Karlsruhe Institute of Technology

INSTITUTE OF INFORMATION SYSTEMS AND MARKETING

Analyzing Smart Meter Data to Cluster Households for Energy Community Load Forecasting

Bachelor Thesis

Lucas Bleher

2229078

At the Department of Economics and Management
at the Institute of Information Systems and Marketing (IISM)
Information & Market Engineering

Reviewer:	Prof. Dr. rer. pol. Christof Weinhardt
Second reviewer:	Prof. Dr. Alexander Mädche
Advisor:	M.Sc. Saskia Bluhm

12th of August 2022

Contents

List of Figures iv

List of Tables v

List of Abbreviations vi

1. Introduction 2

2. Literature and Theory 4

2.1. Definitions 4

2.2. State-of-the-Art: Energy Load Forecasting in the context of Renewable Energy Communities (REC) 4

2.2.1. Residential Household Level 5

2.2.2. Community Aggregation Level 6

2.2.3. Summary 7

3. Data Analysis and Methodology 9

3.1. Underlying Data 9

3.2. Data Understanding and Exploratory Data Analysis 9

3.2.1. Households 9

3.2.2. Days and Daily Patterns 11

3.2.3. Months 12

3.3. Methodology Outline 13

4. Clustering 14

4.1. Methodology: Possible Clustering Approaches 14

4.2. Clustering of Daily Profiles 14

4.2.1. Applied Clustering Algorithms 14

4.2.1.1. k-Means 15

4.2.1.2. Time Series k-Means 17

4.2.1.3. DBSCAN 18

4.2.2. Analysis of the Resulting Clusters 19

4.2.2.1. Domain-specific Description and Characteristics of each Cluster 19

4.2.2.2. Detailed Analysis for selected Households 22

5. Forecasting **26**
5.1. Forecast Methodology 26
5.2. Evaluation Metrics 27
5.3. Forecast Scenarios 28
5.3.1. Daily Forecast 28
5.3.2. Hourly Forecast 31
5.4. Conclusion and Comparison 34

6. Conclusion **35**
6.1. Summary and Discussion 35
6.2. Future Work 35

7. Erklärung **37**

Appendix **38**
A. Figures 38

References **38**
References 41

List of Figures

3.1. Different statistical consumption characteristics of the households 10
3.2. Median representation of every hour over 52 weeks for household one . . . 11
3.3. Boxplots of daily hourly consumption for two selected household 12
3.4. Boxplots of hourly consumption per month for two selected households . . . 13

4.1. Silhouette coefficient for different numbers of K 15
4.2. Cluster centroids for different values of clusters K 16
4.3. Centroids of the k-Means algorithm with DTW ($K = 3$) 18
4.4. Cluster proportions for the k-Means algorithm with $K = 3$ 19
4.5. Composition of Cluster 1, regarding households and days of the week 20
4.6. Composition of Cluster 2, regarding households and days of the week 21
4.7. Composition of Cluster 3, regarding households and days of the week 22
4.8. Cluster allocation for two households with high occupancy and consumption 24
4.9. Cluster allocation for two households with high occupancy but low consumption 24
4.10. Cluster allocation for two households with low occupancy 25

5.1. Effect of clustering on the daily forecast performance of SVR measured by NMAE and NRMSE 30
5.2. Effect of clustering on the hourly forecast performance of SVR measured by NMAE and NRMSE 33
5.3. Feature importance in forecasting for each of the four subgroups 34

A.1. Overview of hourly electricity consumption of all 19 households 38
A.2. Overview of hourly heating consumption of all 19 households 38
A.3. Boxplot daily schedule, HH 06 39
A.4. Boxplot consumption per month, all HH 39
A.5. Day of the week distribution of household 2 for selected clusters 40
A.6. Day of the week distribution of household 14 for selected clusters 40
A.7. Day of the week distribution of household 11 for selected clusters 40
A.8. Comparison of cluster results for 100 days and all 365 days (K=3) 40

List of Tables

4.1. Describing statistics for the three cluster centroids 21
4.2. Demographic variables and consumption statistics [kWh] per household . . 23

5.1. Target and feature variables for the daily forecast 28
5.2. SVR daily forecast results for the four forecasting scenarios 29
5.3. RFR daily forecast results for the four forecasting scenarios 30
5.4. Target and feature variables for the hourly forecast 31
5.5. SVR hourly forecast results for the four forecasting scenarios 32
5.6. RFR hourly forecast results for the four forecasting scenarios 32

List of Abbreviations

ANN .... Artificial Neuronal Network

CBD .... Central Business District

CNN .... Convolutional Neuronal Network

DBSCAN .... Density-Based Spatial Clustering of Applications with Noise

DL .... Deep Learning

DTW .... Dynamic Time Warping

HEM .... Home Energy Management

HEMS .... Home Energy Management Systems

kWh .... Kilowatt-Hour

LSTM .... Long Short-Term-Memory

MAE .... Mean Absolute Error

MAPE .... Mean Absolute Percentage Error

NMAE .... Normalized Mean Absolute Error

NRMSE .... Normalized Root Mean Squared Error

OPTICS .... Ordering Points To Identify Clustering Structure

REC .... Renewable Energy Community

RFR .... Random Forest Regressor

RMSE .... Root Mean Squared Error

RMSPE .... Root Mean Squared Percentage Error

SVR .... Support Vector Regression

Abstract

This thesis addresses how to further improve the prediction of volatile residential electricity loads in the context of energy communities. New decentralized local community concepts including renewable energy production are central components of the energy transition. An important component of these concepts is accurate load forecasting. The latter enables operating home energy management systems and thus increases local efficiency and self-consumption. Various cluster approaches exist in the literature to support models by pre-grouping raw input data. In this work, hourly electricity loads for 19 UK households were investigated. Using the k-Means algorithm, repetitive daily profiles were identified across all households. On the one hand, this allowed investigating typical energy consumption patterns at household level. On the other hand, the clustering supported models in predicting more accurately. The effect of clustering daily load profiles on the prediction performance was tested for two time resolutions. Using the Normalized Root Mean Squared Error, Support Vector Regression and the Random Forest Regressor were compared on the different clustered subgroups.
The main result of the work is that different daily profiles could be identified across a set of households. With the help of these results, the Normalized Root Mean Squared Error could be reduced. For the prediction at daily resolution, this effect was strongly pronounced, while at hourly resolution of the residential electricity loads, a slight improvement could be achieved.

1. Introduction

"Local energy systems can potentially contribute to the overall energy and climate objectives, helping reverse energy consumption and emissions trends worldwide."
– Koirala et al. (2016)

The World Resources Institute (WRI) considers the energy supply of the future and a societal transformation towards more sustainable energy concepts to be among the greatest and most urgent challenges of our time (World Resources Institute, 2018). The tense situation in the energy sector due to climate change was recently further exacerbated by geopolitical conflicts. This clarified once again the great dependence on fossil fuels and their main exporters. That is why, currently many nations are facing major challenges in securing their future energy supply.
In contrast to fossil fuels, renewable energy based on wind, water, or solar energy, for example, are available in almost unlimited quantities, regardless of the amount used. Moreover, they can often be used locally and only have to be transported over short distances. (Etapart AG, n.d.)
It is precisely at this point of local energy production that local district concepts come into the picture. The EU emphasizes that energy communities are an important part of the aimed energy transition until 2050 and that local actors play a crucial role in the transition process. (European Commission, n.d.) A precise definition of Renewable Energy Communities (REC) will be given in Chapter 2.
The "State of the Energy Union report" of the European Commission from October 2021 confirms the enormous potential contribution of local actors and energy communities: At the moment, more than 8400 energy communities, with at least two million participating citizens, exist in the EU. The capacities of these projects led by citizens were estimated to contribute with up to 7% to the national energy production. With about 50%, the largest share was generated by solar photovoltaic systems. (Schwanitz et al., 2021)

At the same time, advances in technology, especially the increasingly widespread smart meters, are shifting the focus from larger aggregation levels to individual households. In this context, applications such as Home Energy Management Systems (HEMS), which optimize energy consumption together with self-production, play an important role. Currently, the main goal of the HEMS is to plan and allocate the next day's household energy operations. This involves taking into account weather conditions and the associated expected electricity generation as well as future electricity demand. The latter depends strongly on the behaviour of the inhabitants. (Yildiz et al., 2018a)
Therefore, accurate predictions of residential load profiles are important for efficient HEMS (Rodrigues et al., 2022). Due to the great uncertainty of the drivers of short-term load profiles at household level, this is a highly topical research field (Jiang et al., 2021). An important component is to understand potential patterns in electricity consumption at the household level. In this thesis, this was addressed by means of a statistical data analysis as well as a detailed cluster analysis. In addition, the pre-grouping of load profiles may

assist models in the challenging prediction. Different cluster approaches, which will be discussed in Chapter 2, have already been applied in literature. This work builds on these results and applies clustering in a modified scenario.
To summarize, the following research questions were investigated in this thesis:

RQ1. Which general patterns of electricity usage can be observed on household level?

RQ2. Does and if how much does clustering improve the forecast performance for the given dataset?

The following work is structured as follows. First, in Chapter 2, relevant literature from the field and its relation to this work are considered. In Chapter 3, the dataset is described, together with a short exploratory data analysis and a methodology outline. Chapter 4 presents the exact clustering approach and results, before Chapter 5 builds on these findings with the prediction models. Finally, Chapter 6 closes the thesis with a concluding discussion as well as future work steps.

2. Literature and Theory

2.1. Definitions

Smart Meter: The term smart meter appeared when microprocessors first were included in static meters. It describes an electronic meter which can collect information on various energy consumption characteristics. Smart meters bring advantages for both customers and electricity suppliers. An important feature is real-time energy recording and the report of this data in regular time intervals. (Koponen et al., 2008)

Renewable Energy Community (REC): The European Union defines a Renewable Energy Community in its final "Clean Energy Package" as a legal entity in accordance with the corresponding national law. It is autonomous, open and voluntary for participation and run by the concerning shareholders. These are living in direct proximity of the renewable energy project for which the legal entity is responsible. The primary focus is not on financial profits but rather on "environmental, economic and social community benefits for its shareholder or members" [p. 5]. (Jeriha, 2019)

2.2. State-of-the-Art: Energy Load Forecasting in the context of Renewable Energy Communities (REC)

In this section, approaches and concepts presented so far in literature are discussed. Moreover, there will be a comparison of obtained results. It is important to distinguish between energy load forecasting in general and load forecasting for Renewable Energy Communities (REC). The focus of this thesis and state-of-the-art presentation is on load analysis and prediction for households in energy communities. RECs have different characteristics and hold other premises than for example Central Business Districts (CBD) or normal power grids (Pirbazari et al., 2021). While power grids operate on a very high aggregation level, RECs are a local concept and rely on interaction with the power grid to operate. CBDs are often much bigger than RECs and therefore balancing effects due to a higher aggregation level can be used (Xu et al., 2017). Furthermore, consumption patterns in the commercial and industrial context are more regular than highly individual and volatile occupancy behaviour (EIA U.S. Government, 2013). Therefore, more and more recent scientific literature is focusing on residential energy communities and forecasting models in their context. However, this analysis also includes publications which involve forecasting for individual households beyond the REC context (e.g. Lusis et al. (2017)). These are just as relevant as findings for stand-alone households also bring interesting insights about household load profiles in the REC context. Therefore, this section will be divided in two parts. First, publications with forecast models on residential household level are considered. Then, literature evaluating forecasting approaches on REC aggregation levels is reviewed. Here, it is distinguished between different sizes of RECs, which are being considered, ranging from a few to hundreds of participating households. However, most of the studies involve much bigger communities than the dataset on hand with 19 households, which is going to be presented later in this study. In general, one should be aware of the big differences between forecasting on individual household and community aggregation level. Due to high volatility and individual characteristics, predictions on household level are much more challenging (Wijaya et al., 2015; Lusis et al., 2017). While older publications are focusing mainly on higher aggregation levels, smart meter based analyses on household level are currently becoming more topical. That is why several recent publications focus on individual household load forecasting for short-term horizons (Jiang et al., 2021). The most common predictor variables introduced in literature are historical load data collected regularly by smart meters, weather and calendar information (e.g. Lusis et al. (2017); Pirbazari et al. (2021)). However, certain studies (e.g. Rodrigues et al. (2022)) did not include exogenous factors in the models and relied completely on historical load data.

2.2.1. Residential Household Level

Due to the increased challenges when forecasting volatile household loads, past studies presented different tools to assist prediction models and gain more insights into driving factors. For example, Humeau et al. (2013) proposed to cluster households according to their energy consumption profiles by applying the k-Means algorithm on 24-dimensional representations of each household. Each dimension stands for the overall average load of the concerning household at the corresponding hour. This idea was further pursued by Yildiz et al. (2018b) who also recommended to include a clustering step before forecasting. By grouping households according to their standard deviation, the authors were able to reduce the Root Mean Squared Percentage Error (RMSPE) per household by 9% on average (Yildiz et al., 2018b).
Shortly afterwards, the same authors introduced "Cluster-Classify-Forecast" (CCF) as a new approach which extends simple Smart Meter based Models (SMBD) by adding information obtained from clustering and classifying daily energy consumption profiles within a household. The increased model complexity was justified by better performance. Another advantage of this approach is the insight into driving components and feature importance for individual load profiles. (Yildiz et al., 2018a)
Dinesh et al. (2019) were performing graph spectral clustering on the very low aggregation level of appliance signals, which were aggregated later to household or community level.

Apart from the positive effects of pre-clustering, Yildiz et al. (2018b) found a considerable impact of data resolution and the forecast horizon on the forecast performance. While Yildiz et al. (2018b) reported the best results for one-hour-ahead forecasting, Lusis et al. (2017) found lower errors for coarser forecast granularities. This is possible due to different raw resolutions in the original dataset. Several studies included calendar information in their models (e.g. Humeau et al. (2013); Yildiz et al. (2018b); Tits et al. (2020)). However, Lusis et al. (2017) reported that the main predictors remain historical load and weather data, as the effect of calendar information included as dummy variables remains little.

Differing from the publications presented before, Aurangzeb (2019) tested 8 different regression models for single household energy consumption in the REC context without clustering households beforehand. For models to learn the highly irregular and complex (sometimes unpredictable) behaviour of single households, more complex non-linear models are necessary (Jiang et al., 2021). This suggestion was backed up by Aurangzeb (2019), who concluded in his study that the non-linear Radial Basis Function (RBF) kernel delivers the best results, compared to other regression models for individual households.

