Characterization of Residential Load Profiles and Evaluation of Their Suitability for Coupling with Photovoltaic and Storage Systems Using Machine Learning Methods

Bachelor Thesis

Kai Firschau

1999902

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

Reviewer:	Prof. Dr. rer. pol. Christof Weinhardt
Advisor:	M. Sc. Sarah Henni

5th of October, 2020

Abstract

Residential energy sharing communities comprising prosumers with photovoltaic- and battery storage systems are a current topic of research, with the objective to assist the transition towards decentralized, sustainable renewable energy generation structures. In this work, the relationship between the characteristics of households' load profiles and their potential suitability to operate as an energy sharing community was investigated using real load profile data from the Chicago area and the profit of 1,000 communities simulated using these load profiles. Firstly, a comprehensive set of electricity parameters derived from load profiles was compiled, partly by extensive cross-literature research, partly by constructing new electricity parameters, aiming to express the properties of high-dimensional load profiles in a concise, yet informative set of indicators. In addition, two approaches for calculating electricity parameters on community level were developed. Secondly, it was investigated whether Supervised Machine Learning models can be trained to predict the profit of a potential community based on the households' electricity parameters. Despite extensive hyperparameter optimization and the usage of several differently configured training datasets, the models were not able to predict the community profit with satisfactory accuracy, wherefore possible reasonings and recommendations for new approaches in further research were proposed. Thirdly, the average load profiles and electricity parameter value distributions of low- and high-performing communities were analyzed in-depth, resulting in the proposal of a set of hypotheses about potentially favorable properties of residential load profiles for operating in an energy sharing community.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Motivation

A crucial part of the global efforts to reduce CO2 emissions is decarbonizing the energy supply by providing electricity from sustainable, renewable energy resources instead of environmentally damaging, fossil-fueled power plants. This entails the process of decentralization: the change in the spatial distribution of the electricity production from very few, centralized and high-capacity supply units, such as nuclear and lignite power plants, towards a large number of decentralized, low-capacity supply units, such as wind power plants and residential photovoltaic plants. Decentralized energy supply structures pose new challenges regarding the optimal temporal and spatial matching of the electricity demand and the increasingly volatile electricity generation [1]. So called energy sharing communities, networks of several residential consumer and prosumer households, can provide an approach to help overcome these challenges, and to foster the transition from centralized energy systems with passive consumers into flexible networks of active users. [2]

In order to accelerate the sustainable development of the energy system, featuring decentralized structures with efficient energy sharing communities, a platform-based sharing economy model for energy sharing communities has been proposed in [1]. The annual electricity costs of 1,000 energy sharing communities, each consisting of five households, partly prosumers and partly consumers, have been simulated and compared to the electricity costs of the households operating individually