau_m \in \{1, ..., K\}$, and then generate the words in a document, $\mathbf{w}_m = (w_{m,1}, ..., w_{m,n})$, mutually independent conditional on this topic τ_m:

$$W_{m,n} \mid \tau_m, \{\boldsymbol{\eta}_k\}_{k=1}^K \overset{\text{iid}}{\sim} \text{Cat}(\boldsymbol{\eta}_{\tau_m}). \tag{5.2}$$

Thus every of the K different topics is characterized by a distribution over words in the dictionary, and every document belongs to one single topic. The number of topics τ, and the word-per-topic distributions, can be estimated with MLE or Bayesian approaches, assuming prior distributions on $\{\tau_m\}_{m=1}^M$ and $\{\boldsymbol{\eta}_k\}_{k=1}^K$.

Probabilistic Latent Semantic Indexing

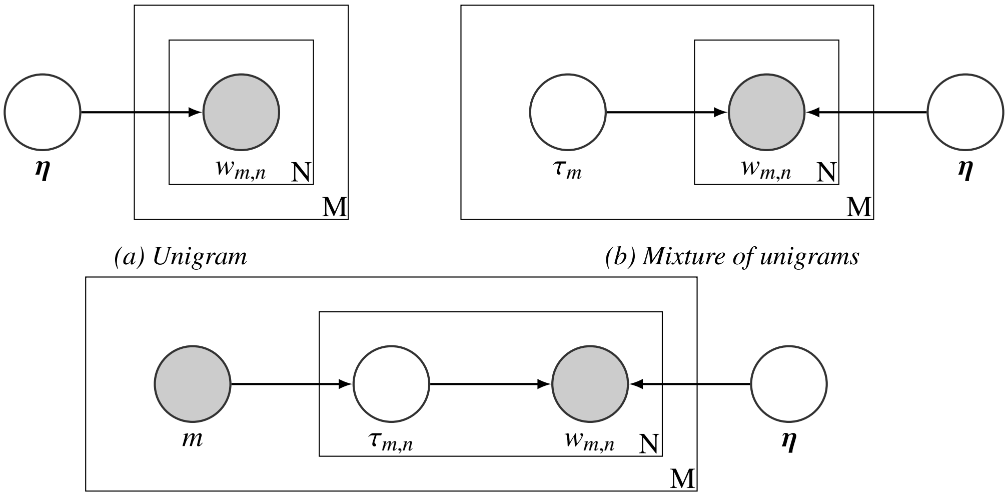
The next step is to allow documents to exhibit a mixture of different topics. The generative model probabilistic latent semantic indexing (pLSI) now assumes that each document has its own specific topic distribution. Each word $w_{m,n}$ is then generated by first drawing a topic $\tau_{m,n}$ from the topic distribution, after which $w_{m,n}$ is generated conditional on $\tau_{m,n}$. This is repeated N times until the document $\mathbf{w}_m$ is completed. This leads to the following hierarchical Bayesian model [95]:

$$W_{m,n} \mid \tau_{m,n}, \{\boldsymbol{\eta}_k\}_{k=1}^K \overset{\text{ind}}{\sim} \text{Cat}(\boldsymbol{\eta}_{\tau_{m,n}}), \tag{5.3}$$

$$\tau_{m,n} \mid m \overset{\text{iid}}{\sim} \text{Cat}(\boldsymbol{\theta}_m), \tag{5.4}$$

with $\text{p}(\tau_{m,n} = k \mid m) = \theta_{m,k}$, the k-th element of the m-th parameter vector. It is assumed that a document $\mathbf{w}_m$ and a word $w_{m,n}$ are conditionally independent given the unobserved topic $\tau_{m,n}$.

This pLSI model is illustrated in Figure 5.2c. Since $\boldsymbol{\theta}_m$ acts as the mixture weights of the topics for the $m-$th document, it seems to capture the multi-topic structure that is desired. However, it is important to note that m is an index into the training corpus C [90]. Therefore pLSI assumes that there is a finite number ($= M$) of possible topic-per-home distributions, exactly defined by the inferred distribution vectors $\{\boldsymbol{\theta}_m\}_{m=1}^M$ for the M documents in the corpus. This makes the model sensitive to overfitting, since the number of parameters which must be estimated grows linearly with the size of the corpus: the parameters for a K-topic pLSI-model, inferred on a corpus of M documents and a dictionary of size V, requires $kV + kM$ parameters to be estimated [96].



(a) Unigram

(b) Mixture of unigrams

(c) pLSI

Figure 5.2: Graphical model representation of the previous latent variable models. The grey circles show the observed variables, while the white circles are unobserved. Each box represents replication. Adapted from [90].

latent Dirichlet allocation (LDA)

Latent Dirichlet Allocation solves the above-mentioned problem of overfitting by treating the parameter $\boldsymbol{\theta}$ as a K-dimensional random variable, rather than one fixed parameter (*mixture of unigrams*) or a set of fixed parameters (*pLSI*), by expanding the Bayesian hierarchical model in Equations (5.3)- (5.4) with a prior distribution on $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}_m \mid \boldsymbol{\alpha} \overset{\text{iid}}{\sim} \text{Dir}(\boldsymbol{\alpha}). \tag{5.5}$$

The (K-1)-dimensional hyperparameter $\boldsymbol{\alpha}$ is assumed constant for all documents. In the next Section this model will be worked out further. Its structure is illustrated in Figure 5.4. In this way, only $K + KV$ paramaters need to be estimated [90], so the model size does not grow in the size of the corpus.

An illustration of how the four mentioned models are linked for a simple case with three words and three topics can be found in Figure 5.3.

5.2.4 Statistical model for LDA

Latent Dirichlet Allocation now assumes that every document $\mathbf{w}_m$ in the corpus C is generated as follows:

1. Draw the parameter $\boldsymbol{\theta}_m \in \mathbb{R}^K$ for the τ-per-document distribution from another Dirichlet distribution:
$$\boldsymbol{\theta}_m \mid \boldsymbol{\alpha} \overset{\text{iid}}{\sim} \text{Dir}(\boldsymbol{\alpha}).$$
2. Then for each of the $n = 1, ..., N$ words $w_{m,n} \in \mathbf{w}_m$:
   (a) draw a topic $\tau_{m,n}$ from a Categorical distribution with parameter $\boldsymbol{\theta}_m$,
   $$\tau_{m,n} \mid \boldsymbol{\theta}_m \overset{\text{iid}}{\sim} \text{Cat}(\boldsymbol{\theta}_m),$$
   followed by;
   (b) draw a word $w_{m,n}$ from a Categorical distribution with parameter $\boldsymbol{\eta}_{\tau_{m,n}} \in \mathbb{R}^V$:
   $$W_{m,n} \mid \tau_{m,n}, \boldsymbol{\eta} \overset{\text{iid}}{\sim} \text{Cat}(\boldsymbol{\eta}_{\tau_{m,n}}),$$

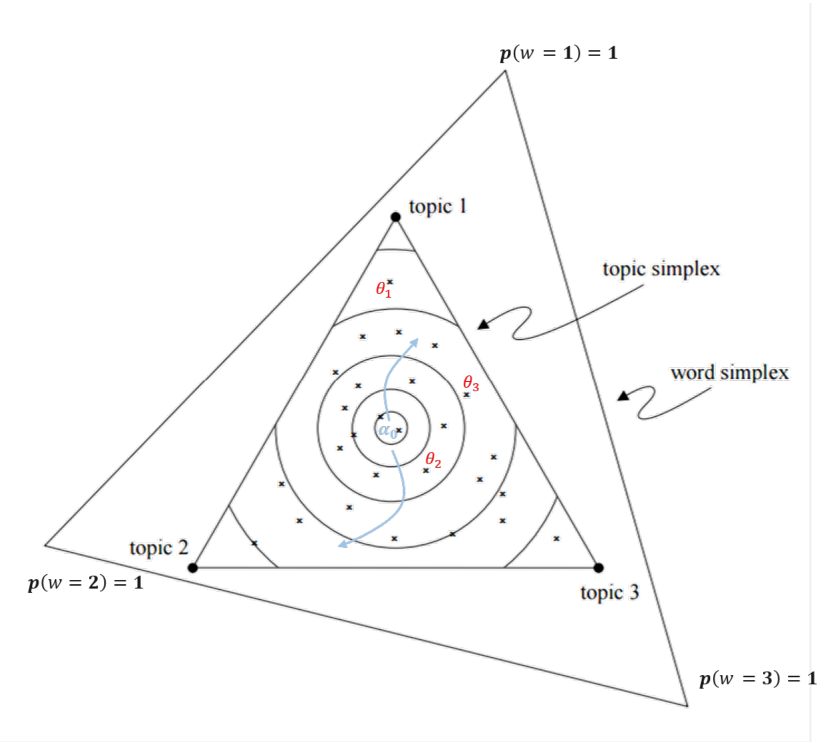


Figure 5.3: A topic simplex for three topics embedded in the w-simplex for three words. Each vertex of the outer triangle corresponds to a deterministic word distribution. The vertices of the inner triangle define the parameter vectors of the word-per-topic distributions. The x's are the topic-distribution parameter vectors θ per document (pLSI). In the LDA-model, these parameters are drawn from a Dirichlet distribution with parameter $\boldsymbol{\alpha}_0$, denoted by the contour lines over the topic simplex. Adapted from [90].

where $\boldsymbol{\alpha}$ is the hyperparameter for the Dirichlet prior and $\boldsymbol{\eta} = \{\boldsymbol{\eta}_k\}_{k=1}^{K}$ – changing notation for convenience – are the unknown parameters for the word-per-topic distributions. This scheme results in the following Bayesian hierarchical model, illustrated in Figure 5.4:

$$W_{m,n} \mid \tau_{m,n}, \boldsymbol{\eta} \overset{\text{ind}}{\sim} \text{Cat}(\boldsymbol{\eta}_{\tau_{m,n}}), \tag{5.6}$$

$$\tau_{m,n} \mid \boldsymbol{\theta}_m \overset{\text{iid}}{\sim} \text{Cat}(\boldsymbol{\theta}_m), \tag{5.7}$$

$$\boldsymbol{\theta}_m \mid \boldsymbol{\alpha} \overset{\text{ind}}{\sim} \text{Dir}(\boldsymbol{\alpha}). \tag{5.8}$$

Now $\boldsymbol{\tau}_m = (\tau_{m,1}, ..., \tau_{,N})$ is defined as the collection of unobserved topics for document $\mathbf{w}_m = (w_{m,1}, ..., w_{m,N})$. The joint distribution of observed document $\mathbf{w}$ and the latent variables, $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$, conditional on (hyper-)parameters $\boldsymbol{\eta}$ and $\boldsymbol{\alpha}$ now becomes:

$$\text{p}(\mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\tau} \mid \boldsymbol{\alpha}, \boldsymbol{\eta}) = \text{p}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \prod_{n=1}^{N} \text{p}(\tau_n \mid \boldsymbol{\theta}) \text{p}(w_n \mid \tau_n, \boldsymbol{\eta}), \tag{5.9}$$

$$= \text{p}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \prod_{n=1}^{N} \theta_{\tau_n} \eta_{\tau_n, w_n}, \tag{5.10}$$

where $\text{p}(\tau = \tau_n \mid \boldsymbol{\theta}) = \theta_{\tau_n}$, the probability of topic τ_n for this document, and $\text{p}(w = w_n \mid \tau = \tau_n, \boldsymbol{\eta}) = \eta_{\tau_n, w_n}$, the probability of word w_n in topic τ_n. Note here that this distribution is independent of the order of words and topics in the document. This assumption is called the *exchangeability* assumption.

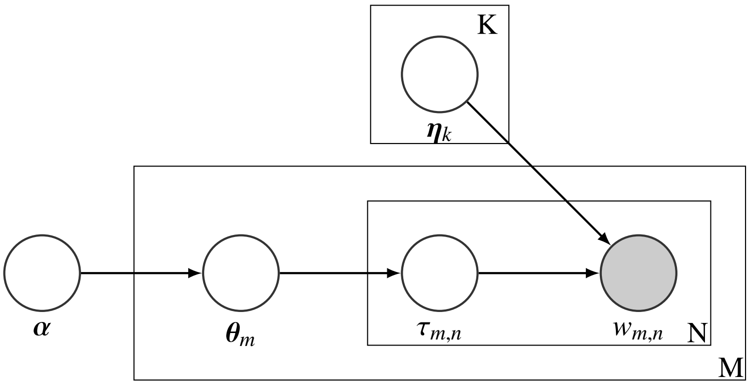


Figure 5.4: Conditionally independent hierarchical model representation of the LDA model [90]. The grey circle is the observed collection of words per document. The lower two boxes represent replicates of "documents"(outer) and "words per document" (inner) and the upper box represent the parameter-vectors for the K words-per-topic-distributions.

5.3 Inference and parameter estimation

Now the goal is to estimate the (hyper-)parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$, and the latent variables $\{\boldsymbol{\tau}_m\}_{m=1}^{M}$ and $\{\boldsymbol{\theta}_m\}_{m=1}^{M}$, based on the observed data $C = \{\mathbf{w}_m\}_{m=1}^{M}$. Blei et al. propose an empirical Bayes method for doing this [90]. Expectation Maximisation and variational inference are combined to make approximations, since the Bayesian hierarchical model in Equations (5.6) - (5.8) is too complex to fit directly. This section treats each of these concepts in the context of LDA.

5.3.1 Empirical Bayes method for parameter estimation

The empirical Bayes method is a method that makes inference about the parameters of a Bayesian hierarchical model – such as LDA – by first estimating the prior distribution from the data, and then using these estimated priors to make inference about the latent parameters[3][97]. Applied to LDA, empirical Bayes consists of two main steps:

1. First estimate the (hyper-)parameters in the model by maximizing the *marginal likelihood* or, equivalently, the *marginal log-likelihood*:

$$\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\eta}} = \underset{\boldsymbol{\alpha},\boldsymbol{\eta}}{\operatorname{argmax}} \log \mathrm{p}\left(C \mid \boldsymbol{\alpha}, \boldsymbol{\eta}\right). \tag{5.11}$$

2. Then for each document $\mathbf{w}$ make inference about the latent variables $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ by drawing from their *joint posterior distribution*:

$$\mathrm{p}\left(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \mathbf{w}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\eta}}\right). \tag{5.12}$$

   Note that. conditional on the observed $\mathbf{w}$ and parameter estimates, these latent variables are independently distributed among documents, and thus can be evaluated individually for each document. Estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\tau}}$ can be inferred by computing the posterior mode or mean from Equation 5.12.

Approximation methods are required for both steps as is motivated in the rest of this section.

[3] This is different from a full Bayesian approach, which chooses some fixed hyper-parameters of the prior distributions – allowing the expression of prior knowledge or ideas about the unknown parameters – and then uses these to compute the posterior distribution.

Marginal likelihood

The marginal likelihood is calculated by integrating out the latent variables in equation 5.9. This leads to the following marginal likelihood of a document:

$$\begin{aligned}
\mathrm{p}(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\eta}) &= \int_{\boldsymbol{\Theta}} \mathrm{p}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \prod_{n=1}^{N} \left[\sum_{\tau_n} \mathrm{p}(w_n \mid \tau_n, \boldsymbol{\eta}) \mathrm{p}(\tau_n \mid \boldsymbol{\theta}) \right] \mathrm{d}\boldsymbol{\theta} && (5.13) \\
&= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \int_{\boldsymbol{\Theta}} \left(\prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \right) \left(\prod_{n=1}^{N} \sum_{k=1}^{K} \prod_{v=1}^{V} (\theta_k \eta_{k,v})^{I(w_n = v)} \right) d\boldsymbol{\theta}, && (5.14)
\end{aligned}$$

with $I(w_n = v)$ the indicator function taking the value 1 if $w_n = v$ and 0 otherwise. Equation 5.14 follows from inserting the probability distribution functions of the categorical and Dirichlet distributions as defined in Appendix C. For the corpus C the likelihood then becomes:

$$\begin{aligned}
\mathrm{p}(C \mid \boldsymbol{\alpha}, \boldsymbol{\eta}) &= \prod_{m=1}^{M} \mathrm{p}(\mathbf{w}_m \mid \boldsymbol{\alpha}, \boldsymbol{\eta}) && (5.15) \\
&= \prod_{m=1}^{M} \left(\int_{\boldsymbol{\Theta}} \mathrm{p}(\boldsymbol{\theta}_m \mid \boldsymbol{\alpha}) \prod_{n=1}^{N_m} \left[\sum_{\tau_{m,n}} \mathrm{p}\left(w_{m,n} \mid \tau_{m,n}, \boldsymbol{\eta}\right) \mathrm{p}\left(\tau_{m,n} \mid \boldsymbol{\theta}_m\right) \right] d\boldsymbol{\theta} \right) && (5.16) \\
&= \prod_{m=1}^{M} \left(\frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \int_{\boldsymbol{\Theta}} \left(\prod_{k=1}^{K} \theta_{m,k}^{\alpha_k - 1} \right) \left(\prod_{n=1}^{N} \sum_{k=1}^{K} \prod_{v=1}^{V} (\theta_{m,k} \eta_{k,v})^{I(w_{m,n} = v)} \right) d\boldsymbol{\theta} \right). && (5.17)
\end{aligned}$$

Equations 5.14 and 5.17 are intractable[4] due to the coupling of $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ [99]. Various approximation algorithms are proposed for LDA [90]. One of these will be introduced in Section 5.3.2.

Joint posterior distribution of the latent variables

To derive the joint posterior probability of the latent variables $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$, Bayes' rule is applied:

$$\mathrm{p}(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{\mathrm{p}(\boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\eta})}{\mathrm{p}(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\eta})} \tag{5.18}$$

. This probability is intractable since the denominator is equal to Equation (5.14).

The intractability of Equations 5.17 and 5.18 imply a need for approximation in both steps of the empirical Bayes method. A wide variety of such approximate inference algorithms for LDA is suggested in literature, including Laplace approximation, variational approximation, and Markov chain Monte Carlo [90]. In this thesis, a combination of expectation maximisation (EM) and variational inference is used.

5.3.2 Expectation Maximisation for latent variable estimation

In models that contain latent variables – like LDA – the explicit computation of the (log-)likelihood as a function of the unknown parameters is often impossible or computationally hard. The expectation maximisation (EM) algorithm tries to solve this problem.

[4] Tractability means there is an algorithm that approximates the solution with error ϵ using $n = n(\epsilon, d)$ samples of the function, where $n(\epsilon, d)$ is polynomially bounded in ϵ^{-1} and d, d the dimensionality [98].

EM-algorithm. In order to approximate some unknown parameters $\boldsymbol{\psi}$ in a model with observed data $\mathbf{x}$ and latent variables $\mathbf{z}$, the EM algorithm approximates the log-likelihood by iterating through the following steps:

1. Choose initial parameter estimate $\boldsymbol{\psi}^{(0)}$ and set $i = 0$.
2. *E-step*. For step i, derive $\mathrm{p}(\mathbf{z}|\mathbf{x}, \boldsymbol{\psi}^{(i)})$, the posterior distribution of the latent variables conditional on the current parameter estimate, and use this to construct the *expected complete log-likelihood*, using notation $\mathbb{E}_{\mathbf{Z}|\mathbf{x},\boldsymbol{\psi}^{(i)}}\left[\mathbf{y}(\mathbf{z})\right] = \int_{\mathcal{Z}} \mathbf{y}(\mathbf{z})\mathrm{p}\left(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\psi}^{(i)}\right) \mathrm{d}\mathbf{z}$:
$$Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(i)}) = \mathbb{E}_{\mathbf{Z}|\mathbf{x},\boldsymbol{\psi}^{(i)}}\left[\log\left(\mathrm{p}(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\psi})\right)\right], \tag{5.19}$$
which is a function of the parameter $\boldsymbol{\psi}$ depending implicitly on $\boldsymbol{\psi}^{(i)}$.
3. *M-step*. Now find the next approximation $\boldsymbol{\psi}^{(i+1)}$ by maximising this expression over all possible values for $\boldsymbol{\psi}$:
$$\boldsymbol{\psi}^{(i+1)} = \underset{\boldsymbol{\psi}}{\operatorname{argmax}}\, Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(i)}). \tag{5.20}$$
4. Iterate steps 2 and 3 until convergence.

The parameters are now estimated as $\hat{\boldsymbol{\psi}} = \boldsymbol{\psi}^{(i)}$, the last estimate before the algorithm converged. inference about the latent variables can be made by computing the posterior mean $\hat{\mathbf{z}} = \mathbb{E}_{\mathbf{Z}|\mathbf{x},\hat{\boldsymbol{\psi}}}\left[\mathbf{z}\right]$, or the posterior mode $\hat{\mathbf{z}} = \underset{\mathbf{z}\in\mathcal{Z}}{\operatorname{argmax}}\, \mathrm{p}\left(\mathbf{z}|\mathbf{x}, \hat{\boldsymbol{\psi}}\right)$. In Appendix D, the EM algorithm is worked out in more detail for the general case, including an example latent variable model that is less complex than LDA.

Now in the context of using empirical Bayes for LDA, the marginal log-likelihood that needs to be maximised is given by:

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\eta}) = \log \mathrm{p}(C \mid \boldsymbol{\alpha}, \boldsymbol{\eta}) = \sum_{m=1}^{M} \log \mathrm{p}(\mathbf{w}_m \mid \boldsymbol{\alpha}, \boldsymbol{\eta}), \tag{5.21}$$

which is intractable. An attempt to solve this with EM is worked out in Appendix D.1. However, the intractability of the joint posterior distribution of latent variables $\boldsymbol{\tau}$ and $\boldsymbol{\theta}$ (Equation (5.9)) leaves computing the expected complete log-likelihood

$$Q(\boldsymbol{\alpha}, \boldsymbol{\eta}; \boldsymbol{\alpha}^{(i)}, \{\boldsymbol{\eta}^{(i)}\}) = \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\tau}|\mathbf{w},\boldsymbol{\alpha}^{(i)},\boldsymbol{\eta}^{(i)}}\left[\log \mathrm{p}\left(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\eta}\right)\right]. \tag{5.22}$$

still infeasible. Therefore, variational inference is proposed as a method to approximate this joint posterior distribution.

5.3.3 Variational inference for EM estimation

The intractability of Equation (5.17) comes from the coupling between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ indicated by the edges between $\boldsymbol{\theta}$, $\mathbf{w}$ and $\boldsymbol{\tau}$ in Figure 5.4. The idea of applying variational inference is now to approximate $\mathrm{p}\left(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\eta}\right)$ by *variational* distribution $\mathrm{q}(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \boldsymbol{\gamma}, \boldsymbol{\phi})$, defined by:

$$\mathrm{q}(\boldsymbol{\theta}_m, \boldsymbol{\tau}_m \mid \boldsymbol{\gamma}_m, \boldsymbol{\phi}_m) = \mathrm{q}(\boldsymbol{\theta}_m \mid \boldsymbol{\gamma}_m) \prod_{n=1}^{N_m} \mathrm{q}(\tau_{m,n} \mid \phi_{m,n}), \tag{5.23}$$

with variational parameters $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$, resulting in a decoupling of $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$. A natural choice for the variational marginal distributions on the latent variables is:

$$\begin{aligned} \boldsymbol{\theta}_m &\sim \mathrm{Dir}(\boldsymbol{\gamma}_m), \\ \tau_{m,n} &\sim \mathrm{Cat}(\phi_{m,n}), \end{aligned}$$

which are document specific and thus will be estimated separately for each document.

The key idea of variational expectation maximisation is now to do the following in every iteration of the EM-algorithm:

1. Find $\boldsymbol{\gamma}_m^{(i)}$ and $\boldsymbol{\phi}_m^{(i)}$ for every document, such that: $\mathrm{q}\left(\boldsymbol{\theta}_m, \boldsymbol{\tau}_m \mid \boldsymbol{\gamma}_m^{(i)}, \boldsymbol{\phi}_m^{(i)}\right)$ is the "best" approximation of $\mathrm{p}(\boldsymbol{\theta}_m, \boldsymbol{\tau}_m \mid \mathbf{w}_m, \boldsymbol{\alpha}^{(i)}, \boldsymbol{\eta}^{(i)})$.
2. Instead of maximising the marginal (log-)likelihood, now maximise some lower bound $L\left(\boldsymbol{\phi}^{(i)}, \boldsymbol{\gamma}^{(i)}; \boldsymbol{\alpha}, \boldsymbol{\eta}\right)$ on this likelihood.