As the domain of Artificial Neuronal Networks (ANN) and Deep Learning (DL) is evolving fast, recent publications were applying these topical techniques in the household load forecasting context (e.g. Jiang et al. (2021); Rodrigues et al. (2022)).
Jiang et al. (2021) used several recent DL architectures on household level. Compared to other models, the latter seem to be more efficient in capturing more uncertain and volatile household load profiles. By combining Convolutional Neuronal Networks (CNN) and Long Short-Term-Memory (LSTM) networks, not only regular consumption behaviour, but also recent characteristics such as short-term abnormalities or shared behaviour patterns across different households could be learned. (Jiang et al., 2021)
Rodrigues et al. (2022) trained ANNs with historical load data to predict load curves of households several hours ahead. A special focus was given to model performance comparison between weekdays and weekend. The authors underlined that the demand differs according to the day of the week. Model performance was slightly worse for weekends.

Other studies took a more holistic approach and highlighted the importance to see individual households as part of RECs or micro-grids (e.g. Tits et al. (2020); Hou et al. (2021); Gong et al. (2021)). In the following section, approaches on these higher aggregation levels are considered.

2.2.2. Community Aggregation Level

A considerable amount of literature also made use of different clustering approaches on community aggregation level (e.g. Wijaya et al. (2015); Flor et al. (2021); Hou et al. (2021)).
Wijaya et al. (2015) combined the advantages of two aggregate forecasting scenarios into an approach called Cluster-Based Aggregate Forecasting (CBAF). The study of Flor et al. (2021) showed how energy behaviour patterns in nearby areas are connected and can be organized into geographical zones to improve forecast performance in each sub-cluster. Besides the classical k-Means algorithm, the Ordering Points To Identify Clustering Structure (OPTICS) algorithm was recently used for power consumption pattern recognition (Hou et al., 2021).

While pattern recognition gets easier and more cyclical on higher aggregation levels, considering whole communities or ensembles of households poses new questions, like determining the community size and respectively the aggregation level (Hou et al., 2021).
Recently, a number of researchers have sought to determine the adequate number of households included in a micro-grid to provide sufficient accurate results for Home Energy Man-

agement (HEM). HEM enables scheduling future load needs to better allocate demand and renewable energy offer. With respect to the Mean Absolute Error Percentage (MAPE), Hou et al. (2021) determined 150 households as sufficient (MAPE below 10%) size. In the recent study by Shaqour et al. (2022) this critical error of 10% was already achieved for an aggregation size of 30 dwellings. A household microgrid meeting these requirements brings advantages compared to single households: The aggregation and cooperation results in a more flexible distribution of the residential electricity and better possibilities to meet the demand efficiently with renewable energies. Finally, this leads to better performance of local energy management (Shaqour et al., 2022).
Previous studies have already highlighted the important effects of REC characteristics (e.g. community size) on the predictability (Tits et al., 2020). It is important to retain that the performance of certain models is highly dependent on the underlying aggregation size (Humeau et al., 2013). Especially for the more challenging lower aggregation levels, model choice is a critical factor (Burg et al., 2021).
However, model choices only driven by quality metrics like the MAPE may be misleading. In practice the added value (e.g. self-sufficiency, cost, carbon footprint) for the community is most important. (Coignard et al., 2021)

Another challenge for RECs is HEM and the efficient energy allocation for households and communities. Gong et al. (2021) used long term load forecasting combined with HEM to reduce consumption peaks and the total load consumption. This was done by load shifting, to efficiently control electric water heaters (EWH) and heating, ventilation, and air-conditioning (HVAC) systems. The same holistic view also dominated the recent study of Pirbazari et al. (2021), which compared solar output and community overall energy usage prediction for 6 equal-sized household communities.

As in Section 2.2.1, recent literature focusing on community aggregation level also makes use of the fast development of deep learning (e.g. (Pirbazari et al., 2021; Hou et al., 2021; Shaqour et al., 2022)).

2.2.3. Summary

The two preceding sections showed similarities and differences between the two perspectives that can be taken when performing energy load forecasting in RECs. The most important point might be the high volatility at household level, which is smoothed out with increasing community aggregation size. Several different models were introduced in literature so far. From classical models like Multiple Linear Regression (Humeau et al., 2013) or Regression Trees (Lusis et al., 2017) over Support Vector Regression (e.g. Humeau et al. (2013); Wijaya et al. (2015)) to highly complex DL architectures (e.g. Hou et al. (2021); Shaqour et al. (2022)), models were tested in different scenarios.

In this thesis, the given dataset (Pullinger et al., 2021) is used to review existing findings on energy load forecasting in the REC context, while also exploring new potential of pre-

clustering of smart meter data for forecast improvement.
An important step is the analysis of the 19 households of the dataset regarding consumer profiles and consumption patterns. Part of the analysis is the application of clustering algorithms to divide daily consumption profiles among all households into groups. In a second step, prediction performance for these defined subgroups of households is evaluated. The hypothesis is that a pre-grouping of the consumption profiles according to their characteristics will support the models in their prediction.

3. Data Analysis and Methodology

3.1. Underlying Data

The given dataset for electricity and heating consumption of households originates from the IDEAL Household Energy Dataset (Pullinger et al., 2021). This original dataset includes raw gas and electricity data for 255 UK households at sensor level. It was internally modified and prepossessed at the Institute of Information Systems and Marketing (IISM). Due to varying observation periods, 19 households, whose time series data overlapped for one year, were chosen. This allowed to analyse a uniform and comparable observation period from 15.06.2017 to 14.06.2018. During this process, outliers and missing values were treated appropriately. Unnaturally high consumption peaks due to transmission problems from the sensors were apportioned to previous time steps without recorded values. The total consumption was unaffected by this.
The energy consumption data used in this study is available in a comma-separated values (CSV) file containing hourly electricity consumption data. The consumption was measured in kilowatt-hours (kWh) and covers a whole year of hourly observations. Each data point refers to the consumption at the corresponding hour and date.
For further analysis, demographic variables, like the number of residents or occupied days and nights of each housing unit, were drawn from the IDEAL dataset.
Weather data was taken from `www.renewables.ninja`, which provides several meteorological variables for Edinburgh (Scotland) in hourly resolution. Since all the selected households are either directly from Edinburgh or the neighbouring Midlothian, the data for Edinburgh was considered representative of the entire data set[1]. These weather variables were considered for the forecasting model and in the explanatory data analysis.
Moreover, calendar and temporal information were integrated via the Python *datetime* module.

3.2. Data Understanding and Exploratory Data Analysis

This section presents first insights obtained from exploratory data analysis which are relevant for the following research process. Furthermore, general patterns of energy use on household level are analysed. Essentially, a distinction between three levels is made here: 1) households, 2) daily schedules and 3) months.

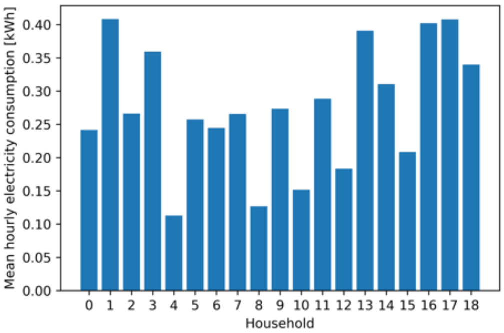
3.2.1. Households

As introduced in Section 2.2, occupancy behaviour and the resulting energy consumption are highly volatile and individual. Figure 3.1 shows mean, overall and maximum electricity consumption per household and confirms varying household summary statistics for the

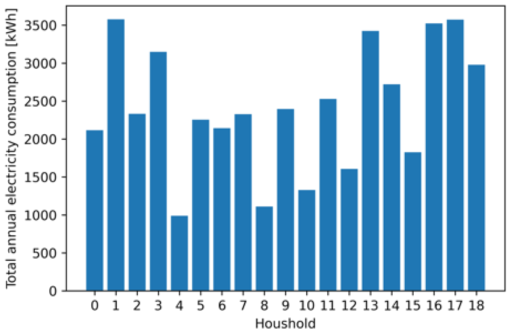
[1] The air-line distance between the two scottish cities is only 9.49 miles (`www.distance.to`).

given dataset. For example, the mean hourly electricity consumption for the observation period of one year is in the range of [0.11 kWh, 0.41 kWh] (Figure 3.1a). Naturally, the same trend can be seen for mean and total electricity consumption. The latter also differs considerably. While household 5 only needs 988 kWh per year, households 2, 17 and 18 consume more than three times as much, with up to 3577 kWh per year (Figure 3.1b). However, the biggest differences can be observed in the maximum hourly peaks. This shows that consumption for households, such as 11 and 17, is highly concentrated at individual times of the day (Figure 3.1c). Moreover, Figure 3.1d summarizes several descriptive statistics for each household in a boxplot. Significantly varying standard deviation of the hourly electricity load can be observed. As introduced in the state-of-the-art (cf. Section 2.2), this is one of the biggest challenges when forecasting residential electricity loads. This aspect is also demonstrated by the individual statistical characteristics for each household.

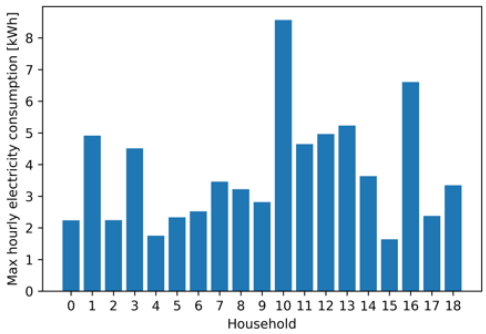
Nevertheless, the question arises how models still can be formed and how they can be assisted in forecasting. This thesis aims to do so by clustering similar daily consumption profiles. The approach is presented in Chapter 4.



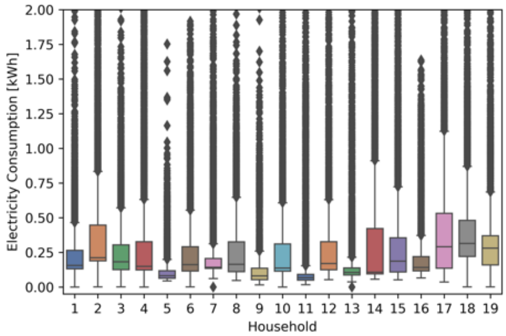
(a) Hourly mean



(b) Total consumption



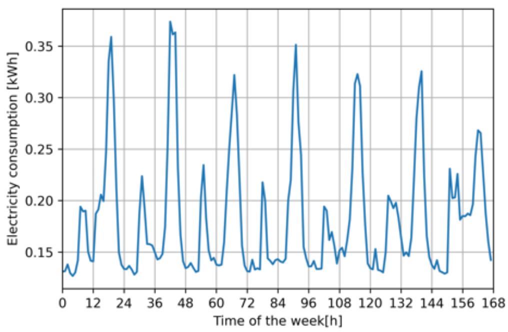
(c) Max hourly consumption



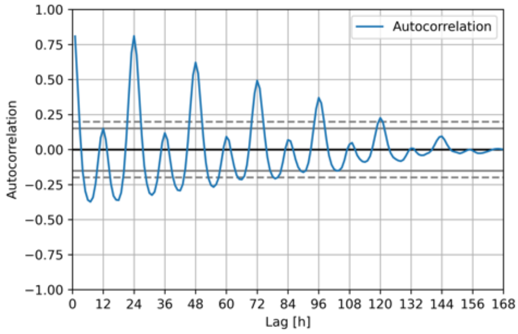
(d) Boxplot: Consumption per Household

Figure 3.1.: Different statistical consumption characteristics of the households

Literature recognizes a high correlation between future household energy usage and historical consumption data (e.g. Pirbazari et al. (2021)). This relationship could be confirmed for the dataset on hand. In order to prove this, observations were divided into weeks, i.e., into time frames of 168 hours each. To compensate for the effect of outstanding weeks, the median of the electricity consumption at the respective hour of the week was selected



(a) Aggregated weekly pattern



(b) Autocorrelation in a week

Figure 3.2.: Median representation of every hour over 52 weeks for household one

over all 52 weeks of the year (Figure 3.2a).

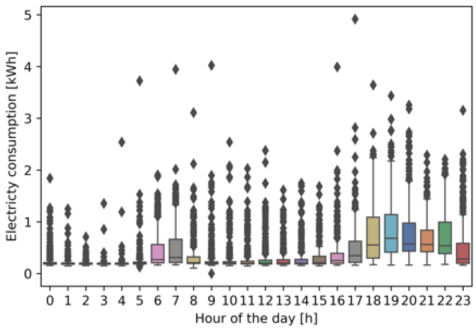
Autocorrelation was used here to measure the correlation of a data point with the preceding observation x hours before. Following the smoothed weekly representation, the previous 168 hours, i.e. the previous week, were taken into account. While the autocorrelation plot for particular stand-alone weeks is highly dependent on the choice of week and household, the mentioned median representation provides a clear and stable plot. Naturally, the exact shape of the autocorrelation plot differs slightly from household to household, depending on the regularity of occupancy behaviour.

Figure 3.2b shows an example of the autocorrelation for household one of the dataset. The course of the plot for this household can be considered representative for the other households, since the general trend concerning autocorrelation is very similar. A periodic trend with a consistently decreasing amplitude can clearly be observed. That means more recent lagged data points are stronger correlated with each other. The blue line indicates autocorrelation, whilst the horizontal lines represent the 95% (solid) and 99% (dashed) confidence interval. In other words, within the horizontal lines the probability for the true value of the given sample to lie in the marked interval is 95% respectively 99% (Cox & Hinkley, 1979).

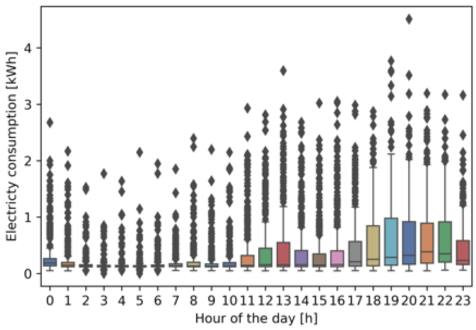
The positive autocorrelation posses its peak for the lags at $hour - (24 \times X)$ (Figure 3.2b). This indicates a high relevance of the consumption at the same hour on the previous days for the future consumption. In other words, from the historical consumption, especially 24 hours before, variables can be created for the model with potentially high explanatory power. Depending on the household, the lagged data points of the three to four previous days are particularly highly correlated. However, one needs to be aware that the peaks are no longer part of the confidence intervals. Nevertheless, within the intervals we can still see relevant autocorrelation. Interestingly, also the 12-hour cycle may be relevant and is considered when performing feature engineering for the models (Section 5.3).