Jordan et al. showed that the marginal log-likelihood of a document can be rewritten as:

$$\log \mathrm{p}(\mathbf{w}_m \mid \boldsymbol{\alpha}, \boldsymbol{\eta}) = L(\boldsymbol{\gamma}_m, \boldsymbol{\phi}_m; \boldsymbol{\alpha}, \boldsymbol{\eta}) + D_{KL}\left(\mathrm{q}(\boldsymbol{\theta}_m, \boldsymbol{\tau}_m \mid \boldsymbol{\gamma}_m, \boldsymbol{\phi}_m) || \mathrm{p}(\boldsymbol{\theta}_m, \boldsymbol{\tau}_m \mid \mathbf{w}_m, \boldsymbol{\alpha}_m, \boldsymbol{\eta}_m)\right), \tag{5.24}$$

using lower bound

$$L(\boldsymbol{\gamma}_m, \boldsymbol{\phi}_m; \boldsymbol{\alpha}, \boldsymbol{\eta}) = \mathbb{E}_{\mathrm{q}}\left[\log \mathrm{p}(\boldsymbol{\theta}_m, \boldsymbol{\tau}_m, \mathbf{w}_m \mid \boldsymbol{\alpha}, \boldsymbol{\eta})\right] - \mathbb{E}_{\mathrm{q}}\left[\log \mathrm{q}(\boldsymbol{\theta}_m, \boldsymbol{\tau}_m)\right], \tag{5.25}$$

with $\mathbb{E}_{\mathrm{q}}[\cdot]$ the expectation with respect to the variational distribution, and the Kullback-Leibler (KL) divergence $D_{KL}\left(\mathrm{q}(x) \parallel \mathrm{p}(x)\right) = \int \mathrm{q}(x) \log \frac{\mathrm{q}(x)}{\mathrm{p}(x)} \mathrm{d}x$ [100]. An important observation here, is that for fixed $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$, maximising the lower bound $L(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\eta})$ is equivalent to minimising the KL-divergence between the variational distribution and the true posterior distribution of the latent variables. Now the variational expectation maximisation (VEM) algorithm is given.

VEM-algorithm.

1. Choose initial $\boldsymbol{\alpha}^{(0)}$ and $\boldsymbol{\eta}^{(0)}$. Set $i = 0$.
2. *E-step.* Now in each step i, find the optimising values of the variational parameters $(\boldsymbol{\gamma}_m^{(i)}, \boldsymbol{\phi}_m^{(i)})$ for each document m:

$$(\boldsymbol{\gamma}_m^{(i)}, \boldsymbol{\phi}_m^{(i)}) = \underset{(\boldsymbol{\gamma}, \boldsymbol{\phi})}{\operatorname{argmin}} D_{KL}\left(\mathrm{q}(\boldsymbol{\theta}_m, \boldsymbol{\tau}_m \mid \boldsymbol{\gamma}, \boldsymbol{\phi}) || \mathrm{p}(\boldsymbol{\theta}_m, \boldsymbol{\tau}_m \mid \mathbf{w}_m, \boldsymbol{\alpha}^{(i)}, \boldsymbol{\eta}^{(i)})\right). \tag{5.26}$$

3. *M-step.* Now maximise the lower bound – i.e. Equation (5.25)– with respect to parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$:

$$(\boldsymbol{\alpha}^{(i+1)}, \boldsymbol{\eta}^{(i+1)}) = \underset{(\boldsymbol{\alpha}, \boldsymbol{\eta})}{\operatorname{argmax}} L(\boldsymbol{\gamma}^{(i)}, \boldsymbol{\phi}^{(i)}; \boldsymbol{\alpha}, \boldsymbol{\eta}), \tag{5.27}$$

   with $L(\boldsymbol{\gamma}^{(i)}, \boldsymbol{\phi}^{(i)}; \boldsymbol{\alpha}, \boldsymbol{\eta}) = \sum_{m=1}^{M} L(\boldsymbol{\gamma}_m, \boldsymbol{\phi}_m; \boldsymbol{\alpha}, \boldsymbol{\eta})$, the lower bound of $\ell(\boldsymbol{\alpha}, \boldsymbol{\eta})$, the marginal-likelihood of corpus C.
4. Steps 2 and 3 are iterated until convergence of the lower bound.

Note that this overall procedure can be viewed as coordinate ascent in L. The minimisation in the E-step is being done with an iterative fixed-point method [90]. In each update of the M-step, $\boldsymbol{\alpha}^{(i)}$ can be approximated using an efficient Newton-Raphson method, while estimates $\boldsymbol{\eta}^{(i)}$ can be computed analytically [90]. The estimation for the (hyper-)parameters is now given by the last estimates before convergence:

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^{(i+1)}, \tag{5.28}$$

$$\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}^{(i+1)}. \tag{5.29}$$

Under the assumption that the true posterior probability is approximated by the variational posterior probability, the later can be used to estimate the latent variables:

$$\hat{\boldsymbol{\theta}}_m = \mathbb{E}_{\boldsymbol{\gamma}_m^*}\left[\boldsymbol{\theta}_m\right] = \frac{\boldsymbol{\gamma}_m^*}{\sum_i \gamma_{m,i}^*}, \tag{5.30}$$

$$\hat{\boldsymbol{\tau}}_{m.n} = \mathbb{E}_{\phi_{m,n}^*}\left[\tau_{m,n}\right] = \phi_{m,n}^*. \tag{5.31}$$

The last step is an elegant result: there is no need to use the variational distribution as an approximation of the true posterior distribution to generate estimates for the latent variables – the second step of empirical Bayes – since these can be computed exactly from the estimated marginal variational distributions.

5.4 Energy consumption lifestyles and LDA

The statistical model on which LDA is based is explained in Section 5.2, and a mathematical procedure for estimating the parameters of this model is introduced in Section 5.3. Now it is time to clarify the analogy between text modelling and energy consumption, before actually fitting the model on the observed data.

5.4.1 Analogy with text modelling

Going back to the start of this chapter, the main issue that needs to be dealt with was structuring the load shape data in a way such that different types of consumers could be classified (classification), by incorporating typical patterns that correspond to behaviour (lifestyle identification) that is more granular and helps understanding what type of consumers are in the dataset (interpretation). This implies modelling energy consumption as a representation of behaviour. Because behaviour is not constant among different days of the week, and also has variability within the same weekdays, the concept of a *lifestyle* is introduced. For example, an energy consumer - a family in this case - can have a 9-5 job during most weekdays (*lifestyle 1*), the children home early on Wednesdays (*lifestyle 2*) and an active lifestyle during the weekends (*lifestyle 3*) with a monthly movie night on Saturday (*lifestyle 4*). Each of these lifestyles has its typical set of load shape(s) that are likely to be observed. A 9-5 working lifestyle will, for example, have a small peak around 8AM, a bigger peak around 6PM, and an occasional third peak at 9:30PM when there is football on the television. Figure 5.5 visualises LDA in the energy context.

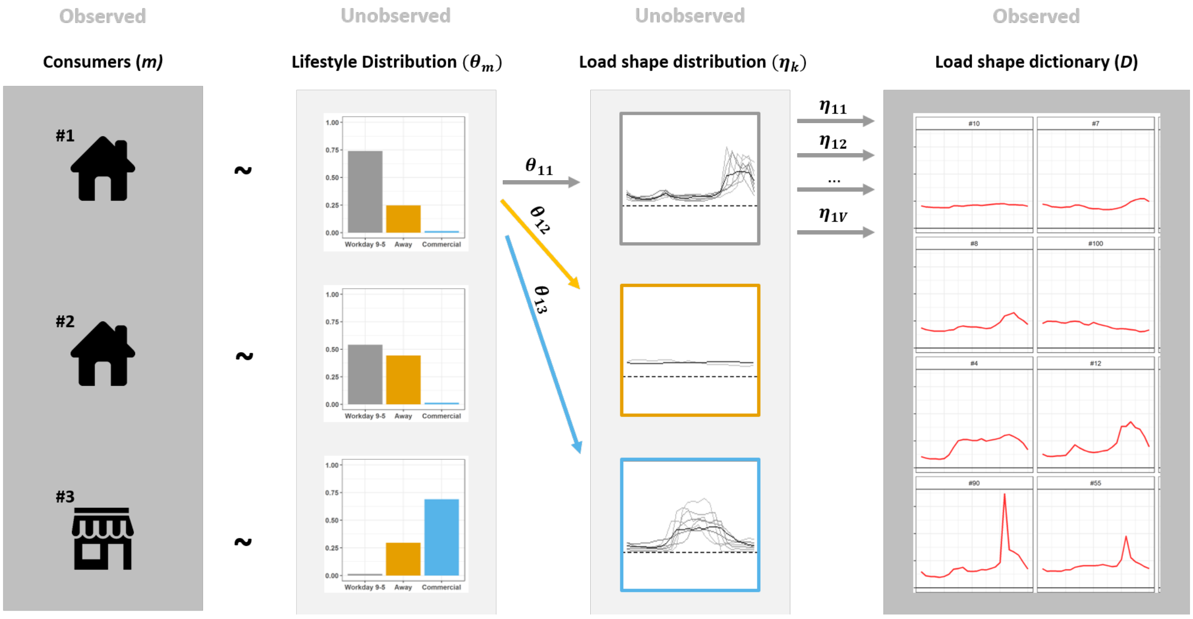


Figure 5.5: Schematic representation of LDA for the energy consumption context. Each consumer is characterised by an unobserved mix of lifestyles, and each of these lifestyles is characterised by a mix of load shapes from the dictionary. In this dummy example, the lifestyles are 9-to-5 workday (grey), away (orange) and commercial (blue).

Terminology

- The *dictionary* D of size V is the collection of cluster centers derived by applying Adaptive K-means to the collection of daily energy consumption data in chapter 4.

$$D = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_V\},$$

  with $\boldsymbol{\mu}_v \in \mathbb{R}^{24}$ the mean of the v-th cluster. This corresponds to a dictionary of words in the context of text analysis.

- The basic unit of discrete data w in this framework now corresponds to an individual *load shape* from the dictionary:

$$w_i = k \quad \Leftrightarrow \quad f(\mathbf{x}_i) = \boldsymbol{\mu}_k \in D,$$

  with $f(\cdot)$ the function mapping daily consumption to a load shape in the dictionary. Load shapes correspond to '*words*' in most LDA-literature.
- The analogy of a *document* is now taken by one single consumer's *load shape series*, consisting of N load shapes denoted by $\mathbf{w} = (w_1, w_2, ..., w_N)$, a series of indices w_i in the load shape dictionary D. When annual energy consumption data is used for analysis, $N = 365$ for all $\mathbf{w}_m$.
- A *corpus* C is now a collection of M load shape series denoted by $C = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$. This is the collection of all observed smart meter data transformed into load shape series.
- The latent *topics* τ in the LDA model are now replaced by the concept of an (energy) *lifestyle*. More specifically, it is assumed that one's energy consumption pattern represents behaviour that is linked to a lifestyle, for example working from 9-5 or a movie night on saturday. These lifestyles and their corresponding load shapes are the key elements for inference in this context.

Generative model

Each annual series of energy consumption data $\mathbf{w}_m$ is now generated assuming the following model:

1. Draw the parameter $\boldsymbol{\theta}_m$ for the τ-per-home distribution from a Dirichlet distribution:

$$\boldsymbol{\theta}_m \mid \boldsymbol{\alpha} \overset{\text{iid}}{\sim} \text{Dir}(\boldsymbol{\alpha}).$$

2. Then for each of the $n = 1, ..., N$ load shapes $w_{m,n} \in \mathbf{w}_m$:

   (a) draw a lifestyle $\tau_{m,n}$ from a Categorical distribution with parameter $\boldsymbol{\theta}_m$:

$$\tau_{m,n} \mid \boldsymbol{\theta}_m \overset{\text{iid}}{\sim} \text{Cat}(\boldsymbol{\theta}_m).$$

   (b) draw a load shape $w_{m,n}$ from a Categorical distribution with parameter $\boldsymbol{\eta}_{\tau_{m,n}}$:

$$w_{m,n} \mid \tau_{m,n}, \boldsymbol{\eta} \overset{\text{ind}}{\sim} \text{Cat}(\boldsymbol{\eta}_{\tau_{m,n}}),$$

with $\boldsymbol{\eta} = \{\boldsymbol{\eta}_k\}_{k=1}^K$ the parameters of the load-shape-per-lifestyle distributions and $\boldsymbol{\alpha}$ the hyperparameter for the Dirichlet prior of $\boldsymbol{\theta}_m$.

5.4.2 Implied assumptions

Several important assumptions in the energy context are implied here.

- The exchangeability assumption implies that the distribution of lifestyles and load shapes are independent of time-of-year and day of week. This is in reality not the case. However, as will be seen in the next section, the identified lifestyles seem to be consistent and interpretable, thus even with this assumption the results are promising.
- The fact that a lifestyle has a distribution over all load shapes in the dictionary implies that an observed load shape can belong to different lifestyles. This assumption is not unrealistic, since different behaviour can lead to the same energy consumption: i.e. the peak at 9 pm can be caused by someone watching football on TV or a student charging his laptop to start studying for an exam.
- The model assumes that lifestyles are uncorrelated. However, it might be reasonable to believe that families with young children expose typical combinations of lifestyles, while a retired couple exposes others. To treat this issue in text modelling, Blei introduced the correlated topic model that includes a correlation structure in the topic-per-home distribution [101]. This is recommended for future model improvement.

5.5 Implementation

In Chapter 4, a dictionary of 633 load shapes was constructed with Adaptive K-means. This dictionary is used to transform each annual energy consumption time series into a discrete series of elements from this dictionary. This transformed series is used as an input for the LDA algorithm that is explained in the previous sections. This is visualized in Figure 5.6.

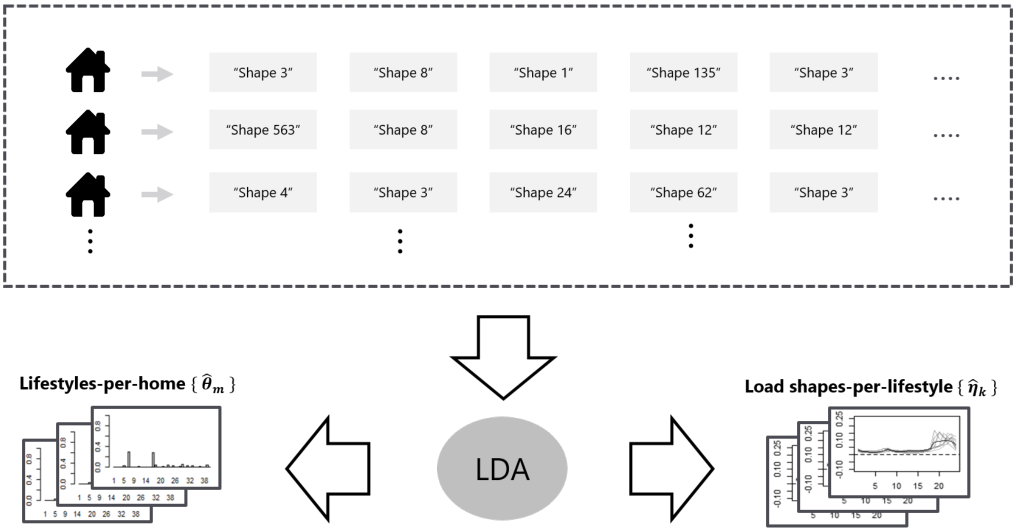


Figure 5.6: Schematic representation of the encoding of all annual consumption time series.

In this section, the transformation process is explained. Furthermore, in the used version of LDA, the number of lifestyles that need to be inferred is still an input parameter of the model. Therefore, the choice of this parameter K will be motivated in the second part of this section.

5.5.1 Data processing

In chapter 4, the NL14conv- and NL15ae-datasets were clustered with Adaptive K-means, such that a set of 633 load shapes was identified that summarized all consumption patterns. This set is called the *load shape dictionary* D of size $V = 633$:

$$D = \{\mu_1, \mu_2, ..., \mu_{633}\}. \tag{5.32}$$

The original dataset contained annual time series **P**, each consisting of 8760 hourly observations for a full year. Each annual consumption is then split into normalized daily consumption vectors $\{(\mathbf{x}_1, ..., \mathbf{x}_{365})_h\}_{h=1}^{5194}$ and then mapped to the index of the nearest load shape in the dictionary as follows:

$$\mathbf{x}_i \rightarrow \mu_{c_i} \rightarrow w_i \tag{5.33}$$

5.5.2 Implentation in R

The **R**-package `Topicmodels` is used for fitting the model. This package provides an interface between **R** and the **C**-code for Latent Dirichlet Allocation models and Correlated Topics Models (CTM) by David M. Blei [90]. In this implementation Variational Expectation Maximization (VEM) is used as an approximation procedure. An implementation that uses Gibbs Sampling is also provided in the package, but because of its computational efficiency VEM is used in further analysis.

This process is visualized in Figures 5.7 and 5.8.

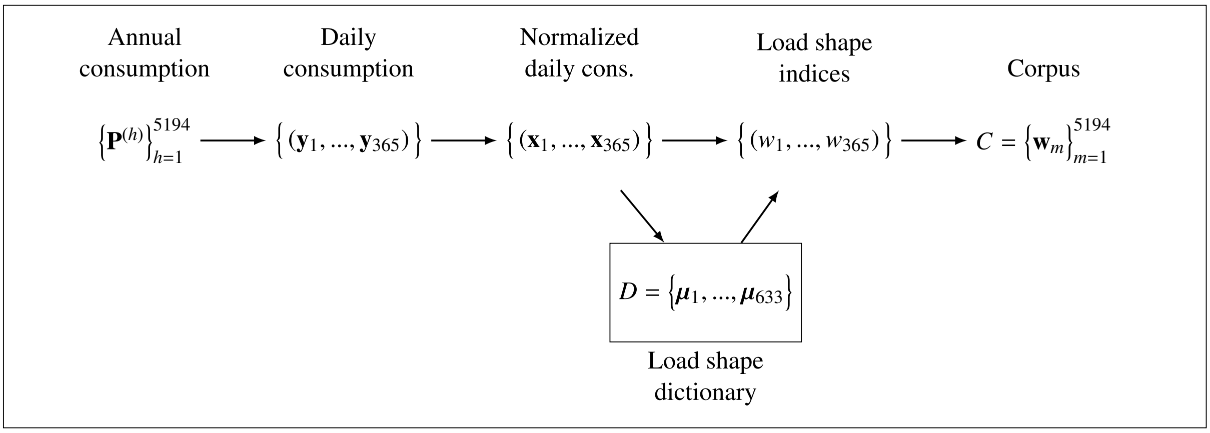

Annual consumption
Daily consumption
Normalized daily cons.
Load shape indices
Corpus
$\{\mathbf{P}^{(h)}\}_{h=1}^{5194}$
$\{(\mathbf{y}_1, ..., \mathbf{y}_{365})\}$
$\{(\mathbf{x}_1, ..., \mathbf{x}_{365})\}$
$\{(w_1, ..., w_{365})\}$
$C = \{\mathbf{w}_m\}_{m=1}^{5194}$
$D = \{\mu_1, ..., \mu_{633}\}$
Load shape dictionary

Figure 5.7: Process of transforming annual time series into corpus C. The annual time series $\{\mathbf{P}\}$ are first chopped into daily consumption vectors $\{\mathbf{y}\}$ and then normalized $\{\mathbf{x}\}$. These vectors were clustered in the previous chapter, resulting in load shape dictionary D. Afterwards, each $\mathbf{x}$ is assigned to the index of the nearest load shape in the dictionary, resulting in corpus C.

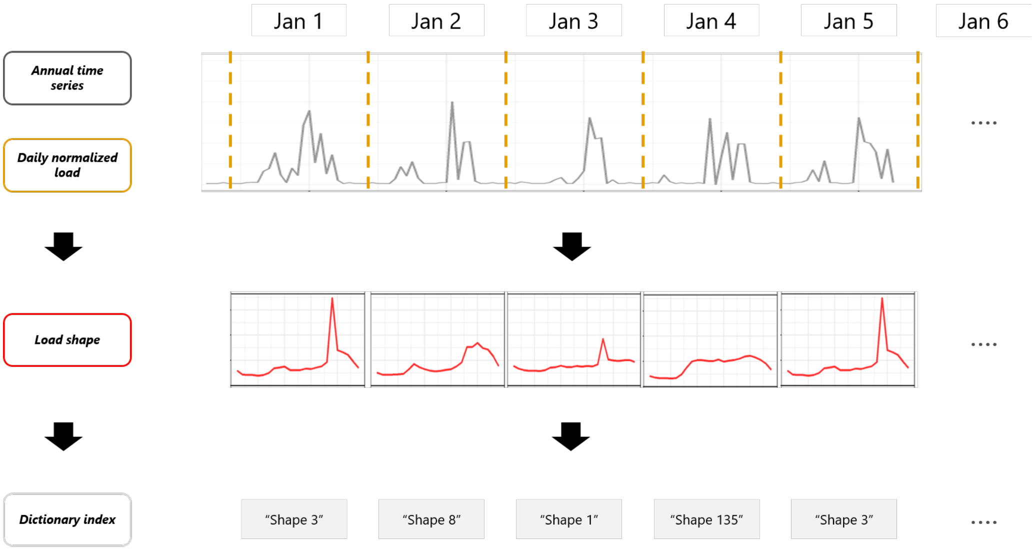


Figure 5.8: Schematic representation of the encoding of an annual consumption time series.

The package requires input in 'DocumentTermMatrix'-format. This is a matrix $DTM \in \mathbb{R}^{M \times V}$, thus rows correspond to consumers and columns to load shapes, which is constructed as follows:

$$DTM = (\mathbf{v}_1, ..., \mathbf{v}_M)^T \;, \tag{5.34}$$

with

$$\mathbf{v}_{m,v} = \sum_{n=1}^{N} I(w_{m,n} = v), \tag{5.35}$$

thus DTM_{ij} indicates the number of occurrences of the j-th load shape in the i-th consumption series.

5.5.3 Choosing the number of lifestyles

The proper Bayesian way to determine the optimal number of lifestyles K would be to assume a prior distribution over the number of topics and include this in the generative model. A Chinese restaurant process is suggested in [90] and worked out in [102] as *hierarchical LDA*, in which not only the number of topics, but also a hierarchical structure among those topics are inferred. An implementation of this method in C^{++} is available on Github. However, no implementation in R is found as of yet, and would take a long time to achieve (and probably also to run). Therefore, two alternative non-Bayesian approaches are taken.

Perplexity and 10-fold cross-validation

In information theory, perplexity is a measure of how well a model predicts a sample and is used to compare different models or parameters choices. If parameter $\hat{\psi}$ is fitted on some training set D_{train}, then the perplexity of a test set D_{test} – which is different from the training set – is defined as the geometric mean per-document likelihood:

$$\text{perplexity}(D_{test}) = \exp\left(-\frac{\sum_{m=\in D_{test}} \log \mathrm{p}(\mathbf{w}_m \mid \hat{\psi})}{\sum_{m \in D_{test}} N_m}\right), \tag{5.36}$$

with $\mathrm{p}(\mathbf{w}_m \mid \hat{\psi})$ the likelihood of the collection of documents for consumer m conditional on the fitted parameters, and N_m the number of words in document $\mathbf{w}_m$.

For 10-fold cross-validation, the data is now split into 10 equal partitions. The model is then fitted on 90% of the data and the perplexity is calculated for the partition that is left out. This is repeated for all 10 partitions. The results for this procedure applied to the corpus C are shown in Figure 5.9a. The perplexity does not have an optimum, but does not decrease much after $K \approx 50$. The elbow method then suggests a choice of $K \in [30, 50]$.

Bayesian Information Criterion

Another evaluation metric is the Bayesian information criterion (BIC), which weighs the model size $n_{par} = K + KV$ (number of fitted parameters) and the likelihood $\hat{L}$ of the Corpus given the fitted parameters:

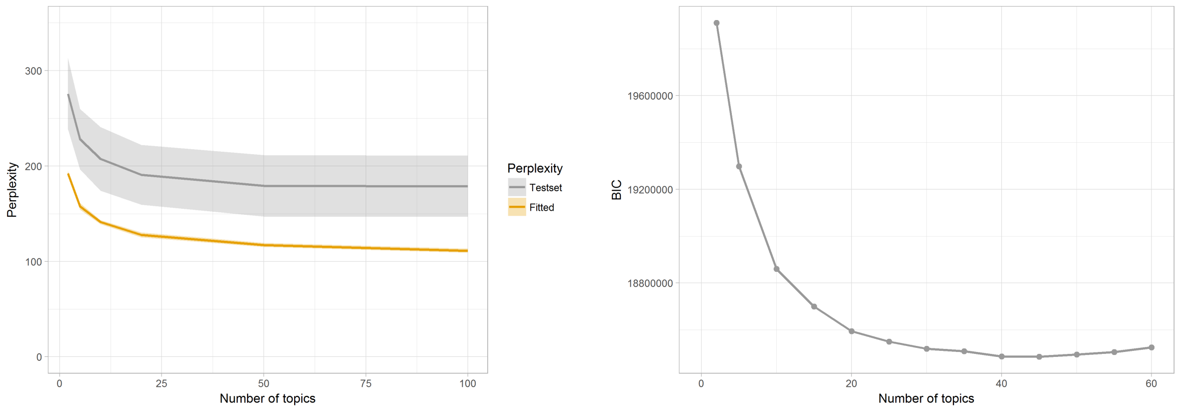
$$\hat{L} = \mathrm{p}(C \mid \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\alpha}}).$$

BIC is defined as follows:

$$\text{BIC}(n_{par}, \hat{L}, m) \quad = \quad \log(m) n_{par} - 2\log(\hat{L}). \tag{5.37}$$

Both should be minimized for ‘optimal’ choice of K. For various model sizes, the results for BIC can be found in Figure 5.9b and Table 5.1. The BIC is at its minimum at $K = 40$.

[5] In Table 5.1 $\log \hat{L}$ and BIC are devided by 10^6



(a) Perplexity *(b) BIC*

Figure 5.9: Perplexity (a) and BIC (b) as a function of the number of fitted lifestyles. The elbow rule applied to the perplexity suggests a choice of $K = 20 - 40$ lifestyles, while the optimum of BIC is at $K = 40$.

Table 5.1: Statistics of runs of LDA with different number of latent topics.[5]

K	n_{var}	$\log \hat{L}$	BIC	$\hat{\alpha}$
2	1,310	-9.950	19.911	0.369
5	3,275	-9.635	19.298	0.241
10	6,550	-9.402	18.860	0.169
15	9,825	-9.307	18.699	0.147
20	13,100	-9.241	18.594	0.126
25	16,375	-9.205	18.550	0.118
30	19,650	-9.176	18.519	0.100
35	22,925	-9.156	18.508	0.090
40	26,200	-9.131	18.485	0.089
45	29,475	-9.116	18.486	0.083
50	32,750	-9.107	18.495	0.076
55	36,025	-9.098	18.505	0.072
60	39,300	-9.094	18.525	0.069

5.6 Results

The LDA model is now fitted on all 5,194 homes with $K = 40$ lifestyles. The output of this are the estimated parameters $\{\hat{\boldsymbol{\theta}}_m\}_{m=1}^{5194}$ and $\{\hat{\boldsymbol{\eta}}_k\}_{k=1}^{40}$, with:

$$\hat{\theta}_{m,i} = \mathrm{p}(\tau_i \mid m), \tag{5.38}$$

the estimated posterior mode of lifestyle τ_i for consumer m, and:

$$\hat{\eta}_{k,j} = \mathrm{p}(w_j \mid \tau_k), \tag{5.39}$$

the estimated posterior mode of load shape w_j for lifestyle τ_k.

5.6.1 Load shapes per lifestyle

To visualize the identified lifestyles, first the characteristic set of load shapes for lifestyle τ is defined as:

$$\Lambda_\psi(\tau) = \left\{\mu_j \in D : \mathrm{p}\left(w_j \mid \tau\right) > \psi\right\}, \tag{5.40}$$

or, in words, all load shapes that have estimated posterior probability greater than ψ for lifestyle τ. For $\psi = 0.05$, three characteristic load shape sets are plotted in Figure 5.10. The characteristic sets for all lifestyles can be found in Appendix E.

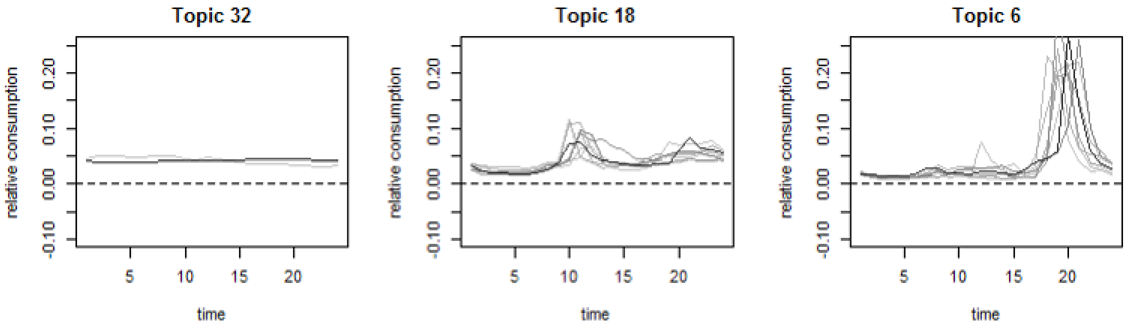


Figure 5.10: $\Lambda_{0.05}(\tau)$ for three inferred lifestyles (= topics). Darker lines indicate higher posterior probability than lighter lines. The numerical similarities of characteristic load shapes of each lifestyle are remarkable, since LDA is agnostic about these similarities: load shapes are only grouped together based on how they are spread among the data.

What is remarkable about the identified lifestyles is that without exception they consist of very homogeneous load shapes, even though the LDA algorithm uses categorical variables as input and thus is agnostic about the numerical similarities of these load shapes. Two load shapes are grouped together in one lifestyle in a similar way as the words ’differentation’ and ’integration’ would be grouped together in one topic - which would probably receive label ’Calculus’ - when text is analysed with LDA. This happens not based on the actual meaning of the words but based on how they are spread through the data: they tend to co-occur through documents.

This observation supports the hypothesis that these identified ’topics’ can actually be interpreted as ‘lifestyles’. Topic 32 in Figure 5.10 is a passive day, Topic 18 someone with higher morning then evening peaks and Topic 6 an evening peaker.

5.6.2 Lifestyles per home

Now the estimated posterior distribution of lifestyles per consumer are considered. For three consumers, $\hat{\theta}_m$ as given in Equation 5.38, are plotted in Figure 5.11.

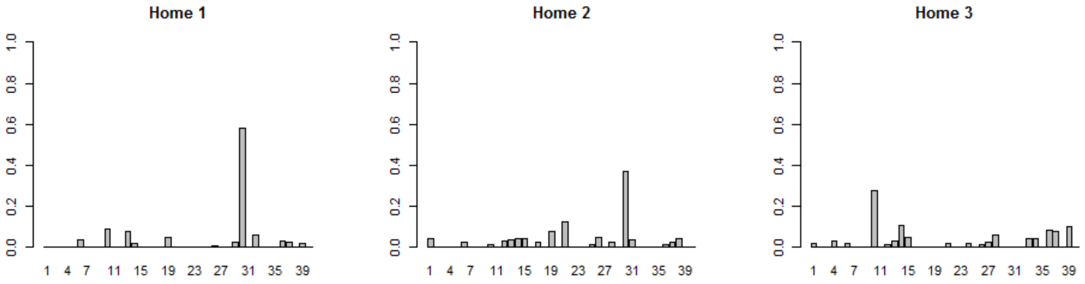


Figure 5.11: Estimated probability vectors $\hat{\theta}_m$ of lifestyles per consumer, for three consumers $m = 1, 2, 3$. Each consumer has its typical set of lifestyles, but whether this is a few or many differs per consumer.

These distributions show that many consumers are characterised by 2 or 3 typical lifestyles with high probability. Others have a less concentrated probability distribution, thus having less constant and more unpredictable energy consumption.

5.7 Conclusion

In this chapter, Latent Dirichlet Allocation (LDA) was introduced as an advanced statistical model to find patterns in large discrete datasets. The original and most common application of LDA is text analysis. In this context, large collections of text are modelled as a mix of topics, with each topic defined by a distribution over words in the dictionary. Due to the complexity of this model, the approximation method Variational Expectation Maximization (VEM) is suggested to make inference about these distributions. In the context of energy consumption modelling, the series of daily energy consumption of a household is seen as a book or document, and lifestyles - typical behaviour patterns - are the 'topics' that need to be made inference about.

In order to apply LDA, the annual energy consumption series are transformed into a series of indices of the daily load shape 'dictionary' constructed in the previous chapter. Applying LDA to this series of discrete data results in estimates for the required distributions of load-shapes-per-life-style and life-styles-per-consumer. BIC and Perplexity are used as measures to optimize the model size. This resulted in 40 lifestyles, each with a remarkable homogeneous set of characteristic load shapes.

In the next chapter, the estimated life-styles-per-consumer distributions $\{\hat{\boldsymbol{\theta}}_m\}_{m=1}^{5194}$ will be used as a feature representation for consumer clustering: the task of grouping consumers together with a similar distribution of lifestyles.

Chapter 6

Lifestyle based consumer segmentation

In the previous chapters, a framework was constructed to identify a set of lifestyles based on energy consumption data. Applying this method to a dataset of annual energy consumption of approximately 5,194 consumers resulted in the identification of 40 lifestyles, each defined by a probability distribution over the elements in the dictionary of load shapes. Furthermore, a distribution over these lifestyles was estimated for each consumer.

In this chapter an attempt is made to segment consumers based on their estimated lifestyle distributions. This approach serves various tasks. The first one is to identify residential and commercial occupancies. Furthermore, assessing the presence of solar panels and other significant NETs is a purpose for which this approach could be suitable. Lastly and foremost, this segmentation could lead to an interpretable grouping of residential consumers based on energy consumption data solely, thus helping to understand what kind of people - e.g. a young family, students or retirees - are living 'behind the meter'.

First the used methodology for clustering is introduced. This includes the used distance metric, the clustering procedure and the dimensionality reduction method for visualisation. The implementation is explained briefly in Section 6.2. Section 6.3 presents and discusses some results of the segmentation. A validation of the developed method will be performed in Chapter 7, by testing how well it segments the all-electric homes from the conventional dataset. In the last section of this chapter, some main conclusions are stated.

6.1 Clustering probability distributions

The goal in this chapter is to group together electricity consumers that have similar energy consumption patterns. However, similarity is not well-defined for a probability distribution. Next to this, a clustering procedure needs to be chosen in order to find the desired clusters effectively. In short, the four following choices need to be made:

- the feature space in which consumers are represented;
- the similarity or distance measure on this space;
- the clustering procedure to group;
- a dimensionality reduction for visualisation purposes.

6.1.1 Representation

In the previous chapter, the posterior lifestyle probabilities were estimated for each consumer. These probabilities are defined by a 40-dimensional vector $\hat{\boldsymbol{\theta}}_m$. Since 'the real' $\boldsymbol{\theta}_m$ will not be considered in any further analysis, the tilde-sign is omitted from here. Therefore, each of the 5,194 consumers is represented by its estimated posterior distribution. This results in datapoints Θ:

$$\Theta = \{\boldsymbol{\theta}_m\}_{m=1}^{5194}, \tag{6.1}$$

with $\Theta \subset \{\boldsymbol{\theta} : ||\boldsymbol{\theta}||_1 = 1, \boldsymbol{\theta} \in \mathbb{R}^{40}_{\geq 0}\}$.

6.1.2 Distance measure

In Chapter 4, normalized daily loads were clustered based on their quantitative similarities with adaptive K-means. In that context, the observation were in a space with some physical interpretation. Thus, using a physical distance measure such as the Euclidean distance made sense. In this chapter, however, the data is in a probability space. Therefore, a measure that operates on this space is preferred. Such measures are the Kullback-Leibler Divergence and the Chi-Squared distance.

Symmetric Kullback-Leibler Divergence

The Kullback-Leibler divergence (KL-divergence) from discrete probability distributions defined by parameters $\boldsymbol{\theta}_j$ to $\boldsymbol{\theta}_i$ is defined by [103]:

$$D_{\text{KL}}(\boldsymbol{\theta}_i||\boldsymbol{\theta}_j) = \sum_k \boldsymbol{\theta}_{i,k} \log \frac{\boldsymbol{\theta}_{i,k}}{\boldsymbol{\theta}_{j,k}}. \tag{6.2}$$

It is clear that $0 \leq D_{\text{KL}}(\cdot||\cdot)$, with equality to zero when the two distributions are identical. Note that $D_{\text{KL}}(\cdot||\cdot)$ is asymmetric. This is not a favourable property, hence the symmetric KL-divergence is defined as:

$$D_{SKL}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \frac{D_{\text{KL}}(\boldsymbol{\theta}_i||\boldsymbol{\theta}_j) + D_{\text{KL}}(\boldsymbol{\theta}_j||\boldsymbol{\theta}_i)}{2}. \tag{6.3}$$

Chi-squared distance

The Chi-squared distance (χ^2-distance) is based on Pearson's chi-squared test statistic [104], and can be seen as a second order Taylor approximation of the KL-divergence:

$$D_{\chi^2}(\boldsymbol{\theta}_i||\boldsymbol{\theta}_j) = \sum_k \frac{\left(\boldsymbol{\theta}_{i,k} - \boldsymbol{\theta}_{j,k}\right)^2}{\boldsymbol{\theta}_{j,k}}, \tag{6.4}$$

or the symmetrised KL-divergence:

$$D_{S\chi^2}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \frac{D_{\chi^2}(\boldsymbol{\theta}_i||\boldsymbol{\theta}_j) + D_{\chi^2}(\boldsymbol{\theta}_i||\boldsymbol{\theta}_j)}{2}. \tag{6.5}$$

Other measures

Some other measures that are considered are the Euclidean distance:

$$D_{\text{EUCL}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = ||\boldsymbol{\theta}_i - \boldsymbol{\theta}_j||_2, \tag{6.6}$$

and the Cosine distance:

$$D_{\text{COS}}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = 1 - \frac{\boldsymbol{\theta}_i \cdot \boldsymbol{\theta}_j}{||\boldsymbol{\theta}_i||_2 ||\boldsymbol{\theta}_j||_2}. \tag{6.7}$$

For each of the distance measures $D(\cdot, \cdot)$, a symmetric distance matrix Δ can be defined such that:

$$\Delta_{i,j} = D(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j). \tag{6.8}$$

6.1.3 Clustering methodology

For a given distance matrix Δ, **hierarchical clustering** is a class of clustering methodologies that seeks a hierarchy of clusters based on the mutual distances between observations [105]. Hierarchical clustering has the distinct advantage that any distance metric can be used. In fact, the actual observations are not required as input, only the distance matrix Δ is used in the clustering procedure. Two general approaches are distinguished:

- **Agglomerative**: A bottom-up approach, starting with every observation in its own cluster ($K = N$) and merging the two closest (in some sense) clusters in every step.
- **Divisive**: A top-down approach, starting with every observation in one cluster ($K = 1$) and splitting one cluster at every step.

Agglomerative clustering is most frequently used, mainly due to its superior computation complexity[1]. Taking into account the total number of current observations (5,194) and the potential number of future observations, only agglomerative hierarchical clustering is considered from here.

At clustering step 0 there are $K_0 = N$ clusters. At a given step l in the clustering procedure, $K_l = N - l$ distinct clusters $\mathbb{C}_l = \{C_1, ..., C_{K_l}\}$ are formed. In the next step, the two clusters C_a and C_b $(a \neq b)$ are merged if they minimise a chosen linkage criterion. Several of these criteria[2] are:

- **Complete linkage**:
$$\underset{(C_a,C_b)\subset \mathbb{C}_l}{\operatorname{argmin}} \max\{D(x,y) : x \in C_a, y \in C_b\}. \tag{6.9}$$
- **Single linkage**:
$$\underset{(C_a,C_b)\subset \mathbb{C}_l}{\operatorname{argmin}} \min\{D(x,y) : x \in C_a, y \in C_b\}. \tag{6.10}$$
- **Average linkage**:
$$\underset{(C_a,C_b)\subset \mathbb{C}_l}{\operatorname{argmin}} \frac{1}{|C_a||C_b|} \sum_{x\in C_a} \sum_{y \in C_b} D(x,y). \tag{6.11}$$
- **Centroid linkage**:
$$\underset{(C_a,C_b)\subset \mathbb{C}_l}{\operatorname{argmin}} D(c_a, c_b). \tag{6.12}$$

  with c_a and c_b the centroids of the respective clusters.

The distance measure $D(\cdot,\cdot)$ can be any symmetric distance measure.

6.1.4 Dimensionality reduction

A dimensionality reduction is performed in order to visualise the multi-dimensional data. Two methods to do this are considered: principal component analysis (PCA) [106] and t-distributed stochastic neighbour embedding (t-SNE) [107]. Next to visualising multi-dimensional data, dimensionality reduction can be deployed to improve clustering results, since it limits the effect of the curse of dimensionality at the price of losing some information. Both applications – visualisation and improving clustering performance – are investigated in Section 6.2.

[1] $O(n^2 \log(n))$ for agglomerative and $O(2^n)$ for divisive clustering [105].

[2] Other criteria, such as Ward's criterion and Minimum energy clustering are left out of this analysis but might be of interest in future research.

principal component analysis (PCA)

Principal component analysis (PCA) is a procedure that uses orthogonal transformation and projections in order to convert a dataset $\mathbf{X} \in \mathbb{R}^{n \times m}$ of n observations in m dimensions, into a set of linearly uncorrelated variables called the principal components. This is done in an iterative way, starting with the first principal component, such that each added variable (component) preserves the largest amount of variance possible. It turns out that the transformation matrix $\mathbf{P} \in \mathbb{R}^{m \times m}$, such that:

$$\mathbf{Y} = \mathbf{X}\mathbf{P}, \tag{6.13}$$

gives the desired transformation, is given by a matrix $\mathbf{P} = (\mathbf{p}_1, ..., \mathbf{p}_m)$ such that $\mathbf{p}_j$ is the eigenvector of the covariance matrix $\mathbf{C} = \mathbf{X}^T\mathbf{X}$, corresponding to the j-th largest eigenvalue λ_j, thus:

$$\mathbf{C} = \mathbf{X}^T\mathbf{X} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}, \tag{6.14}$$

with $\mathbf{\Lambda}$ the diagonal matrix of ordered eigenvalues of $\mathbf{C}$.

To use PCA as a procedure to reduce the dimensionality of the data from m to l, the matrix $\mathbf{P}_l = (\mathbf{p}_1, ..., \mathbf{p}_l)$ is used to transform the data:

$$\mathbf{Y}_l^{PCA} = \mathbf{X}\mathbf{P}_l, \tag{6.15}$$

with transformed data $\mathbf{Y}_l^{PCA} \in \mathbb{R}^{n \times l}$.

t-distributed stochastic neighbour embedding (t-SNE)

As will be shown in the next section, PCA does not always give satisfying results. This is especially the case when the original data contains extreme values or skewed distributions. In such cases, a non-linear dimensionality reduction can bring relief. The chosen method of this kind that does give more satisfying results is t-distributed stochastic neighbour embedding (t-SNE).

Given dataset $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T \in \mathbb{R}^{n \times m}$ and desired reduced dimension l, the approach of t-SNE to find mapping:

$$T_{tsne} : \mathbb{R}^m \rightarrow \mathbb{R}^l, \tag{6.16}$$

such that:

$$T_{tsne}(\mathbf{X}) = \mathbf{Y}_l^{tSNE}, \tag{6.17}$$

$$T_{tsne}(\mathbf{x}_i) = \mathbf{y}_i, \tag{6.18}$$

with $\mathbf{y}_i \in \mathbb{R}^l$, is the following [107]:

1. Compute probability distribution matrix P, with $P_{ij} = \mathrm{p}_{ij}$ proportional to the distance between two observations $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$\mathrm{p}_{j|i} = \frac{\exp\left(\frac{-D(\mathbf{x}_i,\mathbf{x}_j)^2}{2\sigma_i^2}\right)}{\sum_{k \neq l} \exp\left(\frac{-D(\mathbf{x}_l,\mathbf{x}_k)^2}{2\sigma_i^2}\right)}, \tag{6.19}$$

   originally with $D(\cdot,\cdot)$ the Euclidean norm, but any distance measure can be used here. The parameter σ_i can be seen as the kernel bandwidth, chosen in such a way that the perplexity (see Equation 5.36) equals a predefined value, using the bisection method [107]. Now the probability p_{ij} is defined by symmetrising and normalising Equation 6.19:

$$\mathrm{p}_{ij} = \frac{\mathrm{p}_{j|i} + \mathrm{p}_{i|j}}{2N}. \tag{6.20}$$

2. Now the transformed variables $\mathbf{y}_1, ..., \mathbf{y}_n$ need to reflect these similarities as good as possible. In order to achieve this, probabilties q_{ij} are constructed similar to the above:

$$\mathrm{q}_{ij} = \frac{\left(1 + D(\mathbf{y}_i, \mathbf{y}_j)^2\right)^{-1}}{\sum_{k \neq l} \left(1 + D(\mathbf{y}_k, \mathbf{y}_l)^2\right)^{-1}}, \tag{6.21}$$

   with $\mathbf{y}_j$ still unknown. Instead of the Gaussian density function used in 6.19, a Cauchy distribution (Student-t distribution with one degree of freedom) is used here as a measure of similarity. The fat tail of this distribution allows dissimilar objects to be placed far away in the transformed space.

3. Finally, $\mathbf{y}_1, ..., \mathbf{y}_n$ are chosen in such a way that the KL-divergence between distributions P and Q is minimised, with:

$$\underset{(\mathbf{y}_1,...,\mathbf{y}_n) \subset \mathbb{R}^l}{\operatorname{argmin}} \; D_{KL}(P||Q), \tag{6.22}$$

   with:

$$D_{KL}(P||Q) = \sum_{i \neq j} \mathrm{p}_{ij} \log \frac{\mathrm{p}_{ij}}{\mathrm{q}_{ij}}. \tag{6.23}$$

   The minimisation of this divergence is performed using gradient descent.

Note that T_{tsne} operates on the observed data $\mathbf{X}$, but the actual transformation only depends on the distance $\{D(\mathbf{x_i}, \mathbf{x_j})\}$ between each pair of points . An alternative transformation T'_{tsne} could thus be applied to the distance matrix Δ, achieving the same results: $T'_{tsne}(\Delta) = T_{tsne}(\mathbf{X}) = \mathbf{Y}_l^{tSNE}$.

As is shown in the next Section, the dimensionality reduction with t-SNE applied to the dataset $\boldsymbol{\theta}$ results in a representation that shows more structure than the representation in the first 2 principal components.

6.2 Model choices and implementation

Resulting from the previous section, four main model choices are now required:

- the distance measure used for clustering;
- the hierarchical clustering linkage method;
- the number of clusters K to be retrieved;
- the dimensionality reduction procedure.

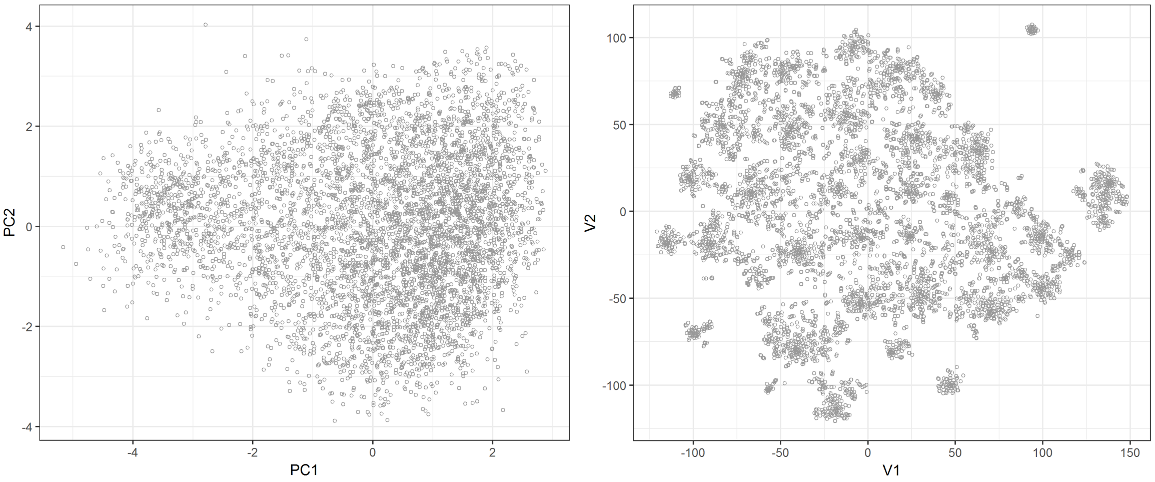
Furthermore, a clustering is performed on both the distance matrix Δ and the transformed data $T'_{tsne}(\Delta)$ and the results are compared.

6.2.1 Distance measure

The first choice that needs to be made is the distance measure from Section 6.1.2. Using the symmetric Kullback-Leibler divergence, $D_{SKL}(\cdot, \cdot)$, gave visually the best results when combined with the t-SNE dimensionality reduction, of which the result is presented in Figure 6.1b. “Visually best results’ is subjective: the 2D-representation shows a scattering of observations that is less uniformly spread out over the space, and thus likely to be more suitable for clustering. More objective criteria for this choice could be used when more context data is available. For example, the distance measure that gives the best prediction of some known lifestyle features could be used. But the choice of the symmetric Kullback-Leibler divergence also makes sense from a more fundamental point of view: The KL-divergence is defined specifically to measure similarities of two probability distributions[103].

6.2.2 Dimensionality reduction

Now the two proposed dimensionality reductions are applied to observed data $\mathbf{\Theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N)^T$. $Y_2^{PCA} = \mathbf{\Theta P}_2$ is visualised in Figure 6.1a, and shows a representation of the data that is somewhat uniformly distributed over this space. In Figure 6.1b, $Y_2^{tSNE} = T'_{tsne}(\Delta)$ is visualised. This representation of the data shows more structure: several clusters of points can be observed. Therefore, the latter of the two seems more suitable for a clustering of the observations. Thus t-SNE is used in further analysis, both for visualisation and clustering improvement.