3.2.2. Days and Daily Patterns

While Section 3.2.1 presented differences between households at a higher level, this section focuses on patterns within a day on household level. Two selected households are presented as representatives. A classical electricity consumption pattern often has a medium morning peak and a clear peak in the late afternoon and evening (e.g. Pirbazari et al. (2021)). This can be explained by the regular routines of our everyday lives and periodic business hours. The aggregated hourly electricity consumption across all 19 households of the dataset confirmed this trend (Figure 3.2). A load peak was observed around 7 a.m. and in the evening hours, with slightly lower consumption during the day. The trend could be even more distinct for bigger datasets due to the smoothing and aggregation effects mentioned in Section 2.2. However, on individual household level one can see differing daily profiles depending on the individual routines of each household. Figure 3.3 shows boxplots for two selected households demonstrating their daily electricity consumption patterns over a year. On the left, the daily consumption of household 2 is shown, which can be interpreted as a typical profile for a working household (Figure 3.3a). The timing of the morning and evening peak is similar as in the aggregated version over all households. On the other hand, Figure 3.3b demonstrates a completely different daily schedule. The first main peak can be observed at 1 p.m. Examining all households, several individual schedules can be observed. For example, the daily schedule of household six is shifted forward by a few hours, with a morning peak at around 5 a.m. and the second peak at 4 p.m. (Figure A.3). A component which is smoothed out by the aggregation done here, is that each household itself can possess several daily schedules according to calendrical or other circumstances. This aspect is taken into account later on in this study when clustering daily consumption profiles across all households.



(a) Boxplot daily schedule, HH 02

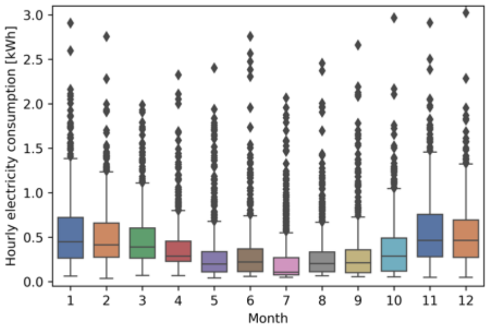


(b) Boxplot daily schedule, HH 04

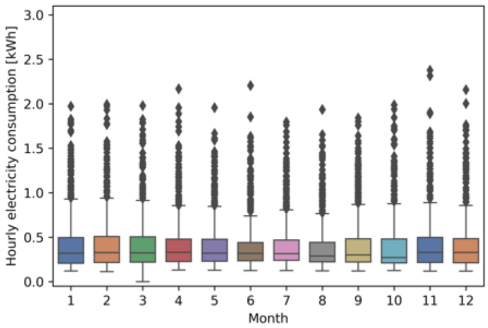
Figure 3.3.: Boxplots of daily hourly consumption for two selected household

3.2.3. Months

From looking at individual days in Section 3.2.2 and weeks in the course of autocorrelation analysis in Section 3.2.1, the perspective is now changed to months. Literature often documented seasonality for electricity consumption. For example, Lee et al. (2014) found increased electricity consumption during Australia's heating and cooling season. For the given dataset once again the situation is different depending on the individual household. In Figure 3.4a we see typical seasonality for household 17. During the winter months, the hourly electricity consumption is slightly elevated. The underlying reasons cannot be explained with the dataset on hand. However, one can imagine the use of electrical



(a) Boxplot consumption per month, HH 17



(b) Boxplot consumption per month, HH 18

Figure 3.4.: Boxplots of hourly consumption per month for two selected households

heating. Contrary to meteorological conditions in Australia, cooling is mostly not needed during summer in Scotland. Therefore, a comparison of the results for these regions is difficult. Interestingly, other households, like household 18 in Figure 3.4b, show nearly no seasonality at all with constant electricity consumption over all 12 months. A possible reason could be the absence of electrical heating and instead the use of gas for heating purposes. At the aggregation level across all households, marginally lower consumption can be observed in the summer months (Figure A.4). However, there is no pronounced seasonality. This initial analysis suggests that meteorological variables such as the current temperature play a minor role in the prediction of electricity consumption for the present dataset. If the heating consumption data is considered for comparison, there is clearly more seasonality. The exploratory data analysis shows a typical "U-shape" for nearly all households (Figure A.2). In the cold months at the beginning and end of the year, heating loads are significantly higher than in summer. However, the detailed analysis of supplementary heating consumption data was beyond the scope of this thesis since the focus was on electricity consumption data.

3.3. Methodology Outline

The detailed methodology is going to be described in the following two chapters Clustering (Chapter 4) and Forecasting (Chapter 5) together with the associated results. However, this section aims to provide the reader with a quick methodological overview for better understanding. As introduced, one of the two addressed research questions was the following: Does and if how much does clustering improve the forecast performance for the given dataset? Therefore, the research process consisted of two main parts. The first important step was to cluster daily electricity load profiles over all 19 households. This included a detailed analysis of the formed clusters together with socio-demographic information about the households. The second important step, which built on the clustering, was the application of forecast models to the different subgroups of daily profiles. Finally, a comparison between the forecast performance of an unspecific model for all daily profiles and a specific forecast model for each cluster respectively each type of day consumption profile was done.

4. Clustering

4.1. Methodology: Possible Clustering Approaches

In literature, different approaches to perform clustering in the energy load forecasting context exist. For example, Humeau et al. (2013) clustered households themselves to apply forecasting to the formed subgroups of households. Yildiz et al. (2018a) clustered daily profiles within a household and fitted a model for each cluster, that is several models per household adapted to the specific cluster. In this thesis, clustering was applied in a different setting with the aim to obtain more generalized models that may be helpful for electricity forecasting in the REC context. Instead of considering days for each household separately, the idea was to obtain typical daily profiles across all 19 households of the dataset. The data was restructured to represent each of the 365 daily load profiles for each of the 19 households. One day is characterized by 24 hourly electricity consumption values. Consequently, the input for the clustering were $365 \times 19 = 6935$ daily profiles. Due to the 24 - dimensional representation of each day, the input data frame for the clustering algorithms possesses 24 columns. The idea behind this approach is that typical daily consumption profiles may appear repeatedly amongst different households, but on different days. The goal of the clustering was to identify K daily profiles among all 6935 days of the dataset and form subgroups of similar days regarding the electricity consumption profile.

4.2. Clustering of Daily Profiles

This section presents the tested clustering algorithms, a detailed analysis of the resulting clusters and the subsequent hypothesis how the clustering results may be used to improve forecast performance in Chapter 5.

4.2.1. Applied Clustering Algorithms

In total, three different clustering algorithms were applied to the mentioned representation of the dataset. In the following, the results for each of them is briefly described. However, the focus will be on the k-Means clustering algorithm, which proved to be the most suitable for the given problem and was used several times in literature for the previously mentioned similar scenarios (e.g. Humeau et al. (2013); Wijaya et al. (2015); Flor et al. (2021)). The focus is on the application of the algorithms to the given problem and the analysis of the results. Mathematical explanations of the algorithms exist numerous times in the literature and are therefore not part of this work.

Data standardization with different scalers like the *StandardScaler* and the *MinMaxScaler* was performed. The *StandardScaler* transforms the data to have mean zero and unit variance, while the *MinMaxScaler* maps the features into the range [0,1] by a linear transformation (Han et al., 2012, pp. 113 - 115). Furthermore, data reduction with principal component analysis was tested. However, the application of the algorithms on the original data representation delivered the best results. The plots for the resulted cluster representations did not vary significantly for different pre-processing settings.

4.2.1.1. k-Means

The well-known and classical k-Means clustering algorithm (MacQueen, 1967) was applied to the daily profile representation of the hourly electricity consumption data described in Section 3.3. For the implementation in Python the machine learning library *scikit-learn* was used (Pedregosa et al., 2011).
A critical point is the parameter K which determines the number of clusters in the data and needs to be specified in advance. Even though guidelines, such as the elbow method and the silhouette coefficient exist, the choice of K for real data is typically never simple and can be ambiguous. Therefore, James et al. (2021, pp. 530 - 532) suggest to try different settings and look for the solution providing the best interpretation of the data. Note that the main goal of clustering is revealing interesting aspects and structures of the data for further implications. (James et al., 2021, pp. 516 - 532)
In this thesis, the goal was to expose information hidden in different daily load consumption profiles to assist forecasting models. Following this argumentation, the silhouette coefficient but mainly the interpretation of the resulting clusters were considered for determining the number of clusters K for the given dataset.
A big advantage of the k-Means algorithm over others, like for example the Density-Based Spatial Clustering of Applications with Noise (DBSCAN), is the possibility to plot the cluster centers vividly (c.f. Section 4.2.1.3). This is particularly interesting for the data representation used here, as the 24-dimensional cluster centroids correspond to the 24 hours of a day. Figure 4.2 shows the clustering results for different values of K, ranging from three to six, while Figure 4.1 presents the silhouette score for values of K ranging from two to 12.

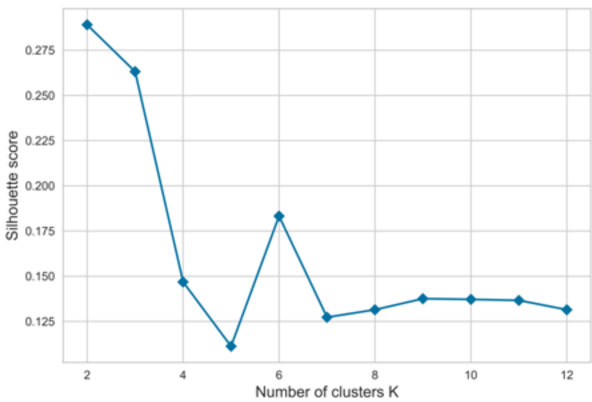
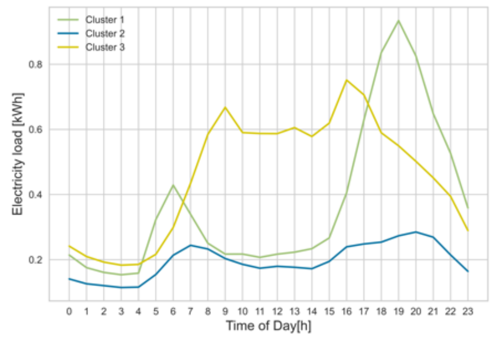
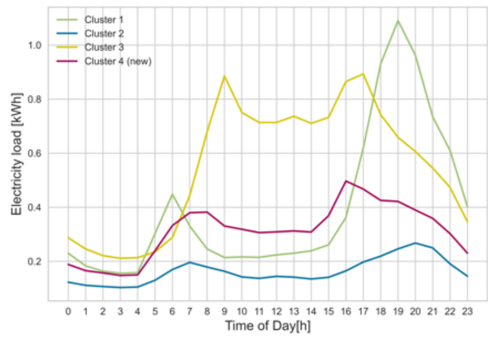


Figure 4.1.: Silhouette coefficient for different numbers of K of the k-Means

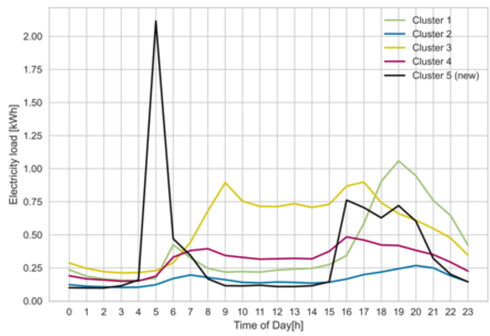
The silhouette coefficient measures the clustering quality by the following equation: $s = \frac{(b-a)}{max(a,b)}$. While a measures the distance of data points within a cluster, b indicates distance between clusters. It is desirable to minimize intra-cluster distance a and maximize inter-cluster distance b at the same time. The score can take values in the interval of [-1,1], where 1 is the best. A negative silhouette coefficient indicates that data points were assigned to wrong clusters. (Belyadi & Haghighat, 2021, pp. 125-168)
With a silhouette coefficient of c = 0.2639 (K = 3), deviating significantly from zero and not being negative, there is potential for clustering structures to be found. (Sari, 2016)
The plot of the silhouette coefficient in Figure 4.1 demonstrates a strongly decreasing score starting from K = 3 upwards. After a small increase for K = 6, the score remains mostly constant. Therefore, a detailed analysis of the clustering results for K's in the range [3,6] was done. This allowed to make an informed decision and to have several decision criteria for the optimal parameter choice. In the course of the analysis, K = 2 was not found to provide a satisfactory solution to be interpreted despite the high silhouette coefficient. This demonstrates very well why it is important to include a domain-specific analysis of the resulting clusters in addition to classical metrics such as the silhouette score. In the course of this, it also became apparent that a higher number of clusters is not useful and usually leads to an over-specification of the clusters. Part of this analysis can be seen in Figure 4.2.



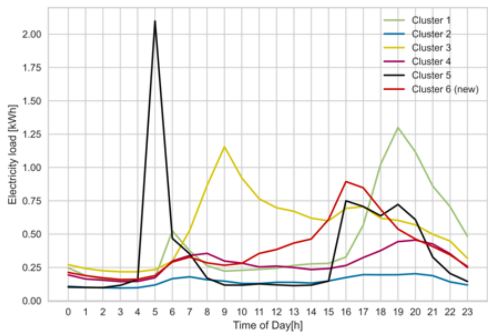
(a) K = 3



(b) K = 4



(c) K = 5



(d) K = 6

Figure 4.2.: Cluster centroids for different values of clusters K

Figure 4.2a shows the plot of the centroids for the configuration with K = 3 which deliv-

ered the most interpretable and coherent solution. In the remaining three subplots, it is apparent how the new clusters formed as the number of clusters was increased. During the transition from K = 3 to K = 4, the three existing clusters change only insignificantly. More importantly, the new cluster does not add significantly more valuable information to the plot. It is mainly a further specification of the second cluster, with slightly higher consumption and a marginally earlier peak in the late afternoon. However, the general trend with two peaks and their timing is the same. Hence, an expansion to K = 4 clusters could not be justified either by the clearly declining silhouette score or by the domain-specific interpretation of the centroids.
The transition to K = 5 clusters is shown in Figure 4.2c. On the one hand, the course of the daily electricity consumption shows a clear distinction from the previous clusters for the early morning hours: The first peak of more than 2 kWh happens very early and concentrated at 5 a.m. On the other hand, a closer look reveals that 94.0% of the daily profiles, within the newly formed fifth cluster, come from household 14. This means that the new cluster was specifically adapted to this outstanding household. However, the aim of clustering in this work was to map general daily consumption profiles across all households (e.g. for a REC) and not to make adjustments for individual households (cf. Section 4.1).
The next transition to K = 6 shows a splitting of the former cluster 3 (Figure 4.2d). In cluster 6 (red), the original daily course is now differentiated into two peaks, one in the late morning and one in the afternoon. Again, this further breakdown is not desired as a daily profile may consist of a combination of several peaks during the day. An overspecification of the clusters to individual households is undesirable and therefore K = 6 was ruled out.

Following this argumentation, it is clear why K = 3 as number of clusters of daily profiles was chosen. The remaining part of the thesis builds on these clustering results. In Section 4.2.2 both the clusters themselves and their composition are analysed in detail.

4.2.1.2. Time Series k-Means

The time series k-Means is a promising alternative to the classical k-Means algorithm, because it takes into account time series specific characteristics which are not considered in the regular algorithm. An implementation of the algorithm is available in the *tslearn* time series software library (Tavenard et al., 2020). The algorithm specified for time series data comes with the possibility to use Dynamic Time Warping (DTW) as distance measure instead of the well established euclidean distance.
DTW originates from speech recognition and is an advanced technique to compare time series sequences to each other. Thanks to a non-linear warping it is possible to match and compare sequences meaningful, even if they vary in speed, length or have differing starting points. (Müller, 2007)
The k-Means algorithm in combination with DTW was already used for clustering residential power load profiles together with spatial analysis by Flor et al. (2021). That is why it was also considered in this study and compared to the classical k-Means approach presented in Section 4.2.1.2. Figure 4.3 presents the centroids for the time series k-Means

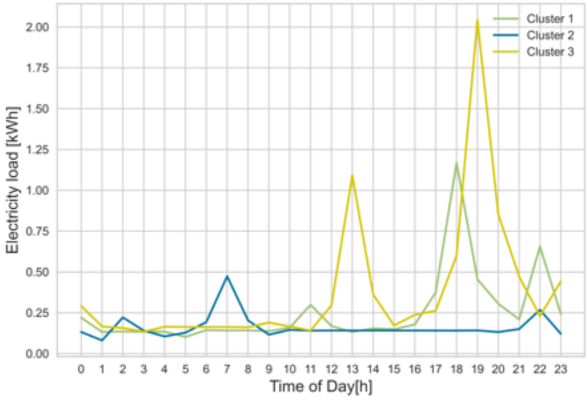


Figure 4.3.: Centroids of the k-Means algorithm with DTW ($K = 3$)

with DTW as distance measure between daily load profiles.
At first glance, the courses of the clusters are far less clear and unambiguous than the results obtained for the classical k-Means algorithm. Nevertheless, some analogies between the clusters can be identified. For better comparability, the same colours were used as in the corresponding Figure 4.2a for clustering with the euclidean distance as distance measure. The first cluster (green) now possesses three peaks, that differ in magnitude. Cluster two (blue) is at a similar level as before, but the peaks are partly more pronounced and deviate somewhat in time. The greatest change can be seen for cluster 3 (yellow). Instead of an increased consumption throughout the day, two peaks can be identified, the magnitude of the second being more than twice as large as before.
The silhouette scores for the clusters obtained by the DTW time series approach were generally significantly worse than in the first approach. For example, for K = 3 the silhouette coefficient for the time series k-Means was only c = 0.1157, whereas it was c = 0.2639 (Figure 4.1) in the initial setting. This was confirmed by a domain-specific analysis of the clusters.
In general, the clusters formed in the initial setting provided a better interpretation of the data. That is why an in-depth evaluation of the clusters was performed for the classical k-Means results and the DTW setting was not further pursued. These results are presented in Section 4.2.2.

4.2.1.3. DBSCAN

The DBSCAN clustering algorithm shall only be mentioned briefly for the sake of completeness, as it turned out to be not suitable for the given problem. The most critical parameter of DBSCAN is *epsilon*, which determines the distance up to which two data points are still considered to lie in the same neighbourhood. It needs to be adapted for each dataset manually. In this study, several parameter configurations were tried, together with the approach to use the k-Nearest neighbour algorithm to get a good order of magnitude for the parameter *epsilon* (e.g. used by Toshniwal et al. (2020)). However, independently of different configurations, the algorithm did not provide meaningful clusters which could

have been used for the described research process of the thesis.
DBSCAN is a density-based clustering algorithm. This type of clustering forms groups out of the data by comparing the density of data points. Typically, observations in regions with low density are labelled as outliers. (Sander, 2010)
This proved to be problematic in the context of electricity load forecasting. Dramatically changing consumption upon different hours of the day or week are as normal as intensive peaks followed by sudden consumption troughs at varying and hardly predictable time (Hou et al., 2021). It does not make sense to declare these peak loads, which often are part of a normal consumption profile, as outliers simply because they may appear in regions with lower density. Unfortunately, for the dataset on hand the DBSCAN algorithm detected a lot of unreal outliers and formed one cluster with high density for the rest of the data points. Moreover, groups formed in addition to this main cluster contained only very few observations (2 - 10 data points). As described in Section 3.1, appropriate preprocessing was applied to the data so that outliers had already been eliminated previously. Consequently, the results of DBSCAN could not further be used in this study.

4.2.2. Analysis of the Resulting Clusters

4.2.2.1. Domain-specific Description and Characteristics of each Cluster

In this section, a closer look at the three resulting clusters is taken and conclusions about the corresponding household electricity consumption are made. Figure 4.2a presents the daily profiles of the final clusters. The associated cluster proportions are shown in Figure 4.4. With a percentage of 63.48%, cluster 2 has the highest number of daily profiles out of a total of 6935 days. Cluster 1 and cluster 3 share the remaining data points relatively evenly with 18.73% and 17.79% of all data points, respectively.

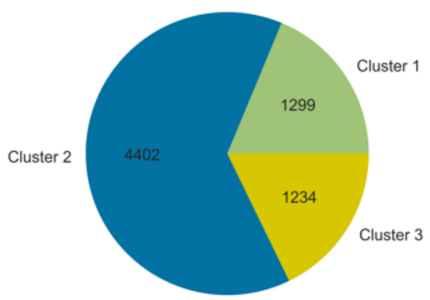
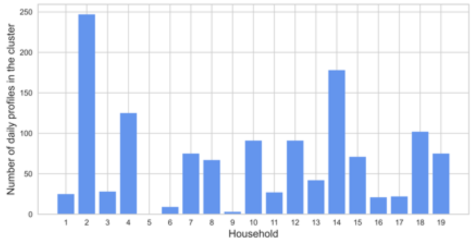
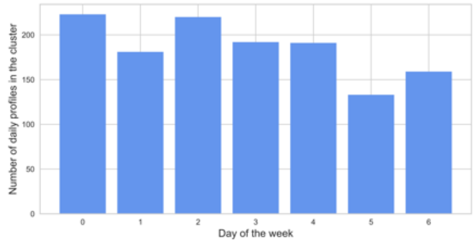


Figure 4.4.: Cluster proportions for the k-Means algorithm with $K = 3$

Cluster 1 (green) has increased consumption early in the morning with a peak of slightly over 0.4 kWh at 6 a.m. Throughout the day, the hourly electricity consumption is only



(a) Household distribution cluster 1



(b) Day of the week distribution cluster 1

Figure 4.5.: Composition of Cluster 1, regarding households and days of the week

half as high, until it increases again significantly towards the afternoon. Around 7 p.m., there is a very strong peak with about 0.9 kWh, which levels off again continuously until 11 p.m. Therefore, cluster 1 represents days when residents are home mainly in the morning and evening. The evening in particular is very relevant. This may be the case for classical workings days. Alternatively, this could also apply for days with general absence during the day, for example, for free time activities.

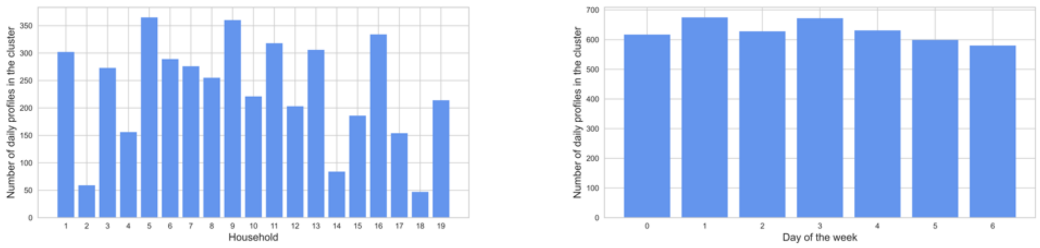
This interpretation of cluster 1 is backed up by the day of the week distribution of the cluster shown in Figure 4.5b. The numbers on the x-axis ranging from zero for Monday to six for Sunday indicate the day of the week. As expected, Saturdays and Sundays have a lower occurrence in the cluster compared to weekdays. It must be taken into account that weekend days occur, with a proportion of $2/7 = 28.57\%$, less frequently in the dataset. However, even then, with a share of 22.48% weekends are underrepresented in cluster 1 compared to an equal distribution.

Nevertheless, it should be noted that the conclusions made across all households are less distinct than for individually selected households. This is logical, because the aim of clustering across all households is to develop a good generalization over diverse usage profiles. Furthermore, one must be aware that a weekend day does not directly imply the presence of the occupants and vice versa.

Figure 4.5a shows the influence of single households for cluster 1. Households 2, 4, 14, and 18 are particularly strongly represented in the corresponding cluster. The interpretation of cluster 1 is especially clear for households 2 and 14. This is explained together with socio-demographic variables of the households in the course of the detailed analysis for selected households in Section 4.2.2.2.

Cluster 2 (blue) has consistently low consumption with two very weak peaks at 7 a.m. and 8 p.m. In general, the hourly consumption fluctuates around 0.2 kWh. That is on a significantly lower level than for the other two clusters. Hence, cluster 2 can be seen as representative for days with minimal electricity consumption, where the dwelling is not occupied most of the day. When clustering daily profiles within a particular household Yildiz et al. (2018b) documented a similar profile.

The weekday distribution in Figure 4.6b is significantly more balanced for cluster two than for the other two clusters. This can be explained by the fact that days with low attendance or consumption can occur equally on all days of the week. Similar applies to



(a) Household distribution cluster 2

(b) Day of the week distribution cluster 2

Figure 4.6.: Composition of Cluster 2, regarding households and days of the week

the distribution across the individual households (Figure 4.6a). There are still individual households, such as 5 or 9, which are particularly strongly represented. However, this is less pronounced than in the remaining clusters. Each household is represented with at least 47 days and more than half of them have at least 250 daily profiles in cluster 2.

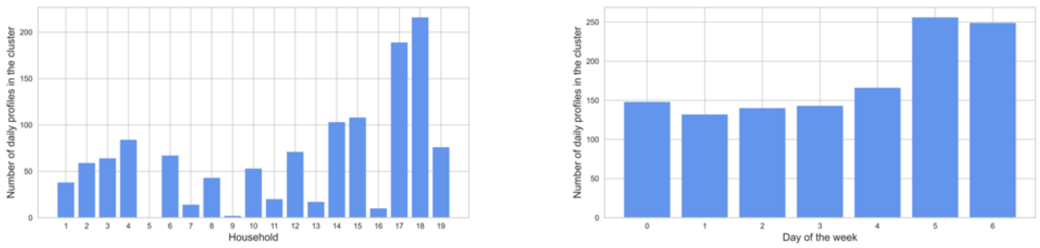
Cluster 3 (yellow) shows a steady high consumption during the day with slight peaks in the late morning at 9 a.m. and early evening at 4 p.m. Between 8 a.m. and 6 p.m. the electricity load fluctuates around 0.6 kWh. Therefore, cluster 3 represents days when residents are home a lot and also use more power-intensive applications for longer periods of time. Moreover, both the daily total consumption and the median hourly consumption are significantly higher. This can be seen together with other describing statistics in Table 4.1. The median hourly load is reported because it is far more robust to single hours with intense peaks than the mean. Peak information is therefore provided trough the maximum hourly load within a day for each cluster.

	Cluster 1	Cluster 2	Cluster 3
Consumption	Partly high	Generally low	High during day
Peaks	strong early morning, very strong evening	minimal morning and evening	light late morning and early evening
Summed daily load [kWh]	8.9512	4.6970	11.0175
Median hourly load [kWh]	0.2586	0.1902	0.5259
Maximum hourly load [kWh]	0.9342	0.2850	0.7516
Minimum hourly load [kWh]	0.1537	0.1145	0.1832
Standard deviation of hourly load [kWh]	0.2307	0.0512	0.1819

Table 4.1.: Describing statistics for the three cluster centroids

For cluster 3, the analysis of the distribution across the days of the week is particularly interesting. In Figure 4.7b one can see that weekend days are significantly more represented than regular weekdays. If it is taken into account that weekdays occur considerably more frequently in the dataset ($5/7 = 0.7143$), the effect becomes even clearer. 40.92% of the daily profiles in cluster 3 are weekends, which is clearly more than the 28.57% in the unclustered full dataset.

The question could rise why weekdays were still found in cluster 3 and weekend days in cluster 1. However, as described in previous sections, one needs to keep in mind the



(a) Household distribution cluster 3 (b) Day of the week distribution cluster 3

Figure 4.7.: Composition of Cluster 3, regarding households and days of the week

high volatility and randomness of residential occupancy behaviour. It is not possible or useful to characterize daily profiles solely based on the type of the day in the week. In the following Section 4.2.2.2, further variables and components are used to investigate the influence of household specific characteristics on the formed cluster. Nevertheless, the analysis performed, especially for cluster 1 and cluster 3, confirmed a considerable influence of the type of weekday on electricity load profiles.
In the day of the week distribution for the third and last cluster, a few households are more strongly represented than others (Figure 4.7a). Again, as for households 2 and 14 in cluster 1, the interpretation is particularly clear for these households. Section 4.2.2.2 will dive deeper into this aspect.