(a) PCA *(b) t-SNE*

Figure 6.1: Transformation of dataset **X** *with PCA (a) and t-SNE (b) visualized in two dimensions. The transformed data shows more structure in the t-SNE transformation than in the PCA transformation and is therefore chosen for further analysis.*

6.2.3 Hierarchical clustering method and parameters

In order to decide on which of the four agglomeration methods mentioned in Equations (6.9) - (6.12) should be used, and to determine the number of clusters K that need to be identified, three clustering evaluation metrics are considered. For a given clustering $C = \{C_1, ..., C_K\}$, these metrics are given below.

- **Silhouette Width.** The Silhouette Width is the average of each observation's Silhouette value $S(i)$, which is a metric for the degree of confidence in the clustering assigment of a particular observation i [81]:

$$Silhouette(C) = \frac{1}{N}\sum_{i=1}^{N} S(i) = \frac{1}{N}\sum_{i=1}^{N}\frac{b_i - a_i}{\max(b_i, a_i)}, \tag{6.24}$$

with:

$$a_i = \frac{1}{|C(i)|}\sum_{j \in C(i)} d(x_i, x_j), \tag{6.25}$$

where $C(i)$ is the cluster to which observation i is assigned. Thus a_i is the average distance from i to each point in its cluster. b_i is defined as the average distance of observation i to the nearest neighbouring cluster:

$$b_i = \min_{C_k \in C \setminus C(i)} \frac{1}{|C_k|}\sum_{j \in C_k} d(x_i, x_j). \tag{6.26}$$

$Silhouette(C)$ lies in the interval $[-1, 1]$ and should be maximised.

- **Dunn Index.** The Dunn Index is the ratio of the smallest distance of two observations in different clusters with the largest distance of two observations within one cluster [82] :

$$Dunn(C) = \frac{\min\limits_{C_k, C_l \in C} \left(\min\limits_{i \in C_k, j \in C_l} d((x_i, x_j) \right)}{\max\limits_{C_k \in C} \left(\max\limits_{i,j \in C_k} d(x_i, x_j) \right)}. \tag{6.27}$$

  $Dunn(C)$ lies within the interval $[0, \infty)$ and should be maximised.

- **Connectivity.** The Connectivity measures to what extent observations are placed in the same cluster as their nearest neighbours. It is defined as [108]:

$$Connectivity(C) = \sum_{i=1}^{N} \sum_{j=1}^{L} \delta_{i,nn_i(j)}, \tag{6.28}$$

  with:

$$\delta_{i,nn_i(j)} = \begin{cases} 0 \text{ if } C(i) = C(j) \\ \frac{1}{j} \text{ if } C(i) \neq C(j) \end{cases}, \tag{6.29}$$

  and L a chosen parameter.

  $Connectivity(C)$ lies within the interval $[0, \infty)$ and should be minimised.

Clustering distance matrix Δ

Each of the four mentioned linkage procedures (single, complete, average and centroid) in Equations (6.9)- (6.12) are now performed on the distance matrix Δ. For all numbers of clusters $K = 2...100$, each of the three metrics in Equations (6.24), (6.27) and (6.28) are computed. These values are plotted in Figure 6.2. The optima for each validation metric can be found in Table 6.1.

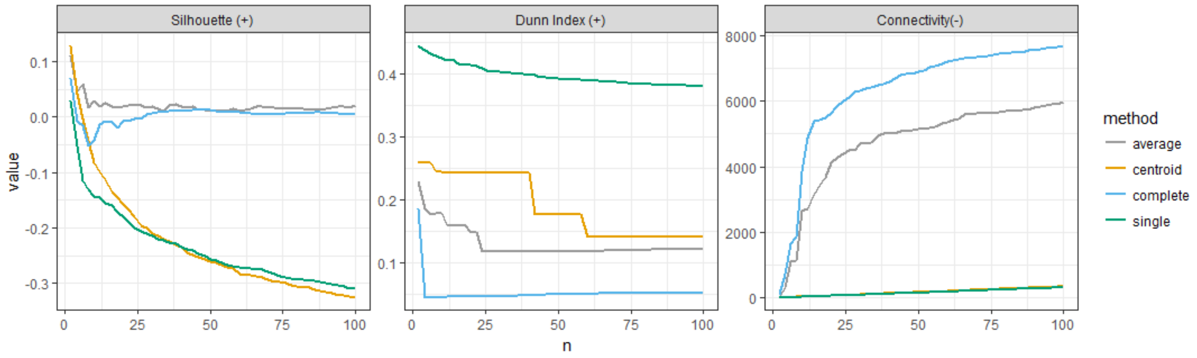


Figure 6.2: Three evaluation metrics for 4 different hierarchical clustering methods. All three evaluation metrics suggest that 2 clusters is optimal, but are not in agreement about the optimal linkage method.

Table 6.1: Optimised evaluation metrics for clustering of $\mathbf{\Theta}$.

	Value	Linkage method	K
Silhouette Width	0.129	centroid	2
Dunn Index	0.445	single	2
Connectivity	3.029	single	2

These results suggest that each of the procedures works best with $K = 2$. The Silhouette Width motivates the use of Centroid Linkage, the Dunn Index motivates Single Linkage and the Connectivity is indecisive between Centroid and Single Linkage.

Clustering applied after t-SNE transformation

Although suggested by the clustering evaluation metrics, $K = 2$ is not a preferred outcome as the number of consumer groups: if only two clusters are chosen, then there are only two consumer types inferred, which is not what is needed for granular customer segmentation. Therefore, the same procedure is repeated on the transformed data $\mathbf{Y}_2^{tSNE}$, with $\mathbf{y}_i \in \mathbb{R}^2$. The results are shown in Figure 6.3 and the optimised values can be found in Table 6.2.

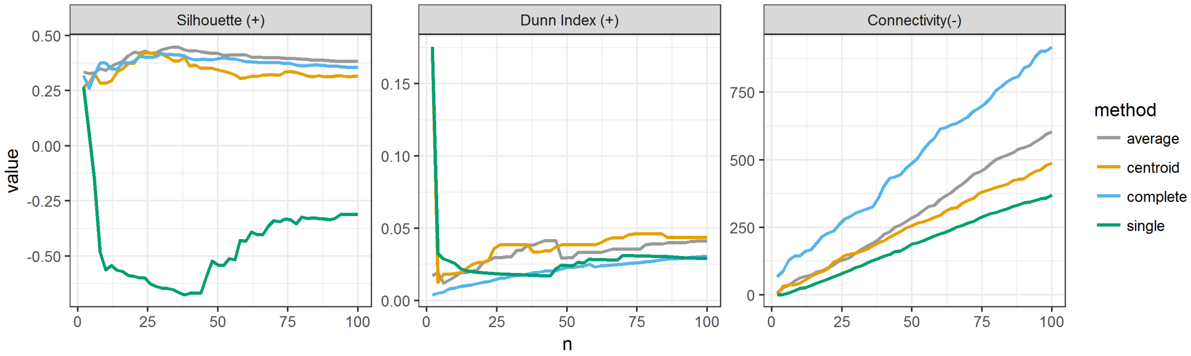


Figure 6.3: Evaluation metrics now applied to the transformed data. Silhouette Index suggests average linkage with $K = 34$ clusters.

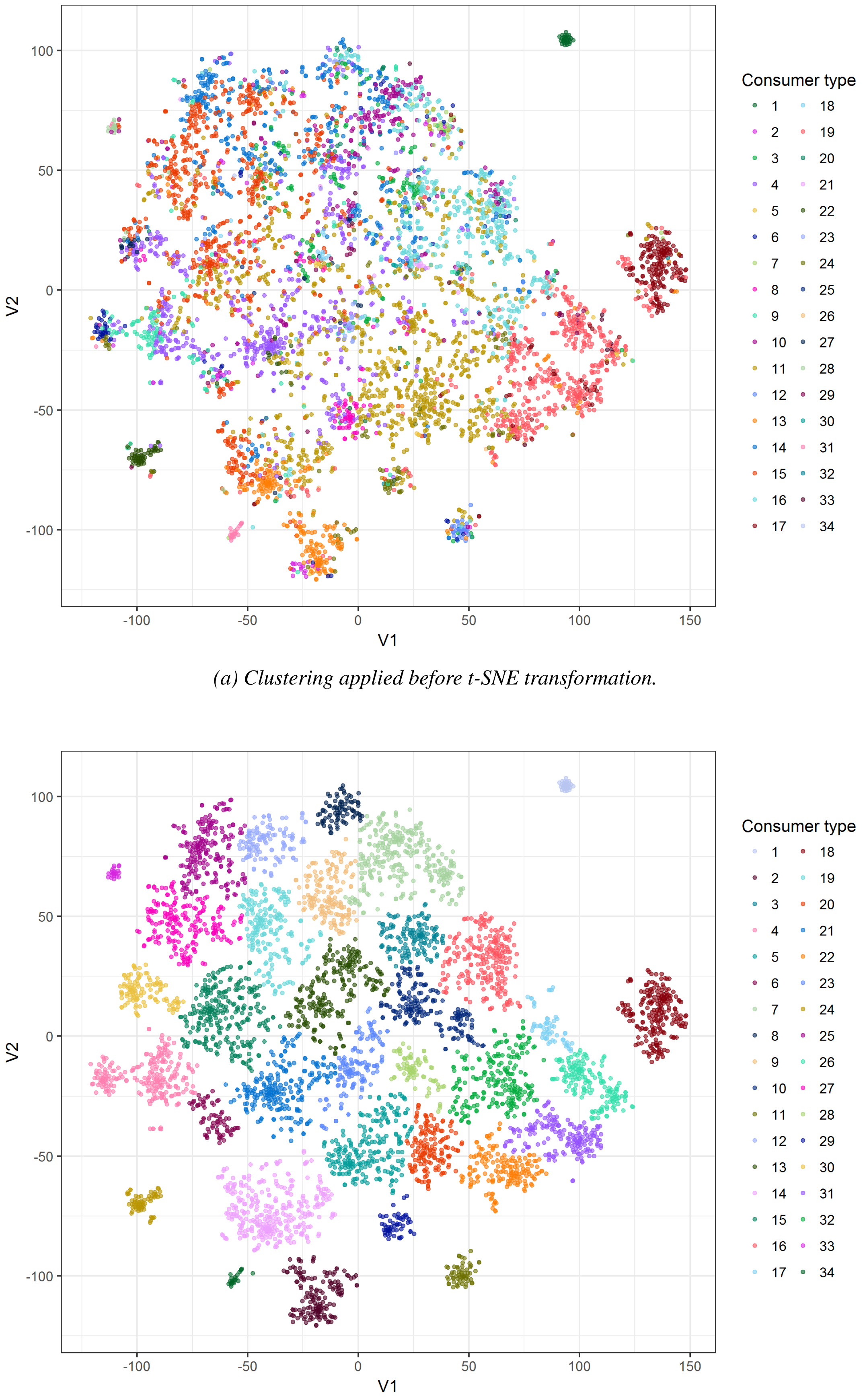
Table 6.2: optimised evaluation metrics for clustering of $\mathbf{Y}_2^{tSNE} = T_{tSNE}(\mathbf{\Theta})$.

	Value	Linkage method	K
Silhouette Width	0.449	average	34
Dunn Index	0.176	centroid	2
Connectivity	0	centroid	2

Now the Dunn Index and Connectivity still give an optimum of $K = 2$ clusters, but the Silhouette Width suggests to use Averege Linkage with $K = 34$ clusters.

6.3 Results

Hierarchical clustering with Average Linkage and $K = 34$ is now performed to both the distance matrix Δ and transformed data $\mathbf{Y}_2^{tSNE} = T'_{tSNE}(\Delta)$ for comparison. The results of this clustering are presented in Figure 6.4. The results are visually very different, with Figure 6.4b showing circular clusters all of approximately equal size. Figure 6.4a shows some large clusters and various smaller ones. Although observations assigned to the same cluster are more scatter than when clustering is applied after t-SNE transformation, they are still relatively close to each other.



(a) Clustering applied before t-SNE transformation.

(b) Clustering applied after t-SNE transformation.

Figure 6.4: Result of hierarchical clustering with Average Linkage on all 5,194 consumers. 34 clusters are identified. Clustering is applied both before (a) and after (b) a t-SNE transformation to 2 dimensions.

Information about cluster size, average consumption, and the characteristic lifestyles of the the retrieved clusters can be found in Appendix F. The characteristic lifestyles $\mathcal{T}_\psi(i)$ of a consumer cluster i are defined as the lifestyles that have high average probability in this cluster, relative to the average probability of these lifestyles over all observations. Thus:

$$\mathcal{T}_\psi(i) = \left\{\tau_j : \bar{\mathrm{p}}_i(\tau_j) > \psi \cdot \bar{\mathrm{p}}_{all}(\tau_j)\right\}, \tag{6.30}$$

with:

$$\bar{\mathrm{p}}_i(\tau_j) = \frac{1}{|C_j|}\sum_{l\in C_j} \mathrm{p}(\tau_j|l) = \frac{1}{|C_j|}\sum_{l\in C_j} \boldsymbol{\theta}_{l,i}, \tag{6.31}$$

the average estimated probability of lifestyle j for all customers in cluster i, and:

$$\bar{\mathrm{p}}_{all}(\tau_j) = \frac{1}{5194}\sum_{l=1}^{5194} \mathrm{p}(\tau_j|l) = \frac{1}{5194}\sum_{l=1}^{5194} \boldsymbol{\theta}_{l,i}, \tag{6.32}$$

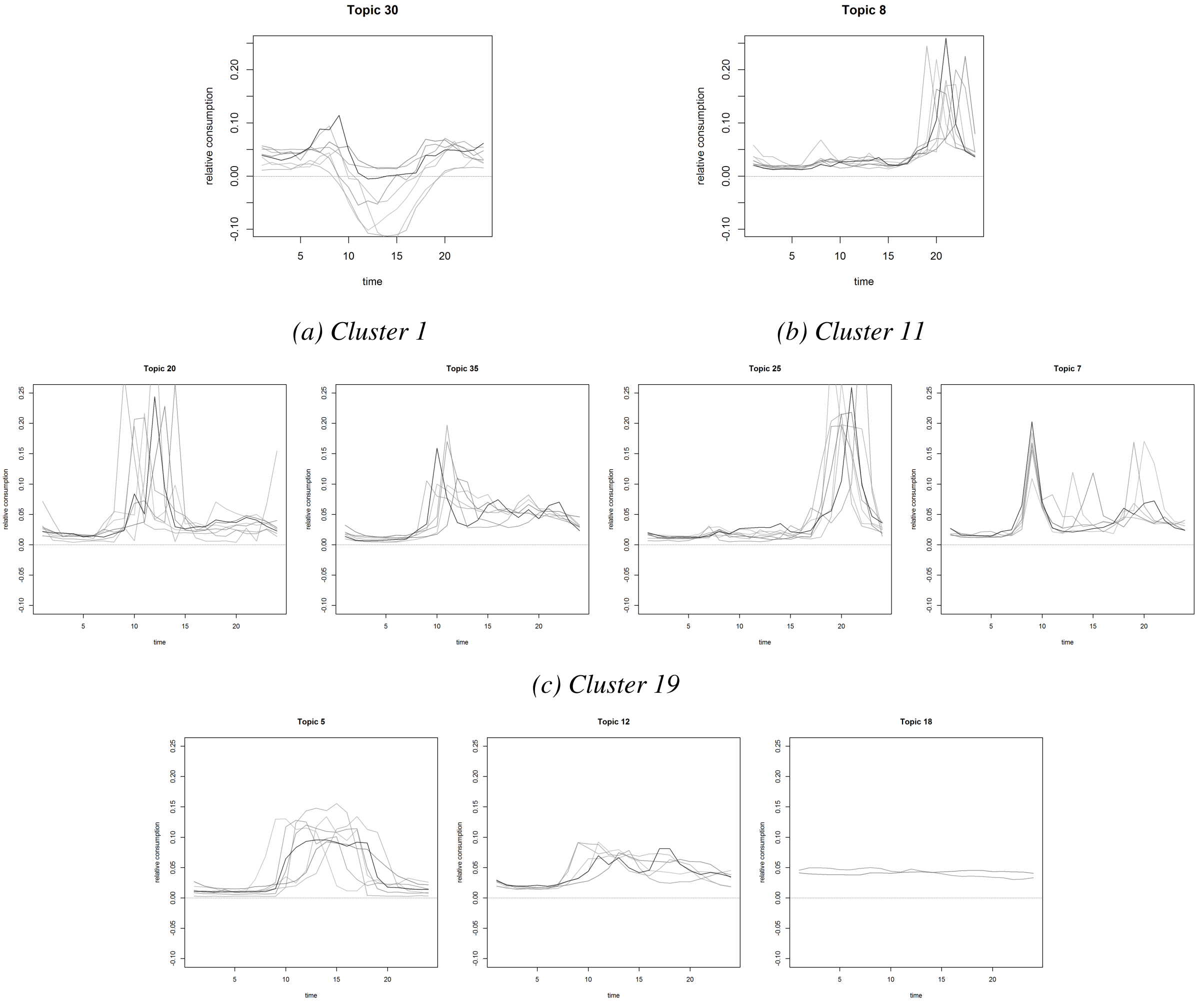
the average estimated probability of lifestyle j for all customers.

Four customer clusters, resulting from hierarchical clustering applied on distance matrix Δ, are briefly investigated further. The characteristic lifestyle of these clusters are visualized in Figure 6.5 and more information can be found in Table 6.3. The total number of observations assigned to cluster i is denoted by $|C_i|$, while the number of all-electric homes in the cluster is denoted by $|C_i^{\mathrm{NL15ae}}|$.

Table 6.3: Summary statistics of several customer clusters, with clustering applied before t-SNE transformation.

i	$\|C_i\|$	$\|C_i^{\mathrm{NL15ae}}\|$	μ_{kwh}	σ_{kwh}	$\mathcal{T}_3(i)$	$\bar{\mathrm{p}}_i(\tau_j)$	$\bar{\mathrm{p}}_i(\tau_j)/\bar{\mathrm{p}}_{all}(\tau_j)$
1	32	27	14.6	13.1	30	0.67	132.7
11	931	1	6.6	4.2	8	0.11	3.5
19	506	0	3.7	4.1	20	0.15	6.3
					35	0.1	6.1
					25	0.12	4.7
					7	0.07	3.7
31	23	0	25.5	24.1	5	0.49	69
					12	0.11	10.3
					18	0.25	4

The biggest cluster is cluster 11, consisting of 931 consumers, almost 20% of all consumers in the dataset. The characteristic load shape shows typical residential consumer patterns with morning and/or evening peaks. Cluster 19 also has typical residential consumption, but a very low average consumption. Cluster 1 has high presence of all-electric consumers and has on average 67% typical PV lifestyles. Cluster 31 is characterized by high average consumption and lifestyles with a high mid-day stable consumption in 50% of the time and a lifestyle with flat consumption patterns in 25% of the time. These are likely to be commercial consumers.



(a) Cluster 1

(b) Cluster 11

(c) Cluster 19

(d) Cluster 31

Figure 6.5: The characteristic lifestyles $\mathcal{T}_3$ for 4 customer clusters, derived with hierarchical clustering applied to Δ. Cluster 11 and 18 show typical residential lifestyles. Cluster 1 is likely to have solar panels, while cluster 31 seems to belong to commercial consumers.

6.4 Conclusions

In this chapter, the estimated lifestyle distributions $\hat{\boldsymbol{\theta}}_m$ per consumer from Chapter 4 were used as a basis for consumer clustering. Symmetric Kullback-Leibler divergence was used as a distance metric between these probability distributions. After comparing various clustering algorithms, hierarchical clustering with average linkage was chosen as the clustering procedure. The clustering was performed twice: once before and once after a transformation to 2 dimensions with t-distributed stochastic neighbour embedding (t-SNE). 34 clusters were identified in both cases. These two approaches led to visually different results, with the latter having more homogeneous cluster sizes. However, a distinction between commercial consumers and several residential types could be seen in both results. Also, the algorithm clustered together a large part of the all-electric consumers in both approaches.

In the next chapter, the developed methodologies from this and the previous chapters are applied to answer several business related questions from DSO Stedin. A validation of the model is also part of this.

Chapter 7

Use-cases for a DSO

In Chapter 3, the energy consumption of all-electric homes was compared with energy consumption data that contained many different consumers. From this analysis it followed that all-electric homes have higher base load, higher peak demand and more simultaneous peaks than homes from the conventional dataset. These three effects add up to a much higher aggregated peak load. The follow-up question was how these factors evolve when different mixes of consumer types are considered. In Chapters 4, 5 and 6, a framework was constructed to identify energy lifestyles from smart meter data, in order to make a clustering of consumers based on their characteristic lifestyles. In this chapter these results are combined to answer the posed question, which links to the main research question: how do aggregated peak load and simultaneity evolve with increasing penetration of all-electric homes?

First, some relevant questions related to the business of Stedin are posed. In the next section some results from the consumer segmentation are presented. In Section 7.2, the consumer segmentation results from the previous chapter are validated based on available context data from the all-electric dataset. Then the simultaneity of different identified consumer types is investigated. In Section 7.3 various consumer mixes are simulated and compared based on simultaneity and aggregated peak demand. Lastly, this chapter ends with some concluding remarks.

7.1 Business questions related to load shape profiles

During the process of developing the model in the previous chapters, various questions relating to the business and operations of Stedin arose:

- What consumer segments can be identified from the available energy consumption data? Can, for example, a clear segmentation of residential and commercial consumers be made?
- How well does the model segment the all-electric homes from the conventional dataset? Can it detect other all-electric homes or buildings with solar panels?
- How do different consumer types compare in aggregated peak load and simultaneity?
- What happens to the aggregated loads, both peak and simultaneity, when different mixtures of consumer types are considered? Does the simultaneity increase/decrease linearly or not?

These questions will be addressed in the remainder of this chapter.

7.2 Identifying commercial consumers and PV presence

This section evaluates how successful the developed method is in identifying four special occupancy segments:

- all-electric homes with PV;
- all-electric homes without PV;
- conventional homes with PV;
- commercial energy consumption.

The first two segments are also functioning as a validation, since this is the only context information available. Of course, many other occupancy segments could be considered, such as *family with young kids* or *retirees*, or homes with heat pumps and electric cars. Validation of this would require more context data. Therefore this analysis is left for future research.

7.2.1 Validating the clustering of all-electric homes

The information available for the NL15ae-dataset is now used for validating the developed model. As mentioned in Section 3.1, the all-electric homes in the dataset were part of a pilot study with new homes, in which each home owner had chosen one out of four propositions, related to the presence of PV and batteries. These are given in Table 7.1.

Table 7.1: Size and PV presence per all-electric proposition.

Proposition	*n*	PV presence	Battery presence
A	24	Yes	Yes
B	3	Yes	No
C	8	No	Yes
D	7	No	No

Some results per proposition can be found in Appendix G. In the remainder of this section only PV presence is considered, identifying battery presence will be left for future analysis. In Figure 7.1 the all-electric homes are plotted in the t-SNE transformed space, coloured by the presence of solar PV. The PV homes (proposition A and B) are all located together in the upper right cluster, while the homes without PV (proposition C and D) are much more scattered. This is summarised in Table 7.2: all 27 homes with PV are in cluster 1, while the 15 non-PV homes are spread among 9 other clusters.

Table 7.2: PV identification of all-electric dataset.

Cluster	Total	no PV (C&D)	PV (A&B)
1	27	0	27
2	1	1	0
3	2	2	0
4	2	2	0
5	2	2	0
6	1	1	0
7	4	4	0
8	1	1	0
9	1	1	0
10	1	1	0

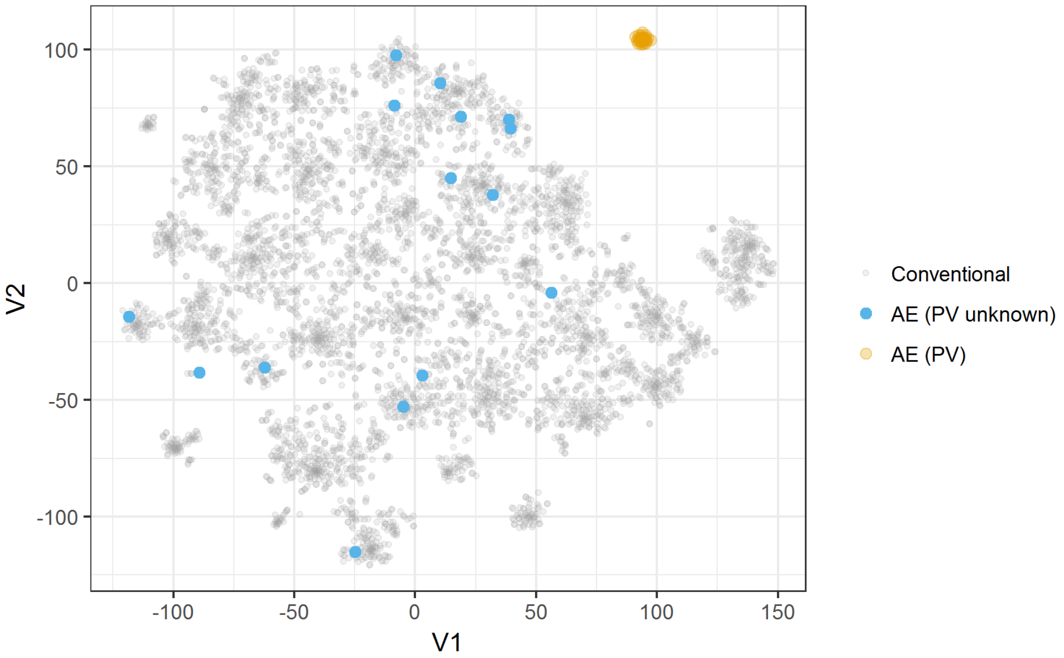


Figure 7.1: Validation of the method using the NL15ae-dataset. The representation shows a cluster of PV-homes in the upper right corner. The non-PV all-electric homes are scattered.