4.2.2.2. Detailed Analysis for selected Households

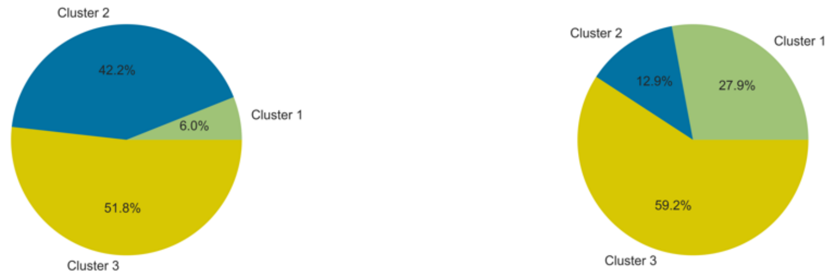
This section further develops the insights gained from the top-level analysis of the three clusters and their characteristics. It is investigated how the daily profiles of single households are distributed over the three clusters with their corresponding day of the week. In a second step, socio-demographic variables, extracted from the original IDEAL data representation (Pullinger et al., 2021), are included in the analysis. The goal is to evaluate how these variables can justify and further explain the clustering results for individual households.
For the analysis, households with extreme occupancy behaviour are of particular interest. On the one hand, certain households in the dataset are nearly always occupied, on the other this is barely the case for others. The degree of occupancy is measured by the variables *occupied days* and *occupied nights*. The authors of the IDEAL dataset chose the following wording for the survey: „In a typical week, how many days would you say your home is occupied during the day; that is, with at least one person in it for most of the day? “ (Pullinger et al., 2021). The same wording applies to the variable *occupied nights*. It must be taken into account that the question is relatively vague and answers may be dependent on the individual interpretation of residents. To check the survey results and keep them up to date, a follow-up survey was done in September 2017. This is relevant for the observation period considered in this study. While most of the residents stuck to their original statement, some have drastically changed their indications concerning the occupancy behaviour. This was taken into account in the analysis.

Demographic variables				Electricity consumption [kWh]		
HH	residents	occupied days	occupied nights	mean hourly	total annual	max hourly
1	2	1	7	0.24	2115.41	2.24
2	2	7	7	0.41	3576.66	4.91
3	3	4	7	0.27	2330.99	2.24
4	1	2	7	0.36	3148.90	4.51
5	2	7	7	0.11	987.99	1.75
6	3	4	7	0.26	2254.15	2.33
7	1	2	7	0.24	2144.06	2.52
8	4	4	7	0.27	2327.90	3.46
9	3	2	7	0.13	1110.64	3.22
10	2	3	7	0.27	2395.84	2.82
11	1	1	7	0.15	1328.44	8.57
12	2	2	7	0.29	2528.12	4.65
13	1	2	5	0.18	1606.97	4.96
14	3	2	7	0.39	3422.85	5.23
15	2	6	7	0.31	2721.05	3.64
16	1	1	7	0.21	1825.65	1.64
17	2	7	7	0.40	3523.32	6.61
18	3	7	7	0.41	3572.30	2.38
19	1	2	7	0.34	2977.92	3.35

Table 4.2.: Demographic variables and consumption statistics [kWh] per household

The dataset on hand with 19 households provides a good representation of different consumer types. The number of residents ranges from one to four and different compositions of the number of occupied days and nights can be observed (Table 4.2). Once again, it must be clarified that the electricity consumption does not have to be directly dependent on these two variables. More complex components, like consumption and sustainability awareness or energy efficiency of the dwelling and the used appliances, also play an important role.

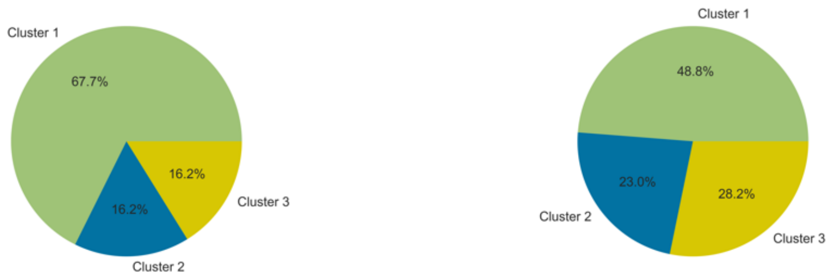
With over 3500 kWh of annual electricity consumption, households 17 and 18 were the two largest consumers in the dataset. The residents stated that the accommodation unit was occupied on 7 out of 7 days as well as nights. Since the consumption level at night is significantly lower for all clusters, the effect of occupied days certainly has a much stronger impact on the electricity consumption. It is therefore logical that households 17 and 18 have particularly many days in cluster 3, which was characterized by high consumption during the day. 51.8% of the days of household 17 and 59.2% of household 18 were assigned to this cluster (Figure 4.8). This is considerably more than for all other households. However, it is logical that also days in the two other clusters exist. On the one hand, not all occupied days, like for example home office days, may lead to intensive electricity usage during the day. On the other hand, there exist 'non-typical' weeks where residents may be less at home. Hence, the comparison to other households of the dataset, where the proportion of days in cluster 3 is much lower, is important.



(a) Household 17 (b) Household 18

Figure 4.8.: Cluster allocation for two households with high occupancy and consumption

It is also interesting to take a closer look at households 2 and 5 (Figure 4.9). Household 2 reported in the first survey, as did households 17 and 18, that the housing unit is mostly occupied on all days. However, at only 16.2% the proportion of days in cluster 3 is significantly lower for this household (Figure 4.9a). While this may be simply due to a different usage pattern with higher focus on the morning and evening, it certainly plays a role that the residents changed their statement in the aforementioned follow-up survey. According to the second statement of September 2017, the dwelling of household 2 was only occupied 2/7 days, instead of 7/7 days (Pullinger et al., 2021).



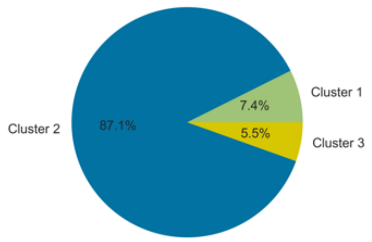
(a) Household 2 (b) Household 14

Figure 4.9.: Cluster allocation for two households with high occupancy but low consumption

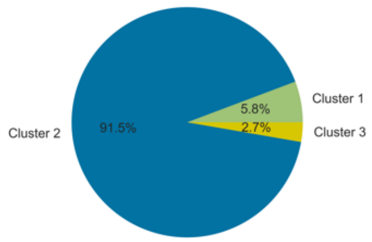
Furthermore, for household 2, the day of the week seems to have a great influence on the consumption profile. 83.0% of the days in cluster 1 are weekdays (Figure A.5a). With a share of 81.4% weekend days, cluster 3 is dominated by Saturdays and Sundays (A.5b). This aligns very well with the interpretation of the formed clusters made in Section 4.2.2.1. Cluster 1 was interpreted as profile for days with high consumption in the morning and evening, while cluster 3 as representative for days with continuous high consumption during the day. For cluster 2 with low consumption, there is no particular tendency regarding the time of the week. However, it is striking that it contains a particularly high number of Fridays for household 2. This again demonstrates the strong tendency towards individual consumption profiles in residential electricity consumption.

Household 14 is another household that strongly justifies the cluster interpretation (Figure 4.9b). In cluster 1, 97.8% of the days are weekdays (Figure A.6a), in cluster 3, 72.8% of the days are on the weekend (Figure A.6b).
A special case is household 5, which has all its 365 daily profiles in the low consumption cluster 2 (Figure 4.6a). Given that the household is occupied daily according to the survey, this may seem surprising at first. However, taking into account the total annual electricity consumption, the clustering result becomes reasonable. Household 5 shows by far the lowest hourly and total electricity consumption (Table 4.2). The very low electricity demand could be explained, for example, by a particularly energy-efficient lifestyle or appliances.

Now that households with very high attendance have been considered, households with more absence are assessed. Household 11 and 16 both indicated only 1 of 7 occupied days during a typical week. This is clearly reflected in the cluster shares (Figure 4.10). For both, a very high percentage of days were assigned to cluster 2. Since it represents days with low consumption, the interpretation is again backed up here. Interestingly, for household 11, the share of 87.1% for cluster 2, corresponds approximately to the share of unoccupied days of $6/7 = 85.7\%$ (Figure 4.10a). Moreover, 90% of the days in the "High during day" cluster (cf. Table 4.1) are actual weekends and 100% of the days were Fridays or weekends (Figure A.7b). Household 16 showed even a slightly higher percentage of days in the low consumption cluster (Figure 4.10b).



(a) Household 11



(b) Household 16

Figure 4.10.: Cluster allocation for two households with low occupancy

Another aspect that was included in the analysis is the influence of national holidays. However, no significant difference could be found compared to usual working days. For most households, there was no deviation of the cluster shares compared to non-holidays. Therefore, assumptions that households might either be absent due to travelling or be at home more could not be confirmed.

5. Forecasting

In this chapter, the obtained clustering results are combined with classical load forecasting models. It is investigated, whether the type of clustering applied here improves the forecast performance. The question is if the pre-grouping of the daily electricity consumption profiles provides the models with sufficient additional information.

5.1. Forecast Methodology

Using the indices of the clustered daily profiles, the days could be divided into three subgroups. In addition, the composition of all unclustered days together was considered as fourth dataset for comparison. Support Vector Regression (SVR) and the Random Forest Regressor (RFR) were applied to all four constellations of the data.
For each of the four forecast settings the corresponding data was split into train, validation and test set. The models were trained on 70% of the days. 10% were reserved for validation and hyperparameter tuning of the involved algorithms. This was done by means of a grid search. Specifications to the tested parameters follow together with the reporting of the results for the two models. The remaining 20% were used for testing and evaluation of the models since it is important to test the model on data which has not been leaked to it beforehand.
Two different split scenarios were tested. In the first one, every household was represented in training, validation and test set with the corresponding percentage. This was not the case for the second one, because the aligned households were consecutively split. The second scenario delivered significantly better results and is therefore considered in this study. Another strong argument for this approach is the better generalizing done by the model. In this setting, the model was fitted on the days of the first households and tested on households the model has not seen before. This prevented overfitting and resulted in a generalized model with better performance.
While it is common to apply shuffling to the input data for classical machine learning algorithms, the original order is often retained for time series data. This is particularly relevant when using models where future predictions depend on past predictions, like in Recurrent Neuronal Networks (RNN). However, algorithms like SVR and RFR do not have such memory. Instead, the time series information of past consumption values was added to the feature matrix by feature engineering (e.g. Table 5.1). Therefore, both approaches can theoretically be applied to the present dataset. Zhang et al. (2018) documented that random sampling can outperform consecutive splitting when less regularity in the residential usage patterns is present. Both options were tested for the daily and hourly forecast scenario (cf. Section 5.3). Similar to the conclusions made by Zhang et al. (2018), for the daily forecast, with less regularity, shuffling the observations outperformed consecutive splitting. At hourly resolution, where temporal sequence plays a more important role, consecutive splitting achieved slightly better results.

Another aspect to consider, is the computational efficiency of the algorithms. The time complexity of Support Vector Machines, both for classification and for regression, increases fast with the number of training vectors. In the worst case, it is $O(n^3)$, making its time complexity cubic. (Ns, 2015)
In general, the prediction can be done for different time granularities. In this thesis, forecast models for daily and hourly resolution were considered. When applying daily forecasting with 6935 days, the cubic time complexity could be handled. However, in the case of forecasting at hourly resolution, there were $6935 \times 24 = 166440$ observations in the unclustered case. This quickly lead to enormous demands on the computing and storage capacity. For this reason, the forecasting analysis at hourly resolution was limited to the first 100 days for each household in the observation period. Consequently, the clustering approach was applied again to this subset.
While making computations manageable without an external server, this constraint at the same time served as yet another validation of the clustering results. Since for the subset of the daily profiles the centroid curves differed only slightly, the clustering structure was confirmed once again and can be considered robust (Figure A.8). (Han et al., 2012, p. 532)

5.2. Evaluation Metrics

There are many metrics available to evaluate the prediction performance of a regression problem. Since these set different weighting priorities, the choice of the right metric is essential. Another important issue is scale dependence, which negatively affects comparability. Therefore, the additional consideration of relative metrics is important. Literature on electricity load forecasting for higher aggregation levels often makes use of the MAPE as performance metric (e.g. Shaqour et al. (2022)). However, when applied to data on a low aggregation level, like hourly electricity consumption on residential household level, values approaching zero are common. Equation 5.1 of the MAPE demonstrates why this is problematic and can lead to unmeaningful high errors. A_t are the actual values and F_t the forecasted ones.

$$MAPE = \frac{100\%}{N} \sum_{t=1}^{N} |\frac{A_t - F_t}{A_t}| \tag{5.1}$$

As the total daily consumption is always sufficiently far enough from zero, there were no problems for the daily forecast. However, for the hourly forecast, A_t in the denominator can be very small and therefore cause the MAPE to be arbitrarily high. The same applies to the RMSPE. Humeau et al. (2013) reported the same issues in the context of household electricity load forecasting and used other relative metrics instead. In this thesis, the Normalized Mean Absolute Error (NMAE) and the Normalized Root Mean Squared Error (NRMSE) were considered. Following Yildiz et al. (2018a), the MAE (Equation 5.2) and RMSE (Equation 5.3) were normalized with the mean of the actual values A_t of the corresponding subgroup. This compensates for the fact that consumption in the individual clusters was of significantly different magnitude. While cluster 2 was characterized by low electricity loads, consumption could take much higher values in cluster 3.

$$NMAE = \frac{MAE}{\overline{A}} = \frac{\frac{1}{N}\sum_{t=1}^{N}|A_t - F_t|}{\overline{A}} \tag{5.2}$$

$$NRMSE = \frac{RMSE}{\overline{A}} = \frac{\sqrt{\frac{1}{N}\sum_{t=1}^{N}(A_t - F_t)^2}}{\overline{A}} \tag{5.3}$$

Following this argumentation, these two normalized performance metrics are appropriate for the comparison of the different subgroups. Moreover, the results of the hourly and daily forecast become more comparable.

5.3. Forecast Scenarios

This section presents the daily and hourly forecast scenario together with the obtained results. To obtain electricity consumption at daily resolution, the hourly data was aggregated to a daily representation by the following formula: $C_d = \sum_{i=1}^{24} C_{h_i}$ (Zhang et al., 2018). C_{h_i} represents the electricity consumption at hour i and C_d represents the total daily electricity consumption of one day.
Following the prediction of the summed daily consumption, the model was applied to hourly electricity consumption data. It should be noted that more detailed and finer predictions are generally more difficult and involve greater errors than predictions for higher granularities (Lusis et al. (2017)). Therefore, the comparison of the models for these two data representations is particularly interesting.

5.3.1. Daily Forecast

The aggregation to the daily representation of the hourly electricity consumption data resulted in a data frame with $365 \times 19 = 6935$ observations. Several features were generated from the original time series. To include lagged variables, which play an essential role for time series data, each household's first 5 days were discarded. This resulted in the final data frame with $6935 - (5 \times 19) = 6840$ observations (Table 5.1). In the first column

dayIndex	Day 0	Day - 1	Day - 2	Day - 3	Day - 4	Day - 5	MeanTemp	DayOfTheWeek	Peak - 1	Peak - 2	Peak - 3	STD - 1	STD - 2	STD - 3	Household
6	6.59	6.75	8.55	6.41	7.08	9.34	0.13	6	0.86	1.26	0.65	0.18	0.31	0.16	1
7	6.02	6.59	6.75	8.55	6.41	7.08	4.76	7	0.99	0.86	1.26	0.24	0.18	0.31	1
8	5.17	6.02	6.59	6.75	8.55	6.41	5.06	1	0.70	0.99	0.86	0.16	0.24	0.18	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6933	12.39	7.47	6.76	8.78	7.58	11.21	-0.04	6	1.63	0.44	1.55	0.31	0.09	0.36	19
6934	10.65	12.39	7.47	6.76	8.78	7.58	0.34	7	2.26	1.63	0.44	0.58	0.31	0.09	19
6935	9.51	10.65	12.39	7.47	6.76	8.78	2.04	1	1.31	2.26	1.63	0.34	0.58	0.31	19

Table 5.1.: Target and feature variables for the daily forecast

'Day 0', the target variable to predict, is displayed. Several lagged time series features were included in the feature matrix. The variables *'Day - X'* reflect the daily electricity consumption of the past X_{th} day. This applies analogously to the standard deviation, which was included for the last three days. The lagged variable *'Peak - X'* indicates the

strength of the peak on the previous days and is intended to support the model in forecasting particularly high electricity demands.
Furthermore, the mean temperature of the last day was included as a meteorological variable. This provided the model with information about the current temperature range. However, air conditioning is not widely used in Scotland. The market research firm *Mintel* documented that only 0.5% of houses and flats in the United Kingdom have an air conditioning system (Leggett, 2006). As heating consumption was recorded separately in the dataset, the temperature variable was not expected to have much predictive power.
A calendrical variable provided via the numbers zero to six information about the day of the week. Moreover, the column *'Household'* integrated a label for each household.
Since the value range of the individual features differs, standardization is indispensable (Han et al., 2012, p. 532). For this purpose, the *StandardScaler* of *sklearn* was used to transform the features to have mean zero and unit variance (Pedregosa et al., 2011).

Table 5.2 shows the results for the four different prediction scenarios using SVR. As introduced in the methodology Section (5.1), the parameters were optimized with a grid search. There, the RMSE was used as comparison metric for optimization.
C indicates the regularization strength. Larger values of *C* emphasize the minimization of the total error and therefore lead to less regularized and generalized models. *Epsilon* defines a tolerance above which a deviation of the prediction from the true value is penalized. (Awad & Khanna, 2015, p. 67 - 72)
The intuition to *epsilon* can be seen clearly, when looking at the tuned parameters for each of the four groups (Table 5.2). Since the consumption level in cluster 2 is generally lower, a relatively low tolerance of 0.2 kWh was chosen. In contrast, for cluster 3, a significantly higher tolerance of 2 kWh was found to deliver the best results. Together with the interpretation of *epsilon* this can be seen as further validation for the clustering results. On all four subgroups, the RBF kernel outperformed the linear and polynomial kernel. The

Support Vector Regression				
Evaluation metric	**All days unclustered** C = 10, eps = 1.1	**Cluster 1** C = 1, eps = 1.3	**Cluster 2** C = 1, eps = 0.2	**Cluster 3** C = 1, eps = 2
RMSE	2.3317	1.8866	1.1276	2.8722
NRMSE	0.3633	0.2081	0.2378	0.2593
MAE	1.6234	1.4386	0.8825	1.9595
NMAE	0.2529	0.1587	0.1862	0.1769

Table 5.2.: SVR daily forecast results for the four forecasting scenarios

MAE and RMSE are not suitable for comparing the groups and should only be considered within a group, taking into account the order of magnitude. Thus, cluster 3 with its high consumption, had the highest MAE and RMSE. Cluster 2, on the other hand, showed less than half the error for these two metrics. Again, this can be seen as a validation of the clustering results as the algorithm succeeded to separate the days by the magnitude of the electricity consumption.
Nevertheless, the MAE is worth to be considered due to its straightforward interpretation.

It indicates how much predictions deviate on average from the actual values. For example, for cluster 2, the predictions deviated on average 0.88 kWh from the actual daily load.

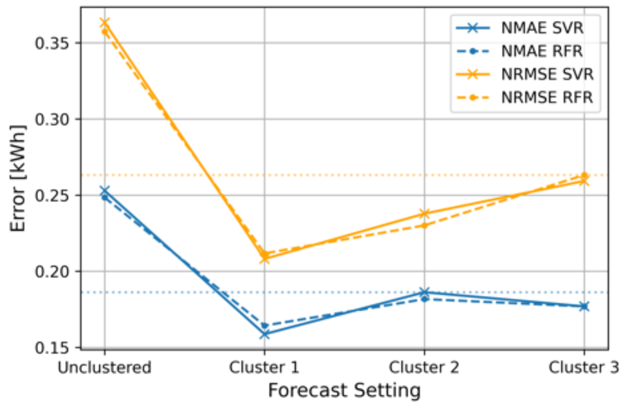


Figure 5.1.: Effect of clustering on the daily forecast performance of SVR measured by NMAE and NRMSE

Now, the comparison of the individual groups using NMAE and NRMSE is particularly relevant for the second research question considered in this thesis. The normalized error is significantly lower for the three subsets formed by the pre-clustering. While Table 5.2 presents the actual metric scores, the improvement achieved through clustering the daily profiles according to their consumption behaviour can be seen clearly in Figure 5.1. The solid line represents the results for SVR, while the scores for RFR are shown by the dashed one. Both models perform similarly and confirm a significant improvement for the daily prediction by the pre-clustering step. To highlight this, an upper error bound for the clustered case is indicated by the two dotted horizontal lines.
The figure also demonstrates a key difference between the MAE and the RMSE. While the MAE treats all magnitudes of errors equally, the RMSE penalizes large errors particularly heavily by squaring them before averaging. This is a possible explanation why the NRMSE increases for cluster 3, while the NMAE decreases.

Random Forest Regressor				
Evaluation metric	**All days unclustered** depth$_{max}$ = 20, leaf$_{minSamples}$ = 10	**Cluster 1** depth$_{max}$ = 11, leaf$_{minSamples}$ = 15	**Cluster 2** depth$_{max}$ = 20, leaf$_{minSamples}$ = 10	**Cluster 3** depth$_{max}$ = 7, leaf$_{minSamples}$ = 10
RMSE	2.2936	1.9196	1.0909	2.9157
NRMSE	0.3574	0.2117	0.2301	0.2632
MAE	1.5940	1.4890	0.8615	1.9597
NMAE	0.2484	0.1643	0.1817	0.1769

Table 5.3.: RFR daily forecast results for the four forecasting scenarios

The tuned parameters of the RFR are reported together with the results in Table 5.3.

$Depth_{max}$ specifies the maximum depth the trees are allowed to have, while $leaf_{minSamples}$ determines the required amount of data points to create a leaf node (Pedregosa et al., 2011).
Interestingly, for the unclustered case and for cluster 2 the RFR performed marginally better, while SVR was the better choice for clusters 1 and 3. However, these differences are minimal and do not change the influence of the clustering on the forecast performance displayed in Figure 5.1. As the analysis from Figure 3.2b already suggested in the course of the exploratory data analysis, previous consumption values had the greatest predictive power.

5.3.2. Hourly Forecast

The next step was to apply the two introduced predictive models to the hourly data resolution. Due to the computational complexity addressed in Section 5.1, the hourly forecast was applied with restriction to the first 100 days of the dataset. This corresponds to $100 \times 24 \times 19 = 45600$ hourly data points. As before, numerous time series features were generated. To do so, the first 120 hours, respectively the first five days, of each household were discarded. This resulted in a data frame with $45600 - (120 \times 19) = 43320$ hourly observations. Consequently, the hourly index starts at the 121^{th} hour on day six.
Table 5.4 shows the corresponding data frame[1], including the features and the target.

hourIndex	target	lagMinus1	lagMinus2	lagMinus3	lagMinus12	lagMinus24	lagMinus48	lagMinus72	stdLast2hours	stdLast3hours	hourOfDay	dayOfTheWeek	currentTemp	household
121	0.1299	0.1512	0.3649	0.5721	0.1330	0.2137	0.1137	0.1479	0.2370	0.2085	0.0	6.0	1.631	1.0
122	0.1253	0.1299	0.1512	0.3649	0.1305	0.1448	0.1230	0.1227	0.2753	0.2088	1.0	6.0	1.918	1.0
123	0.2470	0.1253	0.1299	0.1512	0.3116	0.2600	0.2427	0.1099	0.1862	0.2093	2.0	6.0	2.107	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
45598	0.3136	0.3813	0.5104	0.6312	0.1597	0.3110	0.4090	0.3858	0.4769	0.3840	21.0	2.0	5.125	19.0
45599	0.2071	0.3136	0.3813	0.5104	0.1597	0.1998	0.3484	0.3210	0.4519	0.3786	22.0	2.0	4.615	19.0
45600	0.1604	0.2071	0.3136	0.3813	0.2892	0.1582	0.2485	0.3080	0.4649	0.3762	23.0	2.0	4.456	19.0

Table 5.4.: Target and feature variables for the hourly forecast

The target variable to be predicted is the hourly electricity consumption. Moreover, several lagged hourly electricity loads were included. The exploratory analysis suggested that shortly preceding consumption values and the lags at the same hour of the previous days could be particularly relevant. To also include the variation within the previous hours, the standard deviation of the past two and three hours was included as a variable. It did not prove useful to include hours further back in time.
Since the time of the day can have a significant impact on the electricity consumption at hourly resolution, the variable *'Hour of Day'* was created. Regarding the weather, the *'Current Temperature'* was approximately represented by the temperature of the previous hour, to avoid leaking unknown information into the model. The variable *'Household'* additionally indicates to which household the current hourly consumption belongs. Finally, the corresponding *'Day of the Week'* (range 1 to 7) was added to each consumption value as input information. However, it is assumed that the effect of this variable is limited, since the information about characteristics of individual days has already been considered by the pre-grouping of the daily profiles.

[1]Please note, that the hour index here starts at one for better readability. However, in Python indexing conventionally starts at zero.

For this scenario, the variables were as well standardized due to their different value ranges. At the hourly data resolution, the *MinMaxScaler* proved to be the most suitable, since the relative difference between the individual data points is larger than at daily resolution. The *MinMaxScaler* unifies the value range of all variables to [0,1]. (Pedregosa et al., 2011) In general, the results are comparable for different scaling types. These were only minor improvements as the last tuning.

Table 5.5 presents the error metrics for SVR at hourly electricity data resolution. As already stated in the literature, the error increased significantly at hourly resolution, compared to the daily forecast. At daily resolution, the aggregation effect made the consumption more predictable, while the highly fluctuating hourly consumption values were more challenging for the models.
The parameters of the models were optimized using a Grid Search and adjusted to the four subgroups. Again, the tuned parameters show that the clusters are composed differently. The tolerance *epsilon* was logically at a lower level for the hourly forecast due to the much smaller magnitude of hourly consumption values compared to aggregated daily loads. Moreover, the optimized value for *epsilon* is anew significantly higher for cluster 3 than for the other clusters.

Support Vector Regression				
Evaluation metric	**All days unclustered** C = 10, eps = 0.02	**Cluster 1** C = 10, eps = 0.02	**Cluster 2** C = 1, eps = 0.02	**Cluster 3** C = 10, eps = 0.2
RMSE	0.2746	0.3058	0.1618	0.2920
NRMSE	0.6931	0.6446	0.6362	0.7451
MAE	0.1539	0.1961	0.0879	0.1960
NMAE	0.3885	0.4133	0.3456	0.5001

Table 5.5.: SVR hourly forecast results for the four forecasting scenarios

Random Forest Regressor				
Evaluation metric	**All days unclustered** $\text{depth}_{max} = 20$, $\text{leaf}_{minSamples} = 10$	**Cluster 1** $\text{depth}_{max} = 20$, $\text{leaf}_{minSamples} = 10$	**Cluster 2** $\text{depth}_{max} = 13$, $\text{leaf}_{minSamples} = 15$	**Cluster 3** $\text{depth}_{max} = 7$, $\text{leaf}_{minSamples} = 10$
RMSE	0.2546	0.2947	0.1567	0.2429
NRMSE	0.6426	0.6212	0.6162	0.6197
MAE	0.1426	0.2049	0.0979	0.1532
NMAE	0.3598	0.4319	0.3850	0.3908

Table 5.6.: RFR hourly forecast results for the four forecasting scenarios

Analogue to this, Table 5.6 shows the RFR results of the hourly forecast. At hourly resolution, the difference between the two models was slightly higher. The NRMSE is at a lower level for the RFR than for the SVR for all 4 settings. Particularly, the SVR had problems with cluster 3, with its very high and variable daytime consumption. Even if the difference between the two models measured by the MAE is less clear, the RFR is preferable at the hourly level.