These results are partly satisfying: it shows that the algorithm successfully clusters together all-electric homes that have PV, while the all-electric homes without PV end up in various different clusters. The explanation of the latter probably comes from the remark at the end of Section 4.3.2, which states that the load shape dictionary was constructed based on a sample containing only a very small number (±30) of daily loads from all-electric homes without PV. Therefore it is likely that only a very limited amount of typical consumption of these homes is represented well by any of the 633 constructed load shapes. Thus, when all data is assigned to their nearest load shape, the daily loads of all-electric homes are matched with consumption patterns from conventional homes, possibly with large errors. This results then in LDA not being able to identify the characteristic *all-electric-no-PV*-lifestyle(s). Generating a new dictionary based on a larger sample, and using stratified sampling to compensate for under-representation of all-electric homes, is recommended for future research.

7.2.2 Identifying PV presence in the conventional dataset

Now it is shown that the algorithm is successful in detecting all-electric homes with PV, one might wonder if it can also indicate the presence of PV in the NL14conv-dataset. This is a challenge since:

- no information is available about the presence of PV in this dataset;
- the NL14conv-dataset does not contain negative consumption values, which might result from the way the data is collected or processed previous to this research;
- it is possible that no homes with PV are present in the dataset.

In the previous section consumer cluster 1 was identified as the all-electric PV cluster. However, there are also 5 buildings from the NL14conv-dataset assigned to this cluster. These are given in Table 7.3. This table shows that these 5 consumers all have lifestyle 30 as their most dominant one, with consumer 1573 with even 73% of its load shapes originating from this lifestyle. Lifestyle 30 is shown in Figure 6.5a and shows typical PV consumption load shapes.

These result suggest that this method identified at least one consumer with PV and potentially 4 more.

Table 7.3: Consumers in segment 1 that do not originate from the NL15ae-dataset. The underlined values are the lifestyle and respective probabilities that correspond with a PV lifestyle.

ID	μ_{kwh}	$\mathcal{T}_5(i)$	$\{\text{p}_i(\tau_j) : \tau_j \in \mathcal{T}_5(i)\}$	$\{\text{p}_i(\tau_j)/\text{p}_{all}(\tau_j) : \tau_j \in \mathcal{T}_5(i)\}$
291	16.3	$\underline{30}$, 26, 23, 8	$\underline{0.15}$, 0.28, 0.11, 0.17	$\underline{29.8}$, 12.8, 6.2, 5.7
1422	16.0	$\underline{30}$, 34, 4	$\underline{0.07}$, 0.15, 0.1	$\underline{14.5}$, 6.6, 5.4
1573	22.9	$\underline{30}$	$\underline{0.73}$	$\underline{146.4}$
2502	17.8	$\underline{30}$, 14, 21, 28	$\underline{0.08}$, 0.2, 0.16, 0.13	$\underline{15.4}$, 10.1, 7.4, 5.2
3904	23.7	$\underline{30}$, 26, 17	$\underline{0.15}$, 0.26, 0.06	$\underline{30.6}$, 11.9, 5.5

7.2.3 Segmenting commercial from residential consumers

Since no information is available about the actual occupancy of the consumers in the NL14conv-dataset, it is not possible to test to what extent the developed method can distinguish residential consumers from commercial consumers. However, the characteristic lifestyles from cluster 31 in Figure 6.5d do show consumption patterns that are considered to be typical for commercial applications. Both visualised clustering methodologies have such a cluster, which is isolated in the upper left corner in the t-SNE transformed dataset (see Figure 6.4).

7.3 Aggregation of mixed consumer types

In this section, the effect of the local energy transition is simulated. This is done based on the original energy consumption time series data $\{\mathbf{P}^{(h)}\}$ and the consumer clusters C identified in the previous chapters. Various cluster mixtures are simulated and the development of their simultaneity and aggregated peak loads are assessed.

7.3.1 Sampling procedure

The algorithm is similar to the one explained in Algorithm 1 in Chapter 3. Here the total number of sampled homes N remains constant, but different mixed compositions are considered. Given are:

- clustering $C = \{C_1, ..., C_k\}$;
- observations $\mathbf{P} = \{\mathbf{P}^{(h)}\}_{h=1}^{5194}$;
- the chosen total sample size N;
- the chosen number of samples per step N_{rep}.

For two consumer clusters $u, v \in \{1, ..., k\}$, the following steps are performed:

1. Set initial stratification $(n_u, n_v) = (N, 0)$.
2. Draw stratified samples $S = s_u \cup s_v$, with every $s_i \subset C_i$ a random sample of size n_i from the i^{th} cluster.
3. Calculate the aggregated peak load and simultaneity over sample S like in Equation 3.1:

$$a_S = \max_{t \in T} \sum_{h \in S} P_t^{(h)}, \qquad \textit{(Peak of aggregate)} \qquad (7.1)$$

$$b_S = \sum_{h \in S} \max_{t \in T} P_t^{(h)}, \qquad \textit{(Sum of individual peaks)} \qquad (7.2)$$

$$\gamma_S = \frac{a_S}{b_S}. \qquad \textit{(Simultaneity)} \qquad (7.3)$$

4. Repeat step (2) and (3) N_{rep} times and calculate the mean, standard deviation and quantiles of simultaneity and aggregated peak load.
5. Set $(n_u, n_v) = (n_u - 1, n_v + 1)$ and repeat steps (2)-(4) until $n_u = 0$ and $n_v = N$.

A sampling procedure for $l > 2$ clusters can also be considered, with stratifications $(n_{u_1}, ..., n_{u_l})$, $\sum_{j=1}^{l} n_{u_j} = N$ at every step.

7.3.2 Selected consumer clusters and implementation

The two consumer clusters that are selected for this analysis are the large residential cluster 11 and the all-electric cluster 1 from Section 6.3 and Figure 6.5. The total number of homes simulated (N) is 40 and the sample size per mixture composition is $N_{rep} = 100$.

7.3.3 Results

This simulation is performed on the mentioned consumer clusters. The results are visualised in Figure 7.2. Table 7.4 shows the average simultaneity ($\overline{\gamma}$), average aggregated peak demand ($\overline{a}$) and average sum of individual peaks ($\overline{b}$) for several different compositions. Also, the average effective peak per consumer ($\frac{1}{40}\overline{a}$) and the average individual peak ($\frac{1}{40}\overline{b}$) are given.

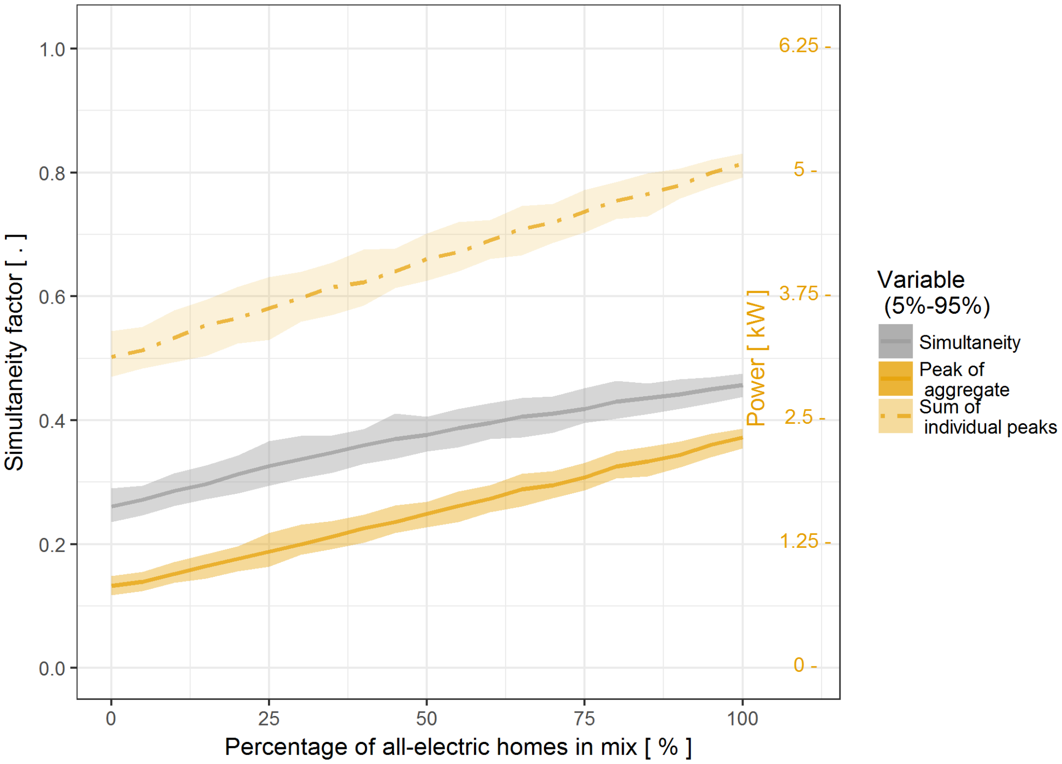


Figure 7.2: Simulation of effect of increased penetration of all-electric homes (with PV) in a community of conventional homes. Both the simultaneity and the aggregated peak increase linear with the penetration ratio.

These results show a linear increase in both simultaneity and aggregated peak load. The simultaneity almost doubles from 0.26 to 0.46, and the average effective peak per home increases from 0.8 to 2.3, almost tripling.

Table 7.4: Results of simulation visualized in Figure 7.2.

$\%C_1$ in mix	$\overline{\gamma}$	$\overline{a}$	$\overline{b}$	$\frac{1}{40}\overline{a}$	$\frac{1}{40}\overline{b}$
0	0.26	33.1	126.6	0.8	3.2
20	0.31	44.3	141.5	1.1	3.5
40	0.35	55.4	158.4	1.4	4.0
60	0.40	68.4	172.7	1.7	4.3
80	0.43	81.3	188.8	2.0	4.7
100	0.46	92.8	203.2	2.3	5.1

It is important to note that the increased simultaneity (+75%) and individual peak demand (+59%) combined lead to a sharp increase in the aggregated peak demand (+188%). These results should be combined with different local energy transition scenarios to assess the risk of deferring grid investments. Also, similar analysis should be done with other mixture compositions to see if this linearity holds in general.

7.4 Conclusions

In this chapter the results of Chapters 3 - 6 were combined to answer several business related questions of Stedin related to the monitoring of the local energy transition and the simulation of its effect.

First, the customer segmentation proved to be successful in identifying all-electric homes with solar panels. Also, five consumers from the NL14conv-dataset were identified as potential homes with solar panels. The all-electric homes without solar did not end up in one single cluster, which is likely to be the result of the sampling procedure applied in Chapter 4. Also, a group of electricity consumers is identified that is likely to be non-residential, such as a small business or office.

In the second part of this chapter, a simulation study on the effect of the local energy transition was performed. This study showed that simultaneity and aggregated peak load both increase linearly in the number of all-electric homes with PV, respectively doubling and tripling.

To conclude, a model has been developed that can support DSOs like Stedin in making better grid investment decisions. An algorithm that finds lifestyles in smart meter data, and detects the presence of technologies like PV, enables monitoring of the status of the local energy transition. Additionally, the developed model can act as a basis for simulating the future impact of this transition on the distribution grid. These two capabilities combined provide new tools to the grid planner and asset manager to prepare the grid for a low-carbon future.

Chapter 8

Conclusion, Discussion & Future Work

In the previous chapters, a framework was developed to monitor the energy transition and simulate its future effects on the distribution grid. This model was applied to two datasets, and used to answer several business related questions of Stedin. This chapter combines the results and insights from the previous chapters to answer the research questions that were posed in the beginning of this thesis. Furthermore, the weaknesses of the developed method are discussed, and directions for both practical application and model improvement are recommended.

Section 8.1 revises the sub-research questions posed in Chapter 1, and uses these to answer the main research question of this thesis. In Section 8.2 the strengths and weaknesses of various aspects of the analysis are discussed. Section 8.3 gives an overview of the potential implications and applications of this work. The chapter ends with various directions for future improvements of the model.

8.1 Conclusion

The main objective of this research is to assess the impact of different technologies and transition scenarios on the aggregated peak load and simultaneity in the grid, based on real-world smart meter data. This section answers the sub-research questions stated in Section 1.3 and combines these insights to answer the main research question:

Can load shapes be used as a basis for simulating the effect of the local energy transition on the aggregated peak demand and simultaneity?

Answering the sub-research questions

SQ1. *What is the current state-of-the-art in the analysis and visualisation of Smart Meter Data?*

With the large-scale roll-out of smart metering infrastructure over the past years, an ever increasing amount of energy consumption data is coming available. This data is becoming an important resource for energy utilities and grid operators to solve the great challenges that the energy transition is posing to the grid. To leverage this value, many techniques for visualisation and analysis are being developed.

A part of this research was done at Stanford University where a variety of these methodologies for analysis and visualisation are being developed. VISDOM is an open-source software tool for visualising smart meter data of individual households. Its main purpose is to quickly visualise large amounts of features and data in order to generate and test hypotheses. Visualisations of the geographical distribution of features, variable-to-variable plots and characteristic load shapes are included in this package. The open-source character of VISDOM makes it adaptable for own purposes. Another software package that is being developed at Stanford is VADER, a tool for grid analysis with a focus on monitoring distributed energy resources.

Various advanced analytical procedures to gain more insight in the way people and buildings are consuming energy are available. Load shape clustering is a methodology that identifies consumption patterns that summarise how electricity is consumed on a daily basis. The literature on methodologies for this is rich, and a selection of the proposed methods are listed in Chapter 2. In several cases, load shapes are used as a basis for consumer segmentation. This thesis proposes an expansion on this method by adding the concept of a *lifestyle*. Consumers are assumed to be represented by a mixture of lifestyles, with each lifestyle defined by a set of load shapes. This extra layer in the model allows for more flexible and interpretable characterization of consumers.

SQ2. *What are the biggest differences in energy consumption statistics between conventional and all-electric communities?*

Two datasets are used in this analysis. The NL14conv-dataset is a relatively large dataset of residential and commercial consumers, geographically spread over the Netherlands. The NL15ae-dataset results of a pilot study of 42 all-electric homes, some of which are equipped with solar panels. In Chapter 3, these two datasets are compared with each other and a selection of Stedin's consumers. Statistics about consumption and peaks were inferred from the data. In order to calculate the simultaneity (a measure for peak synchronisation of a set of consumers) and effective peak (the maximum of the aggregated load divided by the number of homes), simulations with different sample sizes were performed. The results are summarised in Table 8.1.

Table 8.1: Energy consumption average statistics per dataset.

		Annual consumption	Daily 1h-peak	Annual 1h-peak	Simultaneity[1]	Effective peak[1]
Dataset	n	$[MWh]$	$[KWh]$	$[KWh]$	$[\cdot]$	$[KWh]$
NL14conv	$4,735$	3.46	1.16	3.12	0.31	0.97
NL15ae	41	5.91	1.82	4.99	0.49	2.43

[1] Based on simulations of 40 homes

Simultaneity and effective peak were calculated for different sample sizes. This resulted in a relationship between sample size and simultaneity. For both datasets, this simultaneity factor converged for more than 30 homes.

Summarising, the answer to this sub-question is that in comparison with conventional homes, all-electric homes have all of the following:

- 74% higher average consumption;
- 60% higher annual peaks;
- 58% increased simultaneity;
- 151% increased effective peaks.

Thus, because both base and peak load are increasing, and peaks are occurring more synchronised, the effective peak in all-electric communities more than doubles.

SQ3. *How can load shapes be used to summarise the variability in daily consumption patterns in the dataset?*

In Chapter 4, adaptive K-means was deployed as a method to effectively summarise the variability in normalised daily loads. Adaptive K-means allows the modeller to set an upper bound on the compactness of each identified cluster, without the need to determine the number of clusters beforehand. Furthermore, adaptive K-means can analyse data sequentially, meaning that it handles new data in a natural and effective way. Applying this algorithm to a sample of 0,6% of the data, resulted in a dictionary of 633 load shapes, summarising 93% of 1,8 million daily loads in the dataset within specified limits. Analysing the errors made by replacing the observed data by the assigned load shape gives that the maximum error is in the order of 1% of the daily peak, thus the load shapes approximate the actual load well.

Thus, a dictionary of load shapes generated with adaptive K-means summarises the variability in daily consumption patterns in the dataset.

SQ4. *Can load shapes serve as a basis for lifestyle based consumer segmentation?*

Answering this question is done in three steps:

1. model energy consumption as originating from a mix of lifestyles per consumer (Chapter 5);
2. fit this model to the series of load shapes per consumer (Chapter 5);
3. perform a consumer segmentation based on the fitted lifestyle-per-home distributions (Chapter 6).

First, in Chapter 5, latent Dirichlet allocation (LDA) was introduced as an advanced statistical model to find patterns in large discrete datasets. Variational Expectation Maximisation (VEM) was used as an approximation method to fit this model to the data. The input data were series of load shapes per home, with each load shape one of the 633 elements from the load shape dictionary that was constructed with Adaptive K-means. After evaluating various metrics, a total of 40 lifestyles was fitted to the data. This resulted in the following estimated distributions:

- a loadshape-per-lifestyle distribution for each of the 40 lifestyles;
- a lifestyle-per-consumer distribution for each of the 5194 consumers in the dataset.

For each identified lifestyle, the characteristic load shapes were all very similar in shape. This is remarkable since the data that was used as input consisted of categorical variables, and thus the algorithm did not incorporate the numerical similarities between load shapes. This observation supports the hypothesis that the identified structure can be interpreted as an actual lifestyle. Several typical residential lifestyles could be observed. Additionally, one of the clusters seemed to show typical commercial load shapes, while another one had load shapes that are typical for homes with solar panels.

In Chapter 6, a consumer clustering was performed based on the identified lifestyle-per-consumer distributions. As a clustering procedure, hierarchical clustering was performed on the calculated KL-divergence between each couple of vectors. Visualisation of the results in two dimensions was done with t-distributed stochastic neighbour embedding (t-SNE). This clustering technique turned out to be successful in identifying all-electric homes with solar panels. Five conventional homes were indicated likely to have solar panels. Furthermore, a cluster of typical commercial consumers was identified. The all-electric homes without PV were not clustered together, which is likely to be a consequence of the sampling procedure that was used to generate the load shape dictionary.

SQ5. *What is the effect of varying the consumer segment mix within a fixed number of homes on the aggregated peak load and simultaneity?*

In Chapter 7, a simulation study on the effect of local energy transition was performed. Different mixes of previously identified consumer segments were sampled after which aggregated peak load and simultaneity were computed. This resulted in the observation that both the expected aggregated peak demand and the simultaneity increase linearly in the fraction of all-electric homes with PV in the mix, respectively doubling and tripling.

Answering the main research question

Combining the previous insights, it can be concluded that the developed framework for lifestyle-based consumer segmentation is a promising method for clustering together homes and other electricity consumers that share similar energy consumption patterns. Adding these 'lifestyles' as a layer of abstraction in between homes and their electricity consumption adds interpretability of the identified segments. The method was successful in identifying homes with PV and buildings that are likely to be commercially occupied. Mixing several identified segments leads to valuable insight in the development of simultaneity and aggregated peak load as an effect of local energy transition scenarios.

Therefore the answer to the main research question is:

Yes, load shapes can be used as a basis for simulating the effect of local energy transition on the aggregated peak demand and simultaneity.

8.2 Discussion

Needless to say, the famous quote:

> *"Essentially, all models are wrong, but some are useful."* - George Box (1976)

is also applicable to the model developed in this thesis. Therefore, it is necessary to evaluate the assumptions and weaknesses for future improvement of this methodology. This assessment will be done in this section.

Data

The first critical point in this analysis is the data that was used. The quality of the NL14conv-dataset was not very high, showing the following: missing values, absent negative values, difficult-to-handle time formats, and confusing data anomalies. Cleaning the data did solve these issues, but a selection bias might have occurred there. The effect of imputing missing data should be evaluated in future research.

Another important question to ask is whether the used datasets were representative. The NL14conv-dataset originated from a Dutch Energy retailer focusing on a consumer segment that was both low-budget oriented and interested in installing smart meters before this was common practice. The NL15ae-dataset is a small dataset from a pilot study in which several different flexibility propositions were tested. Therefore, both datasets might not be a perfect representation of the groups of energy consumers that they are assumed to represent in the model.

Load shapes

In the generation of load shapes, several methodological choices are made that could potentially influence the analysis. Adaptive K-means avoids choosing the number of clusters prior to the clustering. However, it also results in splitting up clusters with very few observations if they do not meet the criteria, thus spending unnecessary time and increasing the dimensionality without significantly improving the outcomes. Furthermore, the choice of using the ∞-norm is justified by the fact that it poses an absolute bound on the maximum error. However, theoretically, this maximum error could be achieved over the whole load shape without splitting it, causing a very bad fit. Thus using a norm that makes a trade-off between the maximum error and the overall shape – like a (weighted) 2-norm – could improve upon this. However, improving the load shape generation algorithm is not expected to significantly improve overall performance, since the bulk of the data will always be represented by approximately the same load shapes.

A choice that is likely to have affected the outcomes of this analysis was the sampling of daily loads to generate the load shape dictionary, as was already discussed at the end of Section 4.3.2. The combination of the sample size and the fact that the sampling was performed uniformly, resulted in an underrepresentation of all-electric homes in the sample. Therefore, the generated dictionary of load shapes did not capture the energy consumption patterns of all-electric homes sufficiently well to identify all-electric homes without PV. Using a larger and stratified sample will probably solve this issue.

LDA

Some restrictive assumptions are made in the LDA model that should be loosened in future research. First, the bag-of-load-shapes assumption leaves out any time-dependency of the load shape probabilities. This is an oversimplification of reality, since certain lifestyles might be much more likely on a weekend than a weekday, or in summer than winter. Similarly, the weather dependency is not included in the model for now. Lifestyles are now assumed to be uncorrelated, while in reality, this might not be the case. Next, the number of identified lifestyles is currently fitted in an non-Bayesian way, thus not aligning with the rest of the model very well. This could be included in future models by putting a prior distribution on this number of topics. Also, the parameter vectors $\{\boldsymbol{\eta}\}$ are currently of length V, not assigning any probability to load shapes that are not yet included in the dictionary. Smoothed LDA is a expansion on LDA that puts a prior on the $\boldsymbol{\eta}$'s in order to solve this issue. Lastly, the approximation method – VEM – approximates the posterior mode, thus potentially leaving out valuable information about the posterior distribution, like the uncertainty in the estimated parameters.

Lifestyle interpretation

The terminology used speaks of lifestyles, as if they are totally unrelated to technologies or family sizes. This is in reality not the case. For example, the presence of PV might be related to a climate conscious lifestyle, but there is clearly a lot of technology involved too. Moreover electric heating, EV and flexibility solutions will likely be seen back in the results, as they are strongly technology focused. In order to make these identified lifestyles more interpretable as such, they should be correlated with data about the socio-economic status and present technologies.

Consumer clustering

Related to the previous paragraph about lifestyle interpretation, the interpretation of the consumer clusters is currently arbitrary and should be correlated with identified clusters from socio-economic covariates. Another, however small weakness is the absence of a clear interpretation of the axes resulting from t-SNE.

Simulation of segment mixtures

The last subject of criticism is the simulation performed in Chapter 7. This is done based on (assumed) identified lifestyle-based consumer segments. This is one way of approaching this clustering. However, the segments are just based on generated patterns looking at *when* people are consuming energy, while leaving out *how much* energy is being consumed. Two families could have very similar normalised energy consumption data and get grouped into the same segment by the algorithm, but one could be living in a big villa and the other one in a small apartment, with very different absolute energy consumption. Additionally, in an actual subnet there might be a larger variety of lifestyles than two. Therefore, other clustering strategies should be considered as well in the future. A last notion to be made, is that the current model adds up time series of 2014 (NL14conv) and 2015 (NL15ae). The days of the week are aligned – Monday 3-1-2014 is being aggregated with Monday 1-1-2015 – but errors might have resulted from this.

8.3 Recommendations for future work

In this last section some recommendations for future work are presented. Three major recommendations are being made:

1. Validate the developed model on rich and real-life data for improved interpretation of results.
2. Standardise and automatise the model for on-line monitoring of the energy transition.
3. Expand the model for improved performance and simulation purposes.

These recommendations are worked out further below.