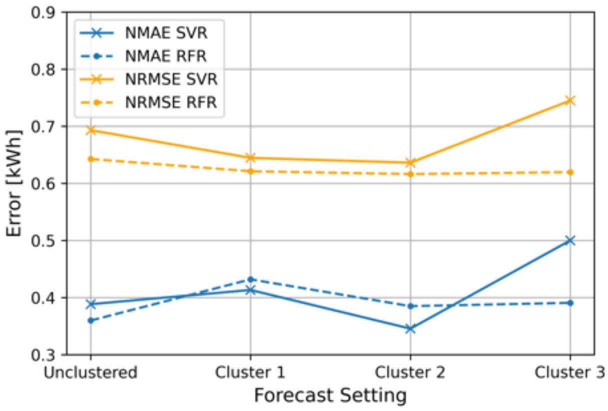
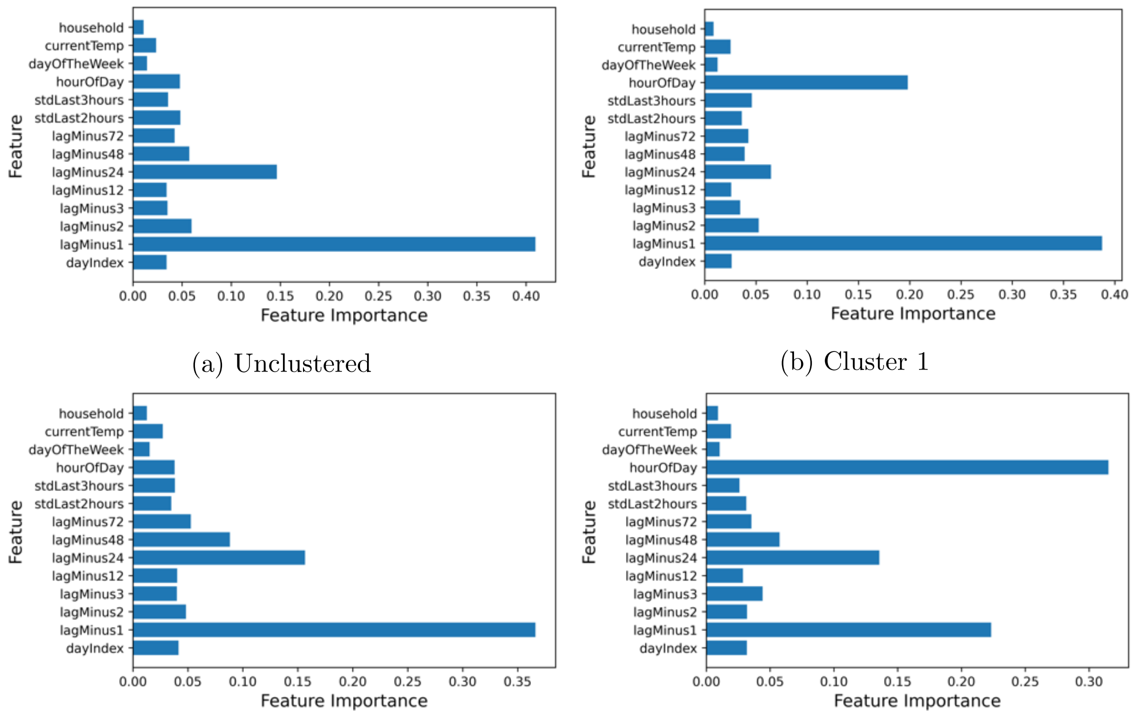


Figure 5.2.: Effect of clustering on the hourly forecast performance of SVR measured by NMAE and NRMSE

Moreover, the error reducing-effect observed at daily resolution was significantly less pronounced. The NRMSE could be reduced by a few percentage points for all clusters by the RFR (Figure 5.2). However, the MAE did not confirm this trend. Apparently, clustering across all daily profiles had a much stronger effect on predicting daily consumption. However, it must also be taken into account in this comparison that hourly electricity consumption at household level is much more difficult to predict.

That the performed clustering had a remarkable influence at hourly resolution, despite the lower improvement, becomes apparent when taking a closer look at the feature importance of the individual models (Figure 5.3). The bar plot shows the importance of each feature considered for splitting in the RFR.
For classification problems with Random Forests, the concept of Gini importance is used for determining the importance of each feature. The Gini index measures the resulting *impurity* of a node after each split. In regression problems, this concept can be applied using summed squares as the impurity measure. The importance of variable X_i thus corresponds to the summed reduction of the squared error that resulted from all splits made with variable X_i. (Nembrini et al., 2018)
For the unclustered dataset and for cluster 2, the electricity consumption of the preceding hour as well as the consumption 24 hours before are the strongest predictors. It is comprehensible that the time of day has low feature importance, since for cluster 2 the dependence of consumption on the time of day is only weak. For clusters 1 and 3, where consumption profiles are strongly dependent on the daytime, the variable is the second, respectively, the most important feature in the model. This confirms the course of the centroids for the individual clusters displayed in Figure 4.2a.
The models were thus able to validate the divergent characteristics of the separated groups. However, this has not led to the same improvement in forecast performance for hourly prediction compared to daily prediction.



(a) Unclustered

(b) Cluster 1

(c) Cluster 2

(d) Cluster 3

Figure 5.3.: Feature importance in forecasting for each of the four subgroups

5.4. Conclusion and Comparison

For the aggregated daily consumption data, a significant reduction in the normalized errors was achieved by clustering. This effect could be confirmed by both, NRMSE and NMAE. In the best case, the NRMSE was reduced from 0.3633 for the unclustered dataset, to 0.2081 for Cluster 1.

Another finding, was that the prediction of electricity consumption of days in cluster 3 was more difficult. As already confirmed in literature, high volatility and random fluctuations during the day increase the error. Accordingly, this was even more pronounced for the hourly resolution. Here, the NMAE for cluster 3 was higher than in the unclustered case. Nevertheless, it can be seen as a success to identify and isolate these days, which are particularly difficult to predict in the course of the day from the data. Using the RFR, even at the hourly resolution, the NRMSE could be slightly reduced. The feature importance varies drastically depending on the subgroup at hourly resolution.

6. Conclusion

6.1. Summary and Discussion

The objective of this work was to investigate the potential of clustering for improving the prediction of household electricity loads. Accurate predictions of consumption for the next hour or day are important when optimizing efficiency in energy communities and especially the HEMS used there. Compared to larger aggregation levels, individual forecasts still have great potential for improvement. In this work, similar results presented in literature were confirmed by showing that clustering positively affects the prediction quality for the dataset on hand. For daily resolution, a reduction of the NRMSE of up to 42%[1] could be achieved. This trend was significantly weaker for the hourly forecast. There, the reduction of the NRMSE by clustering was at best 4.1%[2]. This confirmed the great challenge of hourly forecasts for individual households which needs to be addressed further in future work. Nevertheless, the significant improvement in prediction for daily resolution may offer great potential for further development of energy community concepts and is of particular interest for the HEMS.

To the best of the author's knowledge, the approach of clustering daily profiles across multiple households or an entire community is outstanding (cf. Section 4.1). Substantially, three main daily profiles could be found as a result of the applied clustering. The detailed analysis of the resulting clusters emphasized the high degree of individuality as well as the influence of a wide range of demographic factors on electricity consumption patterns at household level.

6.2. Future Work

There are some limitations in this thesis that could be addressed in future work. Since the dataset studied included 19 households, it would be interesting to apply the used clustering approach to a larger number of households. In this context, the influence of the number of included daily profiles on the clustering quality could be investigated. However, increased computing power would be required. This aspect also plays a role in the prediction for hourly resolution. Due to computational limitations, the prediction for the hourly resolution was limited to the first 100 days of the dataset. With external servers for computation, these limitations could be resolved. An even further generalization could also be achieved by extending the observation period beyond one year. In contrast, it would also be interesting to examine how specific restrictions of the number of daily profiles affect the results.

Further, an integration of the heating consumption data of the dataset into the approach is required. The seasonality is expected to be much higher than in the case of the examined

[1] Comparing the unclustered full dataset with cluster 1, SVR (greatest improvement)
[2] Comparing the unclustered full dataset with cluster 2, RFR (greatest improvement)

electricity data. That means, a holistic integration of all types of energy consumption at household level is recommended. With regard to the HEMS, a comparison of the load forecast with the future generation from the local renewable energy capacities is highly relevant. For this purpose, taking into account meteorological conditions, a model could be created to predict these capacities for the Edinburgh region. Subsequently, it can be analysed how likely the expected loads can be covered by local energy generation.
Mentioning the local conditions for Edinburgh, it is recommended to apply the presented approach for households from different geographic regions. Seasonality and feature importance may vary depending on the local conditions. For example, Pirbazari et al. (2021) found higher seasonality for an australian dataset.

With the challenges addressed, energy communities and load forecasting continue to offer great research potential. It is worth investing in further research, as the concepts play a crucial role for the inevitable energy transition.

7. Erklärung

Ich versichere hiermit wahrheitsgemäß, die Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und die Satzung des Karlsruher Instituts für Technologie (KIT) zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, den 12.08.2022 L. Bleher

Lucas Bleher

Appendix

A. Figures

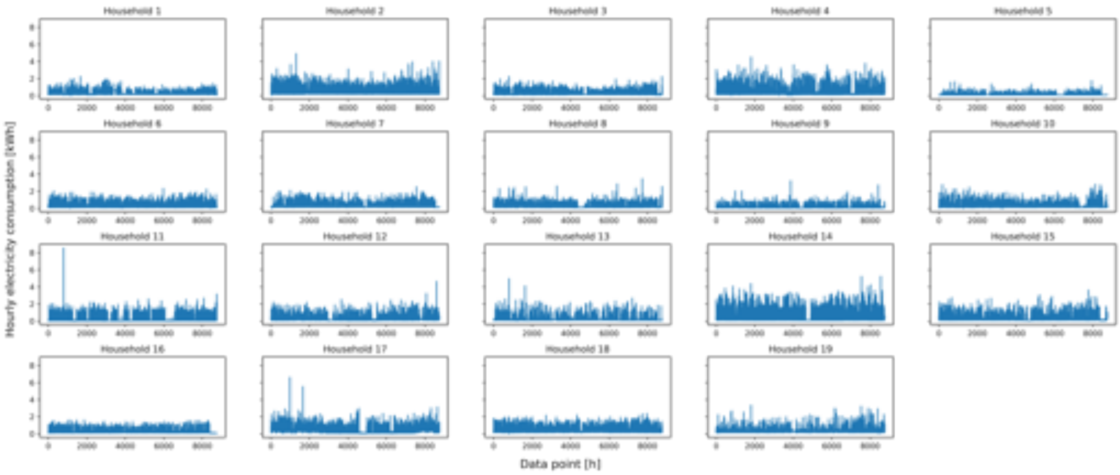


Figure A.1.: Overview of hourly electricity consumption of all 19 households

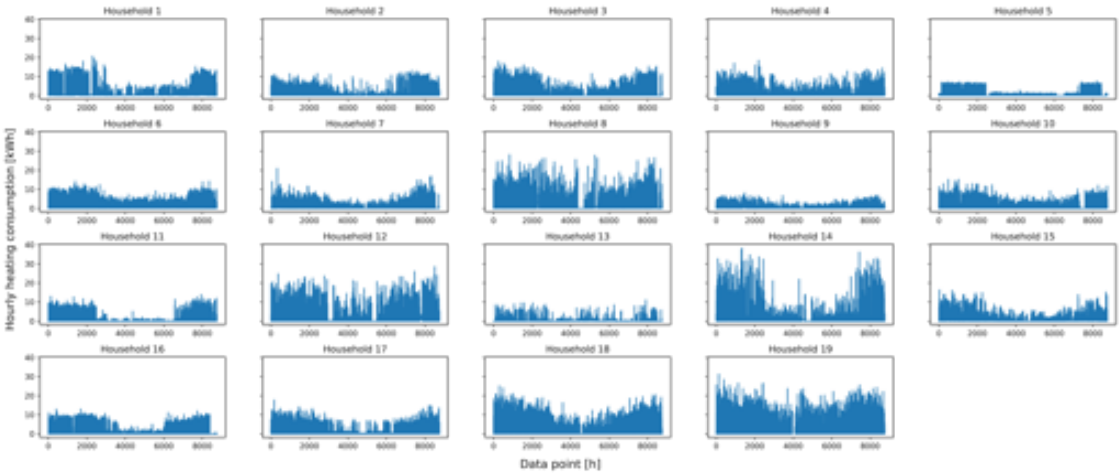


Figure A.2.: Overview of hourly heating consumption of all 19 households

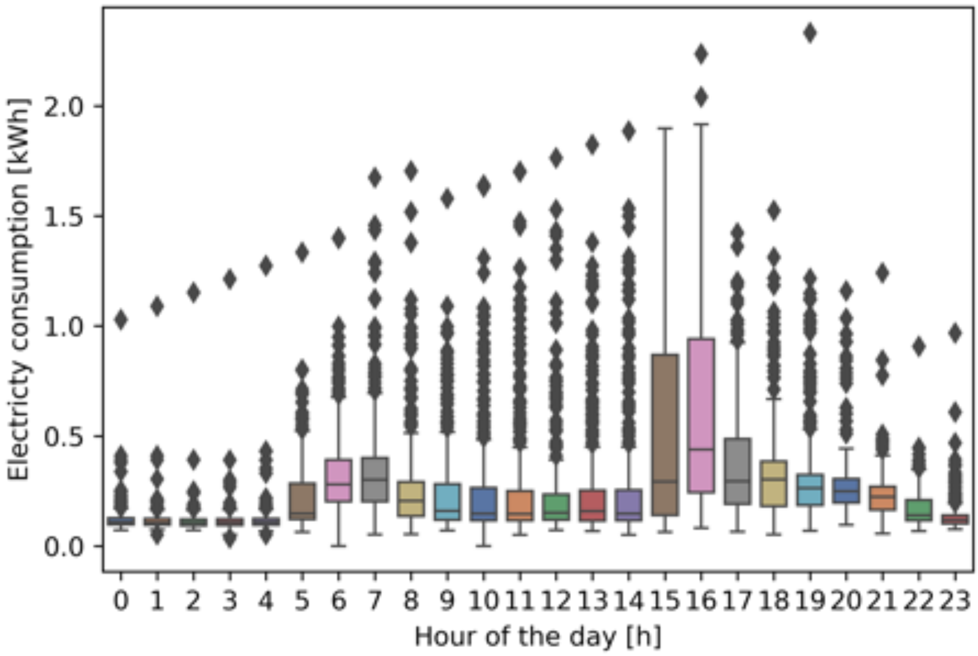


Figure A.3.: Boxplot daily schedule, HH 06

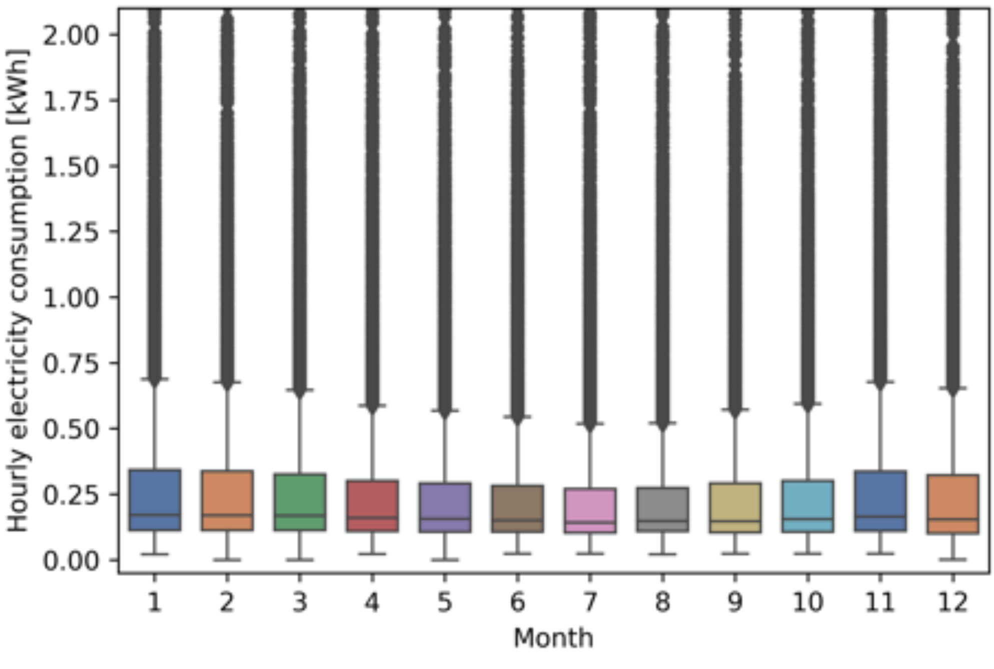
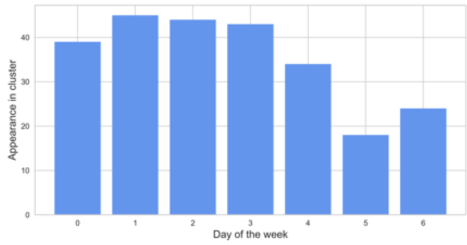
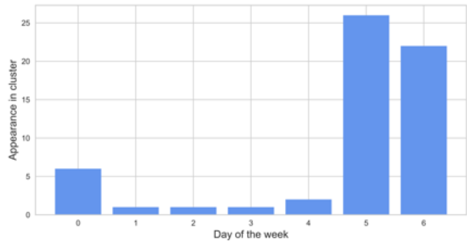