8.3.1 Lifestyle validation and interpretation

The most important steps that are missing in this thesis are the validation and interpretation of the identified lifestyles. An approach to this was worked out together with Professor Ram Rajagopal during a stay at Stanford University. This validation would be performed on an enriched dataset that, next to energy consumption, includes contextual data like occupancy, socio-economic data and technology data. One of such datasets is the CER-dataset that is openly available for research purposes[1]. Next to smart meter data, this dataset contains information about:

- occupancy of the building (i.e. commercial or residential);
- socio-economic information such as family size and the amount of working or elderly people;
- information about building type and presence of different appliances;
- questionnaires about various energy-related topics, such as work schedule.

Now the idea is to try to predict the variables of interest from the – with LDA estimated – lifestyle-per-home distribution vectors $\{\boldsymbol{\theta}_m\}$. For example, if $\boldsymbol{\pi}_m$ is the answer of consumer m to the question *'How many days are 9-5 working days?'*, then a function $f(.)$ is constructed that predicts π from $\boldsymbol{\theta}$:

$$f(\boldsymbol{\theta}) = \hat{\boldsymbol{\pi}}_m \approx \boldsymbol{\pi}_m. \tag{8.1}$$

Inference on the components of $\boldsymbol{\theta}$ can then be made based on this predictor function $f(.)$. For example, if the first component of $\boldsymbol{\theta}$ correlates strongly with $\boldsymbol{\pi}_i$, then the first identified lifestyle corresponds to a 9-to-5 work day.

A less sophisticated but easier approach to interpreting the customer clusters – instead of the lifestyles – is to look at the average values for each variable for the homes within one segment.

8.3.2 Practical applications

Getting from interesting results to actionable insights is often a challenging hurdle in applied research. As this is not different in this thesis, some recommendations for future application are proposed below.

Use cases

Some use cases for DSO Stedin were briefly worked out in Chapter 7. Other potentially promising use cases for DSO's are:

- Estimation of energy consumption in case of limited data availability. This is especially relevant since privacy regulations in the Netherlands allows DSO's to store a maximum of 2 months of data per home. Energy consumption of a home could then be estimated by the actual consumption of homes within the same consumer segment.

[1] `www.ucd.ie/issda/data/commissionforenergyregulationcer/`

- Updating design parameters near real-time by standardised simultaneity calculation. This is already done partly in Chapter 7 and should be expanded for more thorough analysis.
- Monitoring of (local) energy transition by improved identification of technological and behavioural changes.
- Detecting energy theft through identification of suspicious consumption patterns.

For utilities and other actors in the energy market, the developed framework for lifestyle-based consumer segmentation can be an interesting way of inferring consumer information from just energy consumption. Privacy should be taken into account strongly when such an application is considered.

Integration and automation

Currently the implementation requires various scripts, loosely linked together. The implementation is not the most user-friendly, nor well documented. In order to be of practical significance to a DSO like Stedin, it is essential to integrate the developed methodology into existing analytical infrastructure, or to create a tool that can be easily used by asset managers or grid planners in their standard operations. For this, a detailed understanding of the used methodology is not needed, as long as input and output are clearly defined and aligned with current data and parameter practices.

Linked to the subject of integration is the automation of the analysis, such that new batches of smart meter data are analysed (near) real-time. In order to do this effectively, prototype sets of load shapes, lifestyles and customer segments need to be generated based on historical data. Incoming data will then automatically be transformed to each of these variable-spaces. The flexibility of adaptive K-means allows to automatically update the load shape dictionary if new observations cause the compactness criterion to be violated. The expansion of the LDA model with a prior on $\boldsymbol{\eta}$, resulting in the smoothed LDA model, allows the $\boldsymbol{\eta}$ to assign probabilities to these new load shapes. In this way, the model adapts to new data and enables the monitoring of the local energy transition to be performed near real-time. Also, automatic calculation and simulation of a standardised simultaneity measure could be included in order to constantly update design parameters.

8.3.3 Model improvements and scientific research

A wide variety of model improvements can be thought of, and probably a full PhD-program could be filled with working them all out. Some of the improvements that promise the greatest improvement in performance or scientific relevance are proposed below.

Including time- and weather dependencies

An important assumption in the model is that each element w_i in a series of load shapes $\mathbf{w}_m$ is generated independently conditional on the generated lifestyle τ_i on day i, and that each lifestyle τ_i on day i is generated independently conditional on the generated lifestyle distribution $\boldsymbol{\theta}_m$. Dependencies on time (e.g. week/weekend or summer/winter) are not taken into account here. This assumption is called the 'bag-of-words'-assumption in text mining. A model that goes beyond this is developed in [109], in which words are generated conditional on the previous l words. In [110] a method was developed that switches between LDA and a Hidden Markov Model (HMM). This HMM models a document as a collection of chapters, with each chapter its own mixture of topics. A similar approach could be used to model the effect of weather or holidays. Another approach could be to fit two vectors $\boldsymbol{\theta}_{m1}$ and $\boldsymbol{\theta}_{m2}$ for weekdays and weekends separately.

Hierarchical, correlated and dynamic lifestyles

Many other alternatives for the ordinary LDA have been proposed in literature, of which three will be discussed briefly here.

An important assumption in LDA is that the number of topics is assumed to be known and fixed. A solution to this is the Bayesian non-parametric model, in which a prior distribution over the number of topics K is assumed and a posterior mean is estimated. This approach had been extended to the Hierarchical Dirichlet Process (HDP), in which a hierarchical structier of topics is assumed [102], moving from general to more concrete.

Another assumption that potentially influences the results is the assumption that different topics are uncorrelated. This is unrealistic on both text-modelling and energy consumption modelling. To improve this, the Correlated Topic Model (CTM) was introduced [101]. This model is also implemented in the **R**-implementation of LDA.

A third assumption is the exchangeability of documents, thus the ordering of documents not playing a role. In [111] the dynamic topic model was developed, which takes into account the temporal order of documents. Topics are allowed to change over time in this model. In energy consumption context, this could be of interest when longer time series are considered.

Load shape clustering

A last, but potentially very interesting, model improvement is to include the load shape generation in the Bayesian framework and thus simultaneously fitting the load shapes and the lifestyles. This Bayesian framework is worked out in equations (8.2) - (8.6). A 24-dimensional normal distribution is modelled here to generate daily load vector **y**, with a normal and inverse-gamma distribution as prior distributions. The vectors $\{\boldsymbol{\mu}_k\}_{k=1}^K$ can now be seen as the cluster means in the load shape dictionary.

$$\begin{aligned}
y_{m,n} \mid \tau_{m,n}, \{(\boldsymbol{\mu}_k, \Sigma_k)\}_{k=1}^K &\sim \text{Normal}(\boldsymbol{\mu}_{\tau_{m,n}}, \Sigma_{\tau_{m,n}}), && (8.2)\\
\tau_{m,n} \mid \boldsymbol{\theta}_m &\sim \text{Cat}(\boldsymbol{\theta}_m), && (8.3)\\
\boldsymbol{\theta}_m \mid \boldsymbol{\alpha} &\sim \text{Dir}(\boldsymbol{\alpha}), && (8.4)\\
\boldsymbol{\mu}_k \mid \boldsymbol{\mu}_0, \Sigma_0 &\sim \text{Normal}(\boldsymbol{\mu}_0, \Sigma_0), && (8.5)\\
\Sigma_k \mid \lambda, \beta &\sim \Gamma^{-1}(\lambda, \beta), && (8.6)
\end{aligned}$$

with $\Gamma^{-1}(\cdot)$ the inverse-gamma distribution, the conjugate prior distribution of the variance. The variance could also be assumed constant for all load shapes. VEM or Gibbs sampling could be deployed for model identification.

Note that daily loads **y** instead of normalised daily loads **x** are used here. An approach with normalised daily loads could assume a Dirichlet distribution for **x** since all elements sum up to one. A Dirichlet should then also be used as a prior distribution.

This model should be worked out further in future research. Important to not is that the increased complexity introduced could cause the computational complexity to increase.

Bibliography

[1] WEF, “Renewable Infrastructure Investment Handbook: A Guide for Institutional Investors ,” 2017.

[2] E. Gosden, “New record for cheapest offshore wind farm,” 2016. Retrieved from telegraph.co.uk: `www.tinyurl.com/tlgrphWind`.

[3] A. Upadhyay, “SoftBank and Foxconn Bring India Some of World’s Cheapest Solar,” 2017. Retrieved from bloomberg.com: `www.tinyurl.com/bloomberg2017`.

[4] IEA, “Snapshot of global photovoltaic markets 2016,” 2017.

[5] J. Shankleman, “Two charts that show how renewable energy has blown away expectations.” Retrieved from Bloomberg.com: `www.tinyurl.com/blmbrg16`, 2016.

[6] J. Trommelen, “Stroomnet kan zonnepanelen-hausse in groningen niet aan,” 2016. Retrieved from volkskrant.nl: `www.tinyurl.com/vlkskrntZon`.

[7] Stedin Groep, “Jaarbericht 2016.” Retrieved from stedin.net: `www.tinyurl.com/stdnJvslg`, Maart 2017.

[8] ACM, “Toezicht regionale netbeheerders elektriciteit.” Retrieved from acm.nl: `www.tinyurl.com/acmTzcht`, 2016. Date accessed: 2017-01-12.

[9] Y. Koç, *On Robustness of Power Grids*. PhD thesis, TU Delft, Delft University of Technology, 2015.

[10] Wikipedia, “A typical power grid layout.” Available at Wikipedia.com: `www.tinyurl.com/wikiPgrid`, 2010. Accessed: 2016-07-05.

[11] A. H. Fanney, V. Payne, T. Ullah, L. Ng, M. Boyd, F. Omar, M. Davis, H. Skye, B. Dougherty, B. Polidoro, and et al., “Net-zero and beyond! design and performance of nist’s net-zero energy residential test facility,” *Energy and Buildings*, vol. 101, p. 95–109, 2015.

[12] PBL, “Op weg naar klimaatneutrale woningvoorraad in 2050. investeringsopties voor een kosten-effectieve energievoorziening,” 2014.

[13] CBS, “Hernieuwbare elektriciteit; productie en vermogen.” Retrieved from: `www.tinyurl.com/cbsNewPV`, June 2017.

[14] P. I. R. (PIR), “Geïnstalleerd vermogen pv nederland 2015,” 2016. Retrieved from: `www.tinyurl.com/y8at24gq`.

[15] Eurostat, “Share of renewables in gross inland energy consumption, 2014 (%) yb16.” Retrieved from: `www.ec.europa.eu/eurostat/statistics-explained/index.php`, 2016. Accessed: 2017-04-11.

[16] SolarPower Europe, “Global market outlook (2016-2020).” `www.solarpowereurope.org/reports/global-market-outlook-2017/`, 2016.

[17] CE Delft, “Op weg naar een klimaatneutrale gebouwde omgeving 2050,” 2015.

[18] CBS, “Warmtepompen; aantallen, thermisch vermogen en energiestromen.” Retrieved from: `www.statline.cbs.nl/Statweb/publication`, 2016.

[19] ECN, “Nationale energieverkenning 2014,” 2014.

[20] RVO, “Elektrisch vervoer in nederland - highlights 2016.” Retrieved from RVO.nl: `www.tinyurl.com/rvo2016`, 2017.

[21] RVO, “Elektrisch rijden in de versnelling,” 2011.

[22] Y. Parag and B. K. Sovacool, “Electricity market design for the prosumer era,” *Nature Energy*, vol. 1, p. 16032, 2016.

[23] CAISO. Retrieved from caiso.com: `www.tinyurl.com/caisoRnwbl`, 2013.

[24] W. van Westering, A. Zondervan, A. Bakkeren, F. Mijnhardt, and J. van der Els, “Assessing and mitigating the impact of the energy demand in 2030 on the dutch regional power distribution grid,” in *2016 IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC)*, pp. 1–6, April 2016.

[25] T. Ackermann and V. Knyazkin, “Interaction between distributed generation and the distribution network: operation aspects,” in *Transmission and Distribution Conference and Exhibition 2002: Asia Pacific. IEEE/PES*, vol. 2, pp. 1357–1362, IEEE, 2002.

[26] E. J. Coster, “Distribution grid operation including distributed generation,” *Eindhoven University of Technology, The Netherlands*, 2010.

[27] F. Provoost, “Intelligent distribution network design,” *Eindhoven University of Technology*, 2009.

[28] J. von Appen, M. Braun, T. Stetz, K. Diwold, and D. Geibel, “Time in the sun: the challenge of high pv penetration in the german electric grid,” *IEEE power and energy magazine*, vol. 11, no. 2, pp. 55–64, 2013.

[29] U.S. Department of Energy, “Grid 2030: A national vision for electricity’s second 100 years,” 2003.

[30] FERC, “Assessment of demand response and smart metering,” 2008.

[31] Ministerie van Economische Zaken, “Kamerbrief over besluit grootschalige uitrol slimme meters,” 2014.

[32] X. Zhang and S. Grijalva, “A data-driven approach for detection and estimation of residential pv installations,” *IEEE Transactions on Smart Grid*, vol. 7, no. 5, p. 2477–2485, 2016.

[33] V. M. Balijepalli, V. Pradhan, S. Khaparde, and R. Shereef, “Review of demand response under smart grid paradigm,” in *Innovative Smart Grid Technologies-India (ISGT India), 2011 IEEE PES*, pp. 236–243, IEEE, 2011.

[34] RMI, “The economics of demand flexibility,” 2015. Retrieved from: `www.rmi.org/insights/reports/economics-demand-flexibility/`.

[35] N. Yu, S. Shah, R. Johnson, R. Sherick, M. Hong, and K. Loparo, “Big data analytics in power distribution systems,” in *Innovative Smart Grid Technologies Conference (ISGT), 2015 IEEE Power & Energy Society*, IEEE, February 2015.

[36] H. A. Cao, C. Beckel, and T. Staake, “Are domestic load profiles stable over time? an attempt to identify target households for demand side management campaigns,” in *IECON 2013 - 39th Annual Conference of the IEEE Industrial Electronics Society*, pp. 4733–4738, Nov 2013.

[37] M. E. Dyson, S. D. Borgeson, M. D. Tabone, and D. S. Callaway, “Using smart meter data to estimate demand response potential, with application to solar energy integration,” *Energy Policy*, vol. 73, pp. 607–619, 2014.

[38] G. Chicco, R. Napoli, and F. Piglione, “Comparisons among clustering techniques for electricity customer classification,” *IEEE Transactions on Power Systems*, vol. 21, pp. 933–940, May 2006.

[39] J. Kwac, J. Flora, and R. Rajagopal, “Household energy consumption segmentation using hourly data,” *IEEE Trans. Smart Grid IEEE Transactions on Smart Grid*, vol. 5, no. 1, p. 420–430, 2014.

[40] VISDOM, “VISDOM spotlight.” Retrieved from: `www.tomkat.stanford.edu/visdom-spotlight`, 2016. Accessed: 2017-03-20.

[41] VADER, “VADER visualization and analytics of distributed energy resources.” Retrieved from: `www-group.slac.stanford.edu/gismo/research/vader/`, 2016. Accessed: 2017-03-22.

[42] NEDU, “Verbruiksprofielen.” Retrieved from: `www.nedu.nl/portfolio/verbruiksprofielen/`, 2016. Accessed: 2016-08-05.

[43] R. R. B. A. Smith, J. Wong, “A simple way to use interval data to segment residential customers for energy efficiency and demand response program targeting,” 2012.

[44] F. Iglesias and W. Kastner, “Analysis of similarity measures in times series clustering for the discovery of building energy patterns,” *Energies*, vol. 6, no. 2, p. 579–597, 2013.

[45] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader, “Load pattern-based classification of electricity customers,” *IEEE Transactions on Power Systems*, vol. 19, pp. 1232–1239, May 2004.

[46] V. Figueiredo, F. Rodrigues, Z. Vale, and J. Gouveia, “An electric energy consumer characterization framework based on data mining techniques,” *IEEE Transactions on Power Systems*, vol. 20, no. 2, p. 596–602, 2005.

[47] S. V. Verdu, M. O. Garcia, C. Senabre, A. G. Marin, and F. J. G. Franco, “Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps,” *IEEE Transactions on Power Systems*, vol. 21, pp. 1672–1682, Nov 2006.

[48] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, “Two-stage pattern recognition of load curves for classification of electricity customers,” *IEEE Transactions on Power Systems*, vol. 22, pp. 1120–1128, Aug 2007.

[49] Z. Yin, T. Teeraratkul, and N. Tamang, “Appliance based model for energy consumption segmentation,” 2014.

[50] K. C. Armel, A. Gupta, G. Shrimali, and A. Albert, “Is disaggregation the holy grail of energy efficiency? the case of electricity,” *Energy Policy*, vol. 52, pp. 213 – 234, 2013. Special Section: Transition Pathways to a Low Carbon Economy.

[51] K. Ehrhardt-Martinez, K. A. Donnelly, S. Laitner, *et al.*, “Advanced metering initiatives and residential feedback programs: a meta-review for household electricity-saving opportunities,” American Council for an Energy-Efficient Economy Washington, DC, 2010.

[52] A. Zoha, A. Gluhak, M. A. Imran, and S. Rajasegarar, “Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey,” *Sensors*, vol. 12, no. 12, pp. 16838–16866, 2012.

[53] A. Kavousian, R. Rajagopal, and M. Fischer, “Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants’ behavior,” *Energy*, vol. 55, p. 184–194, 2013.

[54] E. C. Kara, M. D. Tabone, J. S. MacDonald, D. S. Callaway, and S. Kiliccote, “Quantifying flexibility of residential thermostatically controlled loads for demand response: a data-driven approach,” in *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, pp. 140–147, ACM, 2014.

[55] A. Albert and R. Rajagopal, “Thermal profiling of residential energy use,” in *2015 IEEE Power Energy Society General Meeting*, pp. 1–1, July 2015.

[56] C. Flath, D. Nicolay, T. Conte, C. V. Dinther, and L. Filipova-Neumann, “Cluster analysis of smart metering data,” *Business & Information Systems Engineering*, vol. 4, p. 31–39, Dec 2012.

[57] T. K. Wijaya, *Pervasive Data Analytics for Sustainable Energy Systems*. PhD thesis, Ecole Polytechnique Federale de Lausanne, Lausanne, 2015.

[58] M. Pedersen and B. Hydro, “Segmenting residential customers: energy and conservation behaviors,” *Proceedings of the 2008 ACEEE Summer Study on Energy Efficiency in Buildings*, vol. 7, pp. 229–41, 2008.

[59] J. Kwac, J. Flora, and R. Rajagopal, “Lifestyle segmentation based on energy consumption data,” *IEEE Transactions on Smart Grid*, p. 1–1, 2016.

[60] A. Albert and R. Rajagopal, “Smart meter driven segmentation: What your consumption says about you,” *IEEE Trans. Power Syst. IEEE Transactions on Power Systems*, vol. 28, no. 4, p. 4019–4030, 2013.

[61] B. Sütterlin, T. A. Brunner, and M. Siegrist, “Who puts the most energy into energy conservation? a segmentation of energy consumers based on energy-related behavioral characteristics,” *Energy Policy*, vol. 39, no. 12, pp. 8137 – 8152, 2011. Clean Cooking Fuels and Technologies in Developing Economies.

[62] T. F. Sanquist, H. Orr, B. Shui, and A. C. Bittner, “Lifestyle factors in us residential electricity consumption,” *Energy Policy*, vol. 42, pp. 354–364, 2012.

[63] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, “Customer characterization options for improving the tariff offer,” *IEEE Transactions on Power Systems*, vol. 18, no. 1, pp. 381–387, 2003.

[64] A. Mutanen, M. Ruska, S. Repo, and P. Jarventausta, “Customer classification and load profiling method for distribution systems,” *IEEE Transactions on Power Delivery*, vol. 26, no. 3, pp. 1755–1763, 2011.

[65] Y. Wang, Q. Chen, C. Kang, and Q. Xia, “Clustering of electricity consumption behavior dynamics toward big data applications,” *IEEE Transactions on Smart Grid*, vol. 7, pp. 2437–2447, Sept 2016.

[66] A. Albert, R. Rajagopal, and R. Sevlian, “Segmenting consumers using smart meter data,” in *Proceedings of the Third ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, BuildSys 11, (New York, NY, USA), pp. 49–50, ACM, 2011.

[67] M. Hayn, V. Bertsch, and W. Fichtner, “Electricity load profiles in europe: The importance of household segmentation,” *Energy Research & Social Science*, vol. 3, p. 30–45, 2014.

[68] T. K. Wijaya, T. Ganu, D. Chakraborty, K. Aberer, and D. P. Seetharam, “Consumer segmentation and knowledge extraction from smart meter and survey data,” *Proceedings of the 2014 SIAM International Conference on Data Mining*, p. 226–234, 2014.

[69] G. K. James Honaker and M. Blackwell, “Amelia ii: A program for missing data.” Retrieved from `www.cran.r-project.org/web/packages/Amelia/Amelia.pdf`, 2015.

[70] F. Wilcoxon, “Individual comparisons by ranking methods,” *Biometrics Bulletin*, vol. 1, no. 6, p. 80, 1945.

[71] Hager, “Zakboek verdelers t/m 125 a,” 2013.

[72] Electrical Installation Wiki, “Estimation of actual maximum kva demand.” Retrieved from `www.tinyurl.com/schneiderSimultaneity`.

[73] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, (Berkeley, Calif.), pp. 281–297, University of California Press, 1967.

[74] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, “Np-hardness of euclidean sum-of-squares clustering,” *Machine Learning*, vol. 75, no. 2, pp. 245–248, 2009.

[75] Y. P. Raykov, A. Boukouvalas, F. Baig, and M. A. Little, “What to do when k-means clustering fails: a simple yet principled alternative algorithm,” *PloS one*, vol. 11, no. 9, p. e0162259, 2016.

[76] D. T. Pham, S. S. Dimov, and C. D. Nguyen, “Selection of k in k-means clustering,” *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 219, no. 1, pp. 103–119, 2005.

[77] R. L. Thorndike, “Who belongs in the family,” *Psychometrika*, pp. 267–276, 1953.

[78] M. Honarkhah and J. Caers, “Stochastic simulation of patterns using distance-based pattern modeling,” *Mathematical Geosciences*, vol. 42, no. 5, pp. 487–517, 2010.

[79] G. Schwarz *et al.*, “Estimating the dimension of a model,” *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[80] H. Akaike, “A new look at the statistical model identification ieee trans auto control 19 (6): 716–723,” *Find this article online*, 1974.

[81] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53 – 65, 1987.

[82] J. C. Dunn, “Well-separated clusters and optimal fuzzy partitions,” *Journal of Cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.

[83] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[84] I. S. Dhillon and D. S. Modha, “Concept decompositions for large sparse text data using clustering,” *Machine Learning*, vol. 42, no. 1, pp. 143–175, 2001.

[85] D. Arthur, B. Manthey, and H. Röglin, "k-means has polynomial smoothed complexity," in *Foundations of Computer Science, 2009. FOCS'09. 50th Annual IEEE Symposium on*, pp. 405–414, IEEE, 2009.

[86] M. Inaba, N. Katoh, and H. Imai, "Variance-based k-clustering algorithms by voronoi diagrams and randomization," *IEICE Transactions on Information and Systems*, vol. 83, no. 6, pp. 1199–1206, 2000.

[87] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[88] G. Hamerly and C. Elkan, "Alternatives to the k-means algorithm that find better clusterings," *Proceedings of the eleventh international conference on Information and knowledge management - CIKM '02*, 2002.

[89] S. K. Bhatia, "Adaptive k-means clustering.," in *FLAIRS Conference*, pp. 695–699, 2004.

[90] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[91] P. Pinoli, D. Chicco, and M. Masseroli, "Latent dirichlet allocation based on gibbs sampling for gene function prediction," in *Computational Intelligence in Bioinformatics and Computational Biology, 2014 IEEE Conference on*, pp. 1–8, IEEE, 2014.

[92] N. Rasiwasia and N. Vasconcelos, "Latent dirichlet allocation models for image classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 11, pp. 2665–2679, 2013.

[93] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, pp. 77–84, Apr. 2012.

[94] K. Nigam, A. K. Mccallum, S. Thrun, and T. M. Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM," *Machine Learning*, vol. 39, no. 2/3, pp. 103–134, 2000.

[95] T. Hofmann, "Probabilistic latent semantic indexing," *Proceedings of the Twenty-Second Annual International SIGIR Conference*, 1999.

[96] D. M. Pennock, S. Lawrence, R. Popescul, and L. H. Ungar, "Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments," *Uncertainty in Artificial Intelligence, Proceedings of the Seventeenth Conference*, 2001.

[97] H. Robbins, "An empirical bayes approach to statistics," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, (Berkeley, Calif.), pp. 157–163, University of California Press, 1956.

[98] E. Novak and H. Woźniakowski, "Tractability of multivariate problems," Feb 2008.

[99] J. M. Dickey, "Multiple hypergeometric functions: Probabilistic interpretations and statistical uses," *Journal of the American Statistical Association*, vol. 78, no. 383, p. 628, 1983.

[100] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Learning in Graphical Models*, p. 105–161, 1998.

[101] D. M. Blei and J. D. Lafferty, "A correlated topic model of science," *The Annals of Applied Statistics*, vol. 1, no. 1, p. 17–35, 2007.

[102] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies," *Journal of the ACM*, vol. 57, no. 2, 2010.

[103] S. Kullback and R. A. Leibler, “On information and sufficiency,” *Ann. Math. Statist.*, vol. 22, pp. 79–86, 03 1951.

[104] K. F. Pearson, “X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling,” *Philosophical Magazine*, vol. 50, no. 302, pp. 157–175, 1900.

[105] L. Rokach and O. Maimon, “Clustering methods,” in *Data mining and knowledge discovery handbook*, pp. 321–352, Springer, 2005.

[106] K. F. Pearson, “Liii. on lines and planes of closest fit to systems of points in space,” *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901.

[107] L. van der Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[108] J. Handl, J. Knowles, and D. B. Kell, “Computational cluster validation in post-genomic data analysis,” *Bioinformatics*, vol. 21, pp. 3201–3212, Aug. 2005.

[109] H. M. Wallach, “Topic modeling: Beyond bag-of-words,” in *Proceedings of the 23rd International Conference on Machine Learning*, ICML ’06, (New York, NY, USA), pp. 977–984, ACM, 2006.

[110] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum, “Integrating topics and syntax,” in *Advances in neural information processing systems*, pp. 537–544, 2005.

[111] D. M. Blei and J. D. Lafferty, “Dynamic topic models,” in *Proceedings of the 23rd International Conference on Machine Learning*, ICML ’06, (New York, NY, USA), pp. 113–120, ACM, 2006.

[112] T. Boggs, “Visualizing dirichlet distributions with matplotlib.” `www.blog.bogatron.net/blog/2014/02/02/visualizing-dirichlet-distributions`, 2015. Accessed: 2017-03-09.

[113] D. Adams and M. Carwardine, *Last Chance to See*. Arrow Books, Collins, 2009.

Appendices

Appendix A
Daily loads of consumers

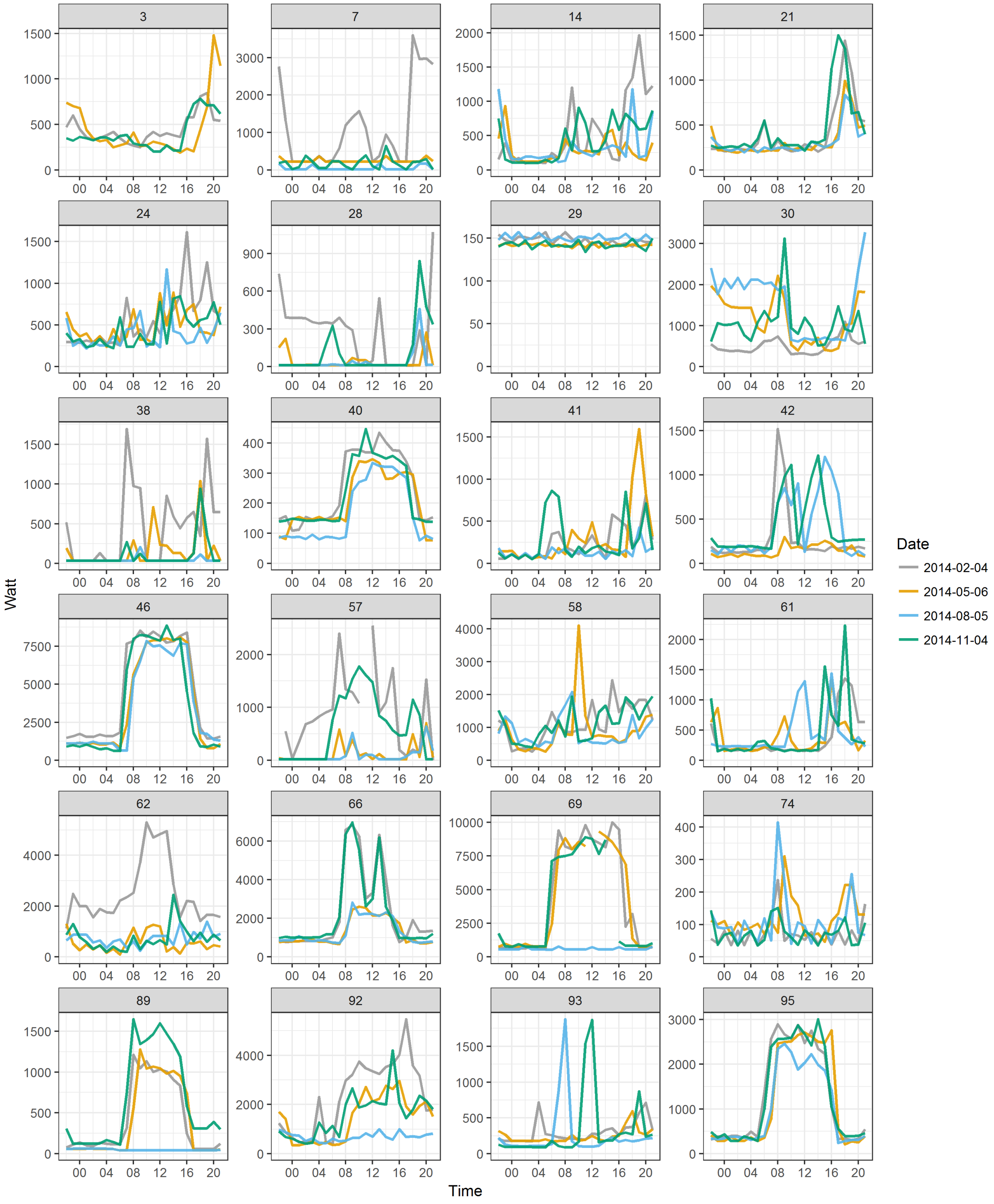


Figure A.1: Daily loads of 24 random consumers on four random Tuesdays from the NL14conv dataset.

Appendix B

Results load shape clustering

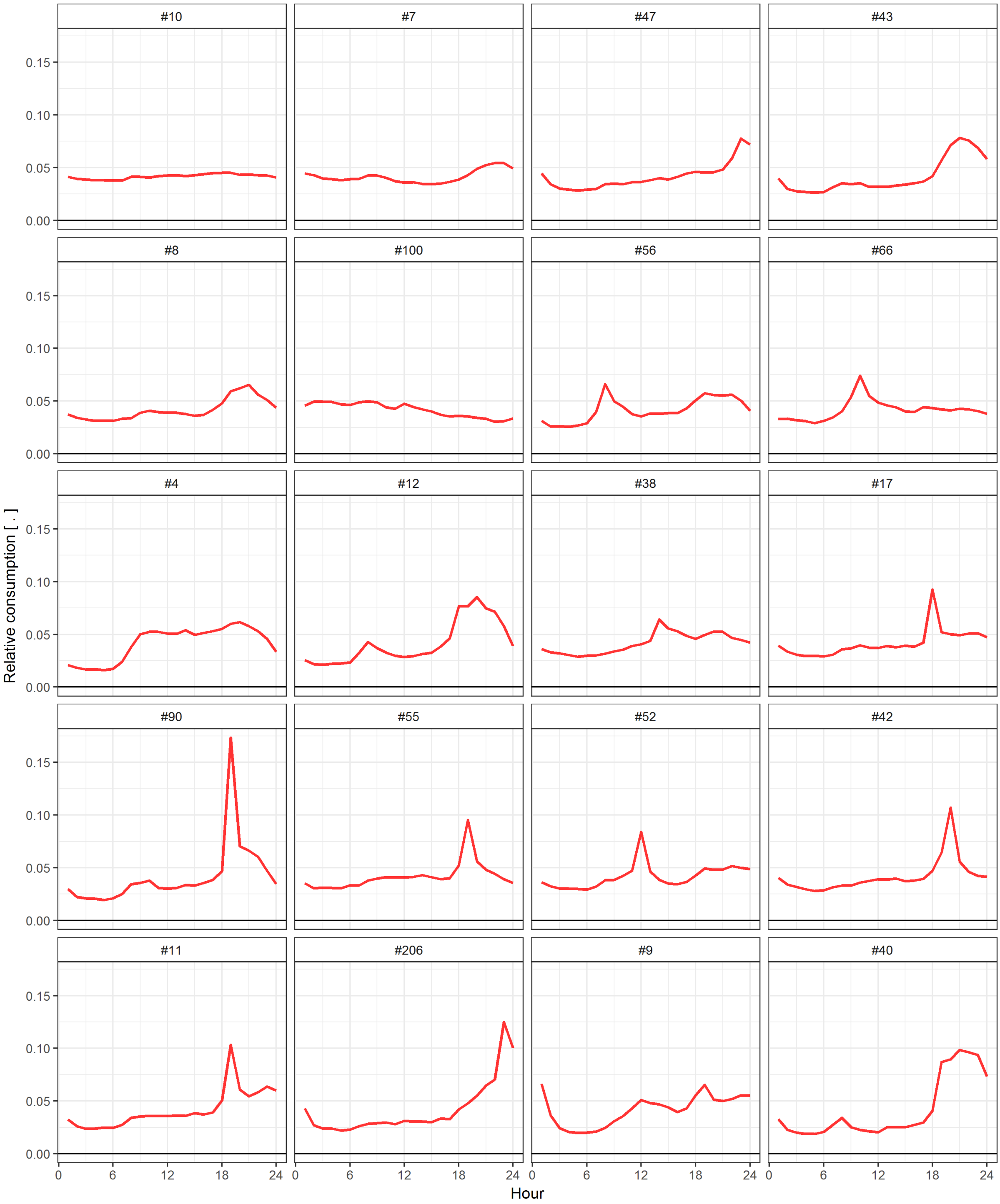


Figure B.1: The 20 most occuring load shapes in the dictionary. Statistics can be found in Table B.1

Table B.1: Statistics of 20 most and least occurring load shapes

order	cluster	n	μ (kWh)	σ (kWh)	rel. n (%)	rel. cum. consumption (%)
1	10	102,031	5.5	11.0	5.7	3.5
2	7	28,992	8.6	11.8	1.6	1.5
3	47	19,602	8.6	9.1	1.1	1.0
4	43	19,136	9.1	8.0	1.1	1.1
5	8	18,996	10.7	11.0	1.1	1.3
6	100	13,056	7.7	9.3	0.7	0.6
7	56	12,284	11.3	13.2	0.7	0.9
8	66	12,173	10.0	11.8	0.7	0.7
9	4	12,165	11.6	9.9	0.7	0.9
10	12	11,813	9.7	6.9	0.7	0.7
11	38	11,089	11.7	12.5	0.6	0.8
12	17	11,085	11.4	8.7	0.6	0.8
13	90	10,884	7.5	4.1	0.6	0.5
14	55	10,709	10.1	9.5	0.6	0.7
15	52	10,159	9.9	9.3	0.6	0.6
16	42	10,116	9.8	8.4	0.6	0.6
17	11	9,325	10.3	6.3	0.5	0.6
18	206	9,250	8.2	5.9	0.5	0.5
19	9	8,807	11.6	8.1	0.5	0.6
20	40	8,686	8.4	6.1	0.5	0.5
...	...	...	...	...	...	...
644	236	118	1.8	2.8	0.01	0.001
645	484	112	-1.4	3.5	0.01	-0.001
646	592	109	7.6	5.6	0.01	0.01
647	227	107	6.3	4.3	0.01	0.004
648	183	93	4.8	3.4	0.01	0.003
649	381	84	5.5	3.5	0.005	0.003
650	186	82	4.9	4.4	0.005	0.002
651	415	82	11.2	6.5	0.005	0.01
652	447	80	6.0	5.5	0.004	0.003
653	531	63	7.8	4.7	0.003	0.003
654	313	58	10.9	4.0	0.003	0.004
655	548	52	5.7	4.2	0.003	0.002
656	93	44	1.7	1.8	0.002	0.000
657	242	43	3.4	3.8	0.002	0.001
658	440	42	4.0	3.1	0.002	0.001
659	378	31	1.7	2.9	0.002	0.000
660	175	30	-4.3	2.7	0.002	-0.001
661	205	25	6.8	3.1	0.001	0.001
662	423	16	2.6	3.7	0.001	0.000
663	483	10	2.8	2.0	0.001	0.000

Appendix C

Used distributions

The two key distributions in this model are the multinomial distribution and the dirichlet distribution. Both distribution and their relationship will be introduced briefly in this section.

C.1 Multinomial distribution.

The multinomial distribution is the (multidimensional) generalization of various related discrete probability distributions: the Bernoulli, binomial and categorical distributions. The **Bernoulli distribution** with parameter p is the probability distribution of a success in a binary experiment, like flipping an (unfair) coin that has probability p to show head: $\mathbb{P}(X = 1) = p$ (and obviously $\mathbb{P}(X = 0) = 1 - p$). The **binomial distribution** with parameters n and p is the distribution of n independent identical Bernoulli trials, e.g. the number of heads after n flips of an unfair coin. The **categorical distribution** with parameter $p \in \mathbb{R}^k_{\geq 0}$, $\sum_i p_i = 1$, is the probability distribution of a discrete random variable that can have k different outcomes, for example throwing an (unfair) k-sided dice once: $\mathbb{P}(X = i) = \mathrm{p}_i$. The **multinomial distribution** with parameters n and $\mathrm{p} \in \mathbb{R}^k_{\geq 0}$ now combines the binomial and categorical distribution: it is the probability distribution of random variable $X \in \mathbb{N}^k$ that models the number of occurrences of each category in n independent categorical trials, e.g. the number each side of an (unfair) k-sided dice has come up after n throws. The probability mass function of the multinomial distribution is given by

$$\mathbb{P}(X = x; n, p) = \begin{cases} \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}, & \text{if } \sum_i x_i = n \\ 0, & \text{otherwise,} \end{cases} \tag{C.1}$$

which is obviously a generalisation of the above-mentioned distributions:

- $n = 1$ and $k = 2$ gives the Bernoulli distribution,
- $k = 2$ gives the binomial distribution, and
- $n = 1$ gives the categorical distribution.

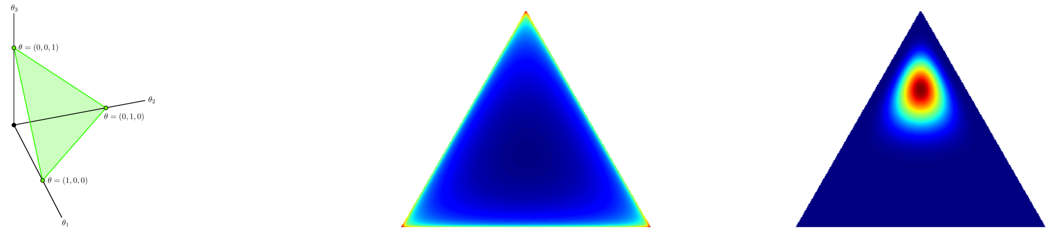
The general notation is $X \sim \text{Multinomial}(n, p)$.

C.2 Dirichlet distribution

The **Dirichlet distribution**, parametrized by the parameter $\boldsymbol{\alpha} \in \mathbb{R}^k$, is a continuous multivariate distribution, such that every realization $y \in \mathbb{R}^k_{\geq 0}$ of stochastic variable $Y \sim \text{Dir}(\boldsymbol{\alpha})$ satisfies $\sum_i y_i = 1$. It is the multivariate generalization of the **beta distribution**. The probability mass function of the Dirichlet distribution is given by:

$$\mathbb{P}(Y = y; \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^{k} y_i^{\alpha_i - 1} \tag{C.2}$$

The Dirichlet distribution is the conjugate prior of the multinomial distribution family. This means that if a stochastic variable is multinomial distributed, and the prior distribution on the parameter vector p is Dirichlet distributed, then the posterior distribution of the parameter is also a Dirichlet. With hyperparameter $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k)$ and observed data $\mathbf{y} = (y_1, ..., y_k)$, with y_i the number of occurrences of category i, this leads to:



(a) Support of a 3D Dirichlet distribution

(b) Probability density of $Dir([.5, .5, .5])$

(c) Probability density of $Dir([5, 5, 15])$

Figure C.1: Examples of the 3D Dirichlet Distribution. On the left, the support of the 3D Dir-distribution - the 2D surface $\{x \in \mathbb{R}^3 : 0 \leq x_i \leq 1, \sum_i x_i = 1\}$ *- is visualized. Next, two Dirichlet probability mass functions on this simplex are plotted.* $\alpha_i < 1$ *gives high probability to the edges of the simplex, thus leading to sparser observations, while* $\alpha_i > 1$ *gives high probability to mixed outcomes [112].*

$$
\begin{aligned}
Y|p &\sim \text{Multinomial}(n, p) && \text{(the likelihood)} \\
p|\boldsymbol{\alpha} &\sim \text{Dir}(\boldsymbol{\alpha}) && \text{(the prior)} \\
p|y, \boldsymbol{\alpha} &\sim \text{Dir}(\mathbf{y} + \boldsymbol{\alpha}) && \text{(the posterior)}
\end{aligned}
$$

The hyperparameter $\boldsymbol{\alpha}$ can now be used to express prior believes about the parameter p that needs to be estimated. If for a certain i, α_i is relatively big compared to the other α_j's, it expresses the believe that the corresponding p_i is relatively big compared to the p_j's. If in general all α_i's are small (typically $\alpha_i < 1$) this enforces sparsity in the multinomial distribution that is inferred, thus observations with one or several big $\hat{p}_i$'s, and many $\hat{p}_j$'s close to zero. This is visualized in figures C.1b and C.1c.

Because of these properties, the Dirichlet distribution is often used in Bayesian modelling as the prior distribution of a multinomial distributed random variable. This is also the main motivation to introduce the Dirichlet distribution in this chapter.

Appendix D

Estimation of latent variable models with EM

In many situation in which parameters need to be estimated, the explicit calculation of the likelihood as a function of the unknown parameters is impossible, because not all required information is available. In other words, the statistical model contains latent variables and Maximum Likelihood can not be used to estimate the unknown parameters directly. An example of such a situation is the case in which it is known that the observed data $\mathbf{x}$ is independently generated from either of two one-dimensional normal distributions with different unknown parameters - $\boldsymbol{\psi}_A = (\mu_A, \sigma_A^2)$ and $\boldsymbol{\psi}_B = (\mu_B, \sigma_B^2)$ - but do not observe from which of the two distributions each data point originates. Using $\phi_{\psi_A}(x)$ as notation for the PDF of the normal distribution with parameters (μ_A, σ_A^2), the likelihood of the complete data then becomes:

$$\mathrm{p}(\mathbf{x}, \mathbf{z}; \boldsymbol{\psi}_A, \boldsymbol{\psi}_B) = \prod_{i=1}^{N} \left(\phi_{\psi_A}(x_i)\right)^{z_i} \left(\phi_{\psi_B}(x_i)\right)^{1-z_i},$$

with

$$z_i = \begin{cases} 1 & \text{if observation } x_i \text{ comes from group A,} \\ 0 & \text{if observation } x_i \text{ comes from group B.} \end{cases}$$

This likelihood is now undetermined since the group assignments z_i are unknown.

A frequently used method to handle such problems is the expectation maximisation (EM)-algorithm. This iterative method consists of an Expectation (E)-step, in which the expected complete likelihood given the data and current parameter estimates $\boldsymbol{\psi}^{(t)}$ is calculated, and a Maximization (M)-step where the expected complete likelihood is maximized to find new parameter estimates $\boldsymbol{\psi}^{(t+1)}$.

EM-algorithm. For observed data $\mathbf{x} \in \mathbb{R}^{d_x}$, missing data (or latent variable) $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^{d_z}$ and unknown parameter $\boldsymbol{\psi} \in \Omega$, the EM-algorithm is summarized by the following steps:

1. *Initialization*. Choose initial parameter estimate $\boldsymbol{\psi}^{(0)}$ and set $i = 0$.
2. *E-step*. For step i, derive
$$\mathrm{p}(\mathbf{z}|\mathbf{x}, \boldsymbol{\psi}^{(i)}), \tag{D.1}$$
the probability distribution function of the missing data (or latent variables) $\mathbf{z}$ conditional on the observed data $\mathbf{x}$ and current parameter estimate $\boldsymbol{\psi}^{(i)}$. Using this conditional probability, form the *expected complete log-likelihood*:
$$\begin{aligned} Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(i)}) &= \mathbb{E}_{\mathbf{Z}|\mathbf{x}, \boldsymbol{\psi}^{(i)}} \left[\log\left(\mathrm{p}(\mathbf{x}, \mathbf{Z} \mid \boldsymbol{\psi})\right)\right] & \text{(D.2)} \\ &= \int_{\mathcal{Z}} \log(\mathrm{p}(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\psi}))\mathrm{p}(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\psi}^{(i)}) d\mathbf{z}, & \text{(D.3)} \end{aligned}$$
which is a function of the parameter $\boldsymbol{\psi}$ and depends implicitly on $\boldsymbol{\psi}^{(i)}$.
3. *M-step*. Now find the next approximation $\boldsymbol{\psi}^{(i+1)}$ by maximizing this expression over all possible values for $\boldsymbol{\psi}$:
$$\boldsymbol{\psi}^{(i+1)} = \underset{\boldsymbol{\psi} \in \Omega}{\operatorname{argmax}}\, Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(i)}) \tag{D.4}$$
4. Repeat step 2 and 3 until some convergence criterion is met, i.g. convergence in parameter estimates, $|\,|\ \boldsymbol{\psi}^{(i+1)} - \boldsymbol{\psi}^{(i)}|| < \epsilon$, or log-likelihood, $|\log \mathrm{p}(\mathbf{x} \mid \boldsymbol{\psi}^{(i+1)}) - \log \mathrm{p}(\mathbf{x} \mid \boldsymbol{\psi}^{(i)})| < \epsilon$, for some $\epsilon > 0$.

Now the parameters are estimated as $\hat{\psi} = \psi^{i+1}$, the last estimate before the algorithm is converged. The latent variables can be estimated by the posterior mean $\hat{\mathbf{z}} = \mathbb{E}_{\mathbf{z}|\mathbf{x},\hat{\psi}}[\mathbf{z}]$, the posterior mode $\hat{\mathbf{z}} = \underset{\mathbf{z}\in\mathcal{Z}}{\operatorname{argmax}}\, \mathrm{p}\left(\mathbf{z}|\mathbf{x},\hat{\psi}\right)$.

D.1 EM for LDA

Now in the context of empirical Bayes for LDA, the (marginal) likelihood that needs to be maximised is given by Equation 5.17. Maximising the likelihood is the same is maximising the log-likelihood:

$$\ell(\boldsymbol{\alpha}, \{\boldsymbol{\eta}\}) = \log \mathrm{p}(C \mid \boldsymbol{\alpha}, \{\boldsymbol{\eta}\}) = \sum_{m=1}^{M} \log \mathrm{p}(\mathbf{w}_m \mid \boldsymbol{\alpha}, \{\boldsymbol{\eta}\}), \tag{D.5}$$

which is still intractable. The EM algorithm for LDA now becomes:

1. Choose initial estimates $\boldsymbol{\alpha}^{(0)}$ and $\{\boldsymbol{\eta}^{(0)}\}$
2. *E-step.* For step i, compute the probability distribution of the latent variables:

$$\mathrm{p}\left(\boldsymbol{\theta}, \boldsymbol{\tau} \mid \mathbf{w}, \boldsymbol{\alpha}^{(i)}, \{\boldsymbol{\eta}^{(i)}\}\right) = \frac{\mathrm{p}\left(\mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\tau} \mid \boldsymbol{\alpha}^{(i)}, \{\boldsymbol{\eta}^{(i)}\}\right)}{\mathrm{p}\left(\mathbf{w} \mid \boldsymbol{\alpha}^{(i)}, \{\boldsymbol{\eta}^{(i)}\}\right)}, \tag{D.6}$$

   and, with the notation $\mathbb{E}_{\mathrm{p}}[\cdot] = \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\tau}|\mathbf{w},\boldsymbol{\alpha}^{(i)},\boldsymbol{\eta}^{(i)}}[\cdot]$, the expected complete likelihood:

$$\begin{aligned} Q(\boldsymbol{\alpha}, \boldsymbol{\eta}; \boldsymbol{\alpha}^{(i)}, \boldsymbol{\eta}^{(i)}) &= \mathbb{E}_{\mathrm{p}}\left[\log \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{\tau}, C \mid \boldsymbol{\alpha}, \boldsymbol{\eta})\right] && \text{(D.7)} \\ &= \int_{\Theta} \int_{\Omega_{\tau}} \log\left(\mathrm{p}(\boldsymbol{\theta}, \boldsymbol{\tau}, C \mid \boldsymbol{\alpha}, \boldsymbol{\eta})\right) \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{\tau} \mid C, \boldsymbol{\alpha}^{(i)}, \boldsymbol{\eta}^{(i)}) \mathrm{d}\boldsymbol{\tau} \mathrm{d}\boldsymbol{\theta} && \text{(D.8)} \end{aligned}$$

3. *M-step.* Find the next approximation for the parameters:

$$\left(\boldsymbol{\alpha}^{(i+1)}, \boldsymbol{\eta}^{(i+1)}\right) = \underset{(\boldsymbol{\alpha},\{\boldsymbol{\eta}\})}{\operatorname{argmax}}\, Q(\boldsymbol{\alpha}, \boldsymbol{\eta}; \boldsymbol{\alpha}^{(i)}, \boldsymbol{\eta}^{(i)}) \tag{D.9}$$

4. Repeat until convergence.

This algorithm cannot be used directly, since Equation D.6 is intractable. A way to handle this issue is applying a convexity-based variational inference method like variational expectation maximisation (VEM), which uses Jensen's inequality to obtain a lower bound on the log-likelihood [100] and then maximizes this lower bound at the M-step of the EM-algorithm.