Figure A.4.: Boxplot consumption per month, all HH

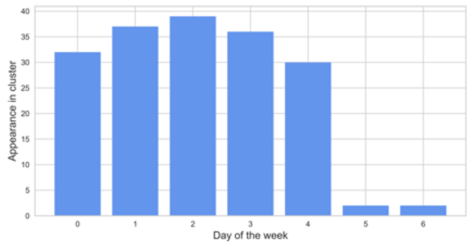


(a) Cluster 1

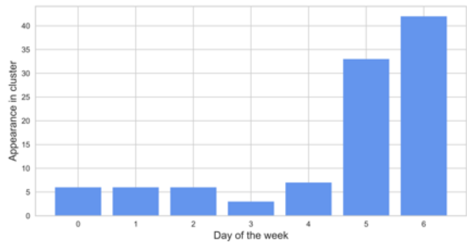


(b) Cluster 3

Figure A.5.: Day of the week distribution of household 2 for selected clusters

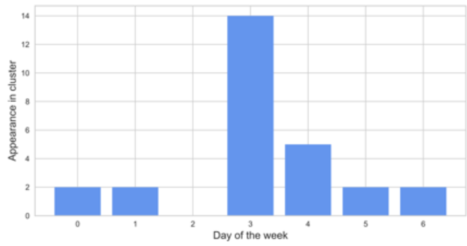


(a) Cluster 1

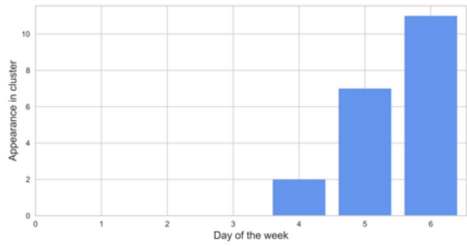


(b) Cluster 3

Figure A.6.: Day of the week distribution of household 14 for selected clusters

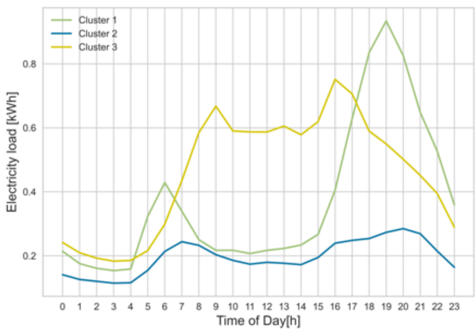


(a) Cluster 1

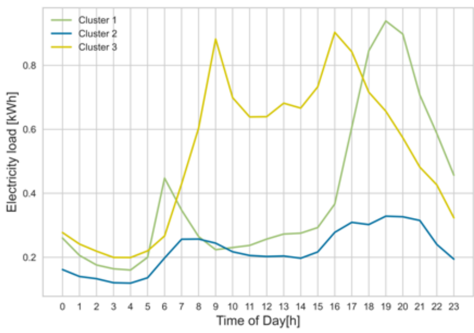


(b) Cluster 3

Figure A.7.: Day of the week distribution of household 11 for selected clusters



(a) Cluster centroids for all 365 days



(b) Cluster centroids for the first 100 days

Figure A.8.: Comparison of cluster results for 100 days and all 365 days (K=3)

References

Aurangzeb, K. (2019). Short Term Power Load Forecasting using Machine Learning Models for energy management in a smart community. In *2019 International Conference on Computer and Information Sciences (ICCIS)* (pp. 1–6). doi: 10.1109/ICCISci.2019.8716475

Awad, M., & Khanna, R. (2015). Support Vector Regression. In M. Awad & R. Khanna (Eds.), *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers* (pp. 67–80). Berkeley, CA: Apress. doi: 10.1007/978-1-4302-5990-9_4

Belyadi, H., & Haghighat, A. (2021). Chapter 4 - Unsupervised machine learning: clustering algorithms. In H. Belyadi & A. Haghighat (Eds.), *Machine Learning Guide for Oil and Gas Using Python* (p. 125-168). Gulf Professional Publishing. doi: 10.1016/B978-0-12-821929-4.00002-0

Burg, L., Gürses-Tran, G., Madlener, R., & Monti, A. (2021). Comparative analysis of load forecasting models for varying time horizons and load aggregation levels. *Energies*, *14*(21), 7128. doi: 10.3390/en14217128

Coignard, J., Janvier, M., Debusschere, V., Moreau, G., Chollet, S., & Caire, R. (2021). Evaluating forecasting methods in the context of local energy communities. *International Journal of Electrical Power & Energy Systems*, *131*, 106956. doi: 10.1016/j.ijepes.2021.106956

Cox, D. R., & Hinkley, D. V. (1979). *Theoretical Statistics.* New York: Chapman and Hall/CRC. doi: 10.1201/b14832

Dinesh, C., Makonin, S., & Bajic, I. V. (2019). Residential power forecasting using load identification and graph spectral clustering. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *66*(11), 1900–1904. doi: 10.1109/TCSII.2019.2891704

EIA U.S. Government. (2013). *Homes show greatest seasonal variation in electricity use.* Retrieved 2022-06-17, from `https://www.eia.gov/todayinenergy/detail.php?id=10211`

Etapart AG. (n.d.). *Fossile brennstoffe vs. regenerative energien.* Retrieved 2022-08-04, from `https://www.etapart.com/de/wissen/energietraeger/fossile-brennstoffe-vs-regenerative-energien`

European Commission. (n.d.). *Energy communities.* Retrieved 2022-08-04, from `https://energy-communities-repository.ec.europa.eu/energy-communities_en`

Flor, M., Herraiz, S., & Contreras, I. (2021). Definition of residential power load profiles clusters using machine learning and spatial analysis. *Energies*, *14*(20), 6565. doi: 10.3390/en14206565

Gong, H., Rallabandi, V., McIntyre, M. L., Hossain, E., & Ionel, D. M. (2021). Peak reduction and long term load forecasting for large residential communities including smart homes with energy storage. *IEEE Access*, *9*(99), 19345–19355. doi: 10.1109/ACCESS.2021.3052994

Han, J., Kamber, M., & Pei, J. (2012). 3 - Data Preprocessing. In J. Han, M. Kamber, & J. Pei (Eds.), *Data Mining (Third Edition)* (pp. 83–124). Boston: Morgan Kaufmann.

doi: 10.1016/B978-0-12-381479-1.00003-4

Hou, T., Fang, R., Tang, J., Ge, G., Yang, D., Liu, J., & Zhang, W. (2021). A novel short-term residential electric load forecasting method based on adaptive load aggregation and deep learning algorithms. *Energies*, *14*(22), 7820. doi: 10.3390/en14227820

Humeau, S., Wijaya, T. K., Vasirani, M., & Aberer, K. (2013). Electricity load forecasting for residential customers: Exploiting aggregation and correlation between households. In *2013 sustainable internet and ict for sustainability (sustainit)* (pp. 1–6). doi: 10.1109/SustainIT.2013.6685208

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). Unsupervised Learning. In G. James, D. Witten, T. Hastie, & R. Tibshirani (Eds.), *An Introduction to Statistical Learning: with Applications in R* (p. 373-418). New York, NY: Springer US. doi: 10.1007/978-1-0716-1418-1_12

Jeriha, J. (2019). Energy community definitions. Retrieved 11.04.2022, from https://www.compile-project.eu/wp-content/uploads/Explanatory-note-on-energy-community-definitions.pdf

Jiang, L., Wang, X., Li, W., Wang, L., Yin, X., & Jia, L. (2021). Hybrid multitask multi-information fusion deep learning for household short-term load forecasting. *IEEE Transactions on Smart Grid*, *12*(6), 5362–5372. doi: 10.1109/TSG.2021.3091469

Koirala, B. P., Koliou, E., Friege, J., Hakvoort, R. A., & Herder, P. M. (2016, April). Energetic communities for community energy: A review of key issues and trends shaping integrated community energy systems. *Renewable and Sustainable Energy Reviews*, *56*, 722–744. doi: 10.1016/j.rser.2015.11.080

Koponen, P., Diaz Saco, L., Orchard, N., Vorisek, T., & Togeby, M. (2008). *Definition of smart metering and applications and identification of benefits.*

Lee, S., Whaley, D., & Saman, W. (2014). Electricity Demand Profile of Australian Low Energy Houses. *Energy Procedia*, *62*. doi: 10.1016/j.egypro.2014.12.370

Leggett, T. (2006). *Air Conditioning (Industrial Report) - UK - August 2006 - Market Research Report.* Retrieved 2022-07-28, from https://reports.mintel.com/display/228786/#

Lusis, P., Khalilpour, K. R., Andrew, L., & Liebman, A. (2017). Short-term residential load forecasting: Impact of calendar effects and forecast granularity. *Applied Energy*, *205*, 654–669. doi: 10.1016/j.apenergy.2017.07.114

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, *5.1*, 281–298. (Publisher: University of California Press)

Müller, M. (2007). Dynamic Time Warping. In *Information Retrieval for Music and Motion* (pp. 69–84). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-540-74048-3_4

Nembrini, S., König, I., & Wright, M. (2018). The revival of the Gini Importance? *Bioinformatics (Oxford, England)*, *34*. doi: 10.1093/bioinformatics/bty373

Ns, A. (2015). Time complexity analysis of support vector machines (SVM) in LibSVM. *International Journal of Computer Applications*, *128*, 975–8887. doi: 10.5120/ijca2015906480

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pirbazari, A. M., Sharma, E., Chakravorty, A., Elmenreich, W., & Rong, C. (2021). An ensemble approach for multi-step ahead energy forecasting of household communities. *IEEE Access*, *9*, 36218–36240. doi: 10.1109/ACCESS.2021.3063066

Pullinger, M., Kilgour, J., Goddard, N., Berliner, N., Webb, L., Dzikovska, M., ... Zhong, M. (2021). The ideal household energy dataset, electricity, gas, contextual sensor data and survey data for 255 uk homes. *Scientific Data*, *8*(1), 146. doi: 10.1038/s41597-021-00921-y

Rodrigues, F. M., Cardeira, C., Calado, J. M. F., & Melicio, R. (2022). Home energy forecast performance tool for smart living services suppliers under an energy 4.0 and cps framework. *Energies*, *15*(3), 957. doi: 10.3390/en15030957

Sander, J. (2010). Density-Based Clustering. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 270–273). Boston, MA: Springer US. doi: 10.1007/978-0-387-30164-8_211

Sari, B. (2016). Identification of Tuberculosis Patient Characteristics Using K-Means Clustering. *Scientific Journal of Informatics*, *3*. doi: 10.15294/sji.v3i2.7909

Schwanitz, V. J., Wierling, A., Zeiss, J. P., Beck, C. v., Koren, I. K., Marcroft, T., ... Dufner, S. (2021). *The contribution of collective prosumers to the energy transition in Europe - Preliminary estimates at European and country-level from the COMETS inventory* (Tech. Rep.). SocArXiv. (type: article) doi: 10.31235/osf.io/2ymuh

Shaqour, A., Ono, T., Hagishima, A., & Farzaneh, H. (2022). Electrical demand aggregation effects on the performance of deep learning-based short-term load forecasting of a residential building. *Energy and AI*, *8*, 100141. doi: 10.1016/j.egyai.2022.100141

Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., ... Woods, E. (2020). Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, *21*(118), 1-6.

Tits, M., Bernaud, B., Achour, A., Badri, M., & Guedria, L. (2020). Impacts of size and history length on energetic community load forecasting: A case study.. doi: 10.1109/COMPSAC48688.2020.00-61

Toshniwal, D., Chaturvedi, N., Parida, M., Garg, A., Choudhary, C., & Choudhary, Y. (2020). Application of clustering algorithms for spatio-temporal analysis of urban traffic data. *Transportation Research Procedia*, *48*, 1046–1059. doi: 10.1016/j.trpro.2020.08.132

Wijaya, T. K., Vasiran, M., Humeau, S., & Aberer, K. (2015). Cluster-based aggregate forecasting for residential electricity demand using smart meter data. In *2015 ieee international conference on big data (big data)* (pp. 879–887). doi: 10.1109/BigData.2015.7363836

World Resources Institute. (2018). *Tackling global challenges.* Retrieved 2022-08-04, from `https://www.wri.org/strategic-plan/tackling-global-challenges`

Xu, L., Pan, Y., Lin, M., & Huang, Z. (2017). Community load prediction: methodology and a case study. *Procedia Engineering*, *205*, 511–518. doi: 10.1016/j.proeng.2017

.10.405

Yildiz, B., Bilbao, J. I., Dore, J., & Sproul, A. (2018a). Household electricity load forecasting using historical smart meter data with clustering and classification techniques.. doi: 10.1109/ISGT-Asia.2018.8467837

Yildiz, B., Bilbao, J. I., Dore, J., & Sproul, A. B. (2018b). Short-term forecasting of individual household electricity loads with investigating impact of data resolution and forecast horizon. *Renewable Energy and Environmental Sustainability*, *3*, 3. doi: 10.1051/rees/2018003

Zhang, X. M., Grolinger, K., Capretz, M. A. M., & Seewald, L. (2018). Forecasting Residential Energy Consumption: Single Household Perspective. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 110–117). doi: 10.1109/ICMLA.2018.00024