Appendix E

Load shapes per life style

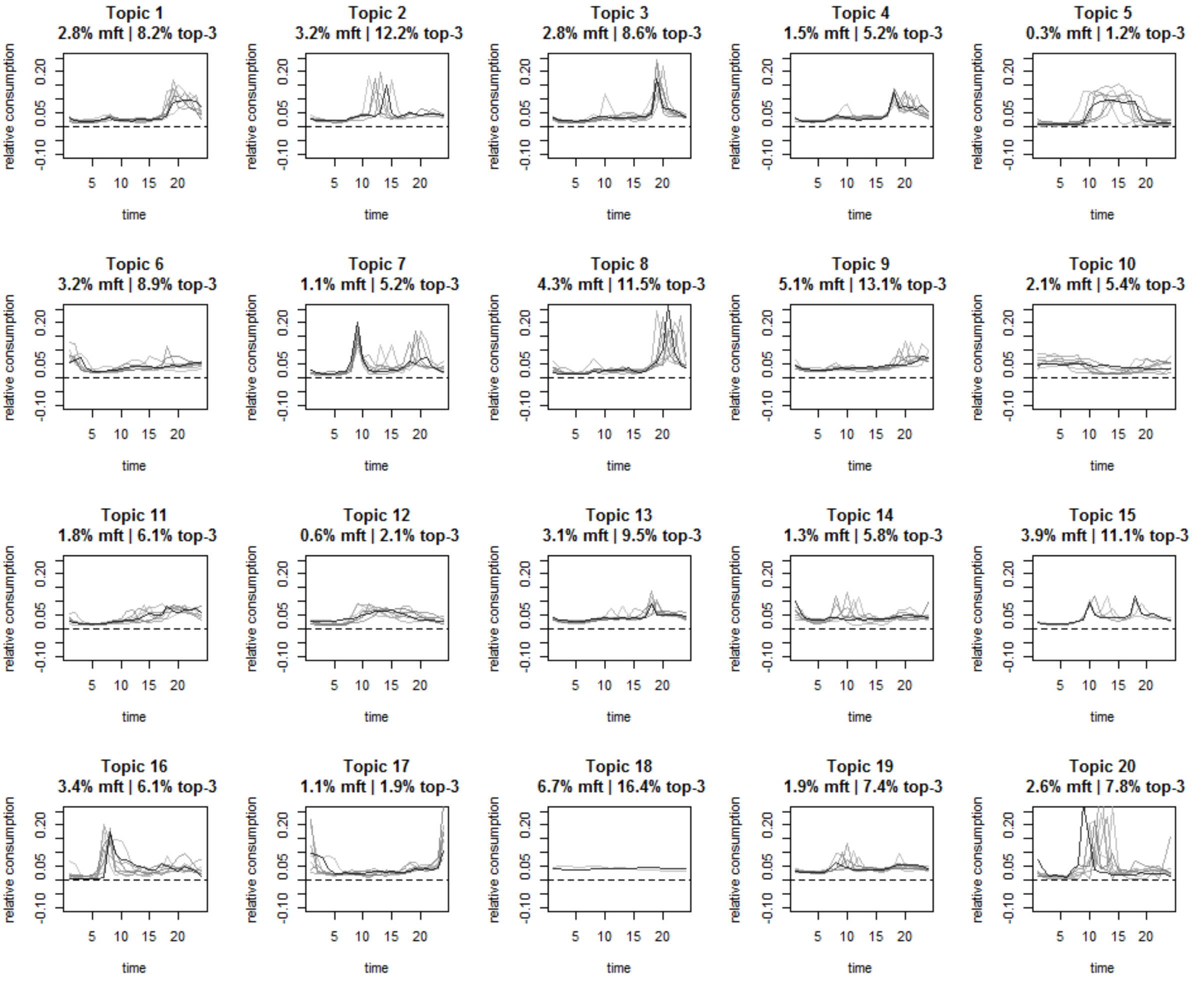


Figure E.1: First 20 out of 40 identifies lifestyles. Each plot shows the characteristic load shapes $\Lambda_{5\%}$ of a lifestyle, with estimated probability > 5%. ‘x% mft’ stands for the percentage of all consumers having this lifestyle as their most frequent one. ’x% top-3’ stands for the percentage of all consumers having this lifestyle in their top-3 of most frequent lifestyles.

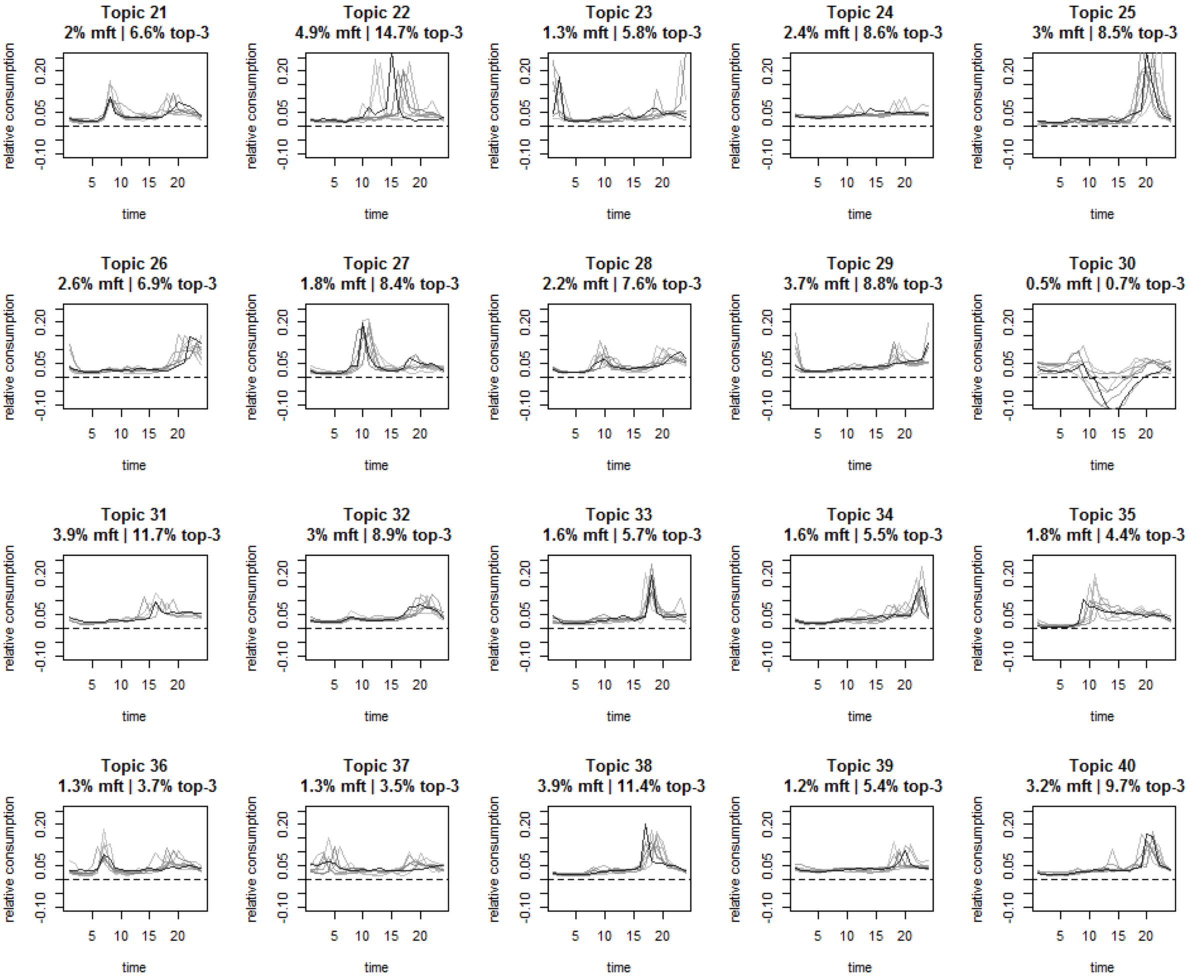


Figure E.2: Last 20 out of 40 identifies lifestyles. Each plot shows the characteristic load shapes $\Lambda_{5\%}$ of a lifestyle, with estimated probability > 5%. 'x% mft' stands for the percentage of all consumers having this lifestyle as their most frequent one. 'x% top-3' stands for the percentage of all consumers having this lifestyle in their top-3 of most frequent lifestyles.

Appendix F
Customer segmentation results

Table F.1: Summary statistics of customer segmentation, applied before t-SNE

i	$\lvert C_i\rvert$	n_{NL15ae}	μ_{kwh}	σ_{kwh}	$\mathcal{T}_3(i)$	$\{\overline{\mathrm{p}}_i(\tau_j) : \tau_j \in \mathcal{T}_3(i)\}$	$\{\overline{\mathrm{p}}_i(\tau_j)/\overline{\mathrm{p}}_{all}(\tau_j) : \tau_j \in \mathcal{T}_3(i)\}$
1	32	27	14.6	13.1	30	0.67	132.7
2	56	1	16.5	11.7	10, 14, 21	0.15, 0.11, 0.08	7.7, 5.7, 3.6
3	147	2	12.5	7	40, 28, 32, 4	0.12, 0.09, 0.1, 0.06	4, 3.6, 3.5, 3.1
4	561	1	10.9	6.1	29, 6	0.12, 0.09	4, 3.3
5	30	1	8.4	5.8	7, 28, 14	0.12, 0.13, 0.07	6.5, 5.1, 3.8
6	71	5	11.3	5.9	37, 36	0.2, 0.05	15.8, 3.4
7	32	1	13.3	9.1	21, 16, 15	0.3, 0.09, 0.11	13.5, 3.7, 3.4
8	78	1	8.7	6.5	26, 17	0.29, 0.04	13.4, 3.4
9	125	1	11.6	9.8	6, 37, 14	0.23, 0.08, 0.07	8.4, 6, 3.3
10	194	1	11.9	6.7	15, 12	0.2, 0.03	5.9, 3.1
11	931	1	6.6	4.2	8	0.11	3.5
12	37	0	6.7	5.2	36	0.31	20.1
13	248	0	7.2	14	10, 18	0.19, 0.3	9.3, 4.8
14	397	0	10.3	5.9	13	0.14	4.7
15	672	0	12.1	9.7	24, 9	0.11, 0.13	3.8, 3.1
16	447	0	8.8	4.5	38	0.15	4.4
17	210	0	4.1	4.4	16, 20	0.38, 0.12	15.4, 4.9
18	22	0	9.5	7	28	0.29	11.4
19	506	0	3.7	4.1	20, 35, 25, 7	0.15, 0.1, 0.12, 0.07	6.3, 6.1, 4.7, 3.7
20	30	0	10.2	4.8	33, 31, 38, 4	0.23, 0.15, 0.15, 0.06	11.4, 4.5, 4.5, 3.4
21	24	0	6.6	3	33, 3, 4	0.21, 0.21, 0.1	10.6, 7.1, 5.3
22	69	0	9.8	5.2	17, 29	0.46, 0.15	44, 5.3
23	38	0	9.1	4.8	1, 4, 11	0.31, 0.14, 0.08	12.2, 7.5, 3.6
24	100	0	6.6	5.6	23, 37, 17	0.19, 0.06, 0.03	10.4, 4.4, 3.1
25	7	0	8.6	3.6	2, 1	0.48, 0.1	13.1, 3.7
26	4	0	3.4	2.6	25, 5, 7, 27	0.29, 0.06, 0.14, 0.2	11.1, 9.1, 7.5, 7.2
27	26	0	7.4	7.2	11	0.36	16
28	23	0	19	17.1	12, 24, 11	0.42, 0.11, 0.07	39.2, 3.9, 3.1
29	30	0	5.4	5.1	35, 12, 28	0.14, 0.04, 0.08	8.7, 3.9, 3.2
30	8	0	7.8	5.4	39, 3	0.28, 0.33	13.8, 11.1
31	23	0	25.5	24.1	5, 12, 18	0.49, 0.11, 0.25	69, 10.3, 4
32	2	0	5.2	2.8	23, 16, 30, 10, 17	0.41, 0.31, 0.04, 0.12, 0.04	22.4, 12.5, 7.8, 6.2, 3.7
33	11	0	4	5.5	7, 35, 2	0.26, 0.09, 0.11	14.5, 5.4, 3.1
34	3	0	22.8	19.3	24, 32, 39, 37	0.24, 0.2, 0.12, 0.07	8.6, 7, 5.7, 5.6

Table F.2: Summary statistics of customer segmentation, applied after t-SNE

i	$\lvert C_i \rvert$	n_{NL15ae}	μ_{kwh}	σ_{kwh}	$\mathcal{T}_3(i)$	$\{\overline{\mathrm{p}}_i(\tau_j) : \tau_j \in \mathcal{T}_3(i)\}$	$\{\overline{\mathrm{p}}_i(\tau_j)/\overline{\mathrm{p}}_{all}(\tau_j) : \tau_j \in \mathcal{T}_3(i)\}$
1	28	27	14.1	13.5	30	0.74	148.5
2	153	1	8.5	12.3	10	0.33	16.3
3	138	2	10.7	5.7	40, 32	0.29, 0.06	9.1, 2.1
4	226	2	10.8	9.4	6, 37	0.24, 0.11	8.9, 8.7
5	227	2	7.6	5.8	26, 8, 1	0.2, 0.11, 0.06	9.1, 3.5, 2.4
6	71	1	11.5	7.4	14, 6	0.22, 0.08	11.3, 2.8
7	291	4	11.1	6.4	15, 21	0.21, 0.1	6.3, 4.5
8	84	1	12.8	6.6	4, 33	0.27, 0.04	15.1, 2
9	154	1	13.6	6	31, 4	0.24, 0.04	7.3, 2
10	174	1	7.5	3.9	3, 33, 38	0.21, 0.13, 0.07	7.2, 6.3, 2
11	59	0	8.4	6.8	36	0.31	19.9
12	120	0	11.3	12.8	19, 21, 24, 36	0.24, 0.06, 0.06, 0.03	9.5, 2.6, 2.2, 2
13	219	0	9	5.6	28, 2	0.12, 0.15	4.5, 4.1
14	318	0	6.7	10.4	18	0.38	6.1
15	260	0	10	8	9	0.29	6.8
16	234	0	9.7	4.6	38, 40	0.26, 0.07	7.8, 2.1
17	59	0	6.5	5.2	27, 35, 7, 5, 20	0.32, 0.07, 0.07, 0.02, 0.06	11.4, 4.4, 3.7, 2.5, 2.3
18	210	0	4.5	5	16, 20, 7, 27	0.4, 0.1, 0.05, 0.06	16.2, 4.2, 2.6, 2.2
19	190	0	9.6	6.5	32	0.21	7.4
20	152	0	6	3.9	8, 23	0.27, 0.04	9, 2.1
21	227	0	11	5.2	29	0.26	9.1
22	167	0	4.3	4.5	25, 8, 5, 20, 22, 35, 23	0.31, 0.07, 0.02, 0.05, 0.09, 0.03, 0.04	11.9, 2.4, 2.2, 2.1, 2.1, 2.1, 2.1
23	132	0	8.8	5	1, 3	0.24, 0.07	9.4, 2.5
24	63	0	9.7	5.6	17, 29, 6, 23	0.53, 0.11, 0.07, 0.04	51.1, 3.7, 2.4, 2.2
25	201	0	12.7	7.7	13, 24, 19	0.3, 0.08, 0.06	9.6, 2.8, 2.5
26	160	0	4.6	4.5	35, 7, 20, 25	0.22, 0.15, 0.08, 0.06	13.7, 8, 3.3, 2.1
27	203	0	12.3	10.7	24, 39, 12	0.21, 0.1, 0.02	7.7, 4.7, 2
28	77	0	9.5	4.5	34	0.26	11.6
29	53	0	6.4	4.9	23, 17, 37, 22	0.3, 0.03, 0.03, 0.09	16.6, 3, 2.6, 2
30	113	0	9.5	7.2	11, 9, 6	0.33, 0.1, 0.05	14.5, 2.3, 2
31	169	0	3.1	3.9	20, 25, 22, 7	0.32, 0.09, 0.12, 0.04	13.2, 3.5, 3, 2
32	218	0	5.5	4	22	0.28	6.6
33	22	0	22.4	18.9	12, 5, 19	0.52, 0.03, 0.06	48.4, 4.1, 2.3
34	22	0	21	19.3	5, 12, 18	0.52, 0.05, 0.25	72.3, 5.1, 4

Appendix G

Identifying different all-electric propositions

Table G.1: Before TSNE

Segment	Total	Prop. B[1]	Prop. C	Prop.D	Prop.A[1]
1	27	3	0	0	24
2	1	0	0	1	0
3	2	0	0	2	0
4	1	0	0	1	0
5	1	0	0	1	0
6	5	0	4	1	0
7	1	0	0	1	0
8	1	0	0	1	0
9	1	0	1	0	0
10	1	0	1	0	0
11	1	0	1	0	0
Total	42	3	7	8	24

[1] Known to have PV installed

Table G.2: After TSNE

Segment	Total	Prop. B[1]	Prop. C	Prop.D	Prop.A[1]
1	27	3	0	0	24
2	1	0	0	1	0
3	2	0	1	1	0
4	2	0	1	1	0
5	2	0	0	2	0
6	1	0	0	1	0
7	4	0	3	1	0
8	1	0	0	1	0
9	1	0	1	0	0
10	1	0	1	0	0
Total	42	3	7	8	24

[1] Has solar panels installed

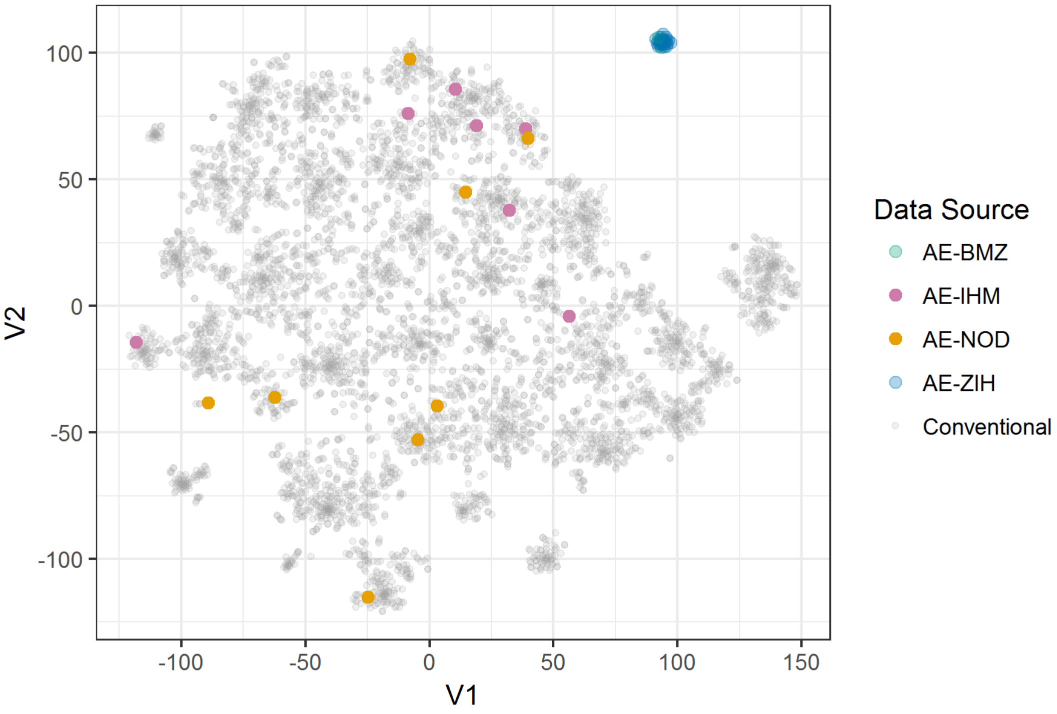


Figure G.1: Plot of all all-electric homes (coloured) and conventional homes (grey) after dimensionality reduction with t-SNE. The plot shows that all-electric homes with PV (blue and green) are clustered together well, while the all-electric homes without PV (purple and orange) are scattered around the data.

R-code

Since...

1. printing unnecessary amounts of paper is bad for the environment,
2. it is much more useful to have actual code than a printed copy of it,
3. people can now potentially collaborate on this project,
4. this gives me one more day to clean up my code, and
5. who looks at these appendices anyway?

I decided to upload the code to *www.bitbucket.org*. If you are interested in having a look at my code out of academic interest, curiosity, or because you are my supervisor who actually wants to check if I implemented all the code myself, please do not hesitate to send an email to `runevandermeijden@gmail.com` and I will give you access to the repository. In order to fill up this space, I have a few "fun facts" for you, as a show of appreciation for the fact that you ploughed through all of this.

Did you know that...

- ... there is a website that teaches your cat how to program in **R**? Cat lovers are also welcome[1] .
- ... *glycophilia* is the scientific term for the compulsive collecting of sugar bags, the ones you get with your cup of coffee? People suffering from glycophilia are called *sucrologists*. The Guinness World Record of the largest sugar package collection belongs to Ralf Schröder, who has collected a total of 14,502 different sugar packets over the past 30 years[2]. However, a fierce debate is currently held in the sucrologists community, since Marianne Dumjahn from Germany claims to have collected a whopping total of 398,572 different sugar packages[3]. This means she has collected more than 35 new sugar packets each day since the moment she started collecting 30 years ago. Amazing!
- ... the Kakapo (*Strigops habroptila*) is an almost extinct bird living in New Zealand. The unfortunate fact is that this bird, in the time it did not have any natural enemies, has forgotten how to fly. Apparently, a seriously worried kakapo will run up a tree, fly like a brick and smash on the ground [113].

[1] See `www.rforcats.net/`.

[2] See `www.guinnessworldrecords.com/world-records/largest-collection-of-sugar-packets`

[3] See `www.alternativerecords.co.uk/recorddetails.asp?recid=108&page=cat`

Outline articles for future publication

This last chapter presents the outline of two articles based on the work in this thesis. The first article presents the latent lifestyle model and fits it to the data. The second article focuses on applications of the model for the DSO. No time was left to work out these papers in more detail. Also, some analysis is still left to be done.

Lifestyle based consumer segmentation

The first article focuses on the main contribution of this research: the development of a latent lifestyle model that capture patterns of electrical energy consumption that seem to resemble lifestyles. Therefore, this article is mainly presenting the work in Chapter 5. The load shapes generated in Chapter 4 are serving as an input to the model and are therefore also included in the article. The sections of this article are:

1. **Introduction**
   The introduction briefly paints the contextual landscape (Section 1.2) and relevant research (Section 2.2). It concludes with the positioning of this research (Section 2.3).
2. **The load shape dictionary**
   In this section, the load shape dictionary is generated based on a stratified and sufficiently large sample. The methodology is adaptive K-means, as developed by Kwac et al. [39, 59]. This section summarises Chapter 4 and presents its results.
3. **Latent lifestyle model**
   In this section, LDA is introduced as a method to discover patterns in collections of discrete data (Section 5.2 and [90]). The analogy with energy consumption modelling is drawn and the underlying assumptions are explained. Furthermore, the statistical model is introduced here, making a link to the load shape dictionary (Section 5.4).
4. **Inference and parameter estimation**
   This section introduces empirical Bayes as the method for making inference on the parameters and latent variables, and variational expectation maximisation as approximation method. Therefore, it is a condensed version of Section 5.3 from this thesis.
5. **Implementation**
   This section elaborates on the pre-processing of the data, the implementation in **R**, and the model choices. The number of lifestyles is determined as was done in Section 5.5.3
6. **Results**
   This section shows some of the results, like the characteristic load shapes per lifestyle, and the lifestyle-per-home distribution, as was done in Section 5.6. The results are briefly discussed.
7. **Conclusions**
   The conclusions in Sections 5.7 and 8.1 are combined in this section..

8. **Discussion and future work**
   The drawbacks of this model are discussed (Section 8.2), after which suggestions for validation (Section 8.3.1), application (Section 8.3.2) and improvement (Section 8.3.3) are given. The part about application links to the second article, by mentioning several potential applications for the grid operator, like monitoring the local energy transition and simulating its impact.

Latent lifestyle model for energy consumption analysis

The second article presents the work in Chapter 6 and Chapter 7, focusing on three practical applications of the model: (1) customer segmentation, (2) monitoring of the local energy transition, and (3) simulating the effect of the energy transition on a local scale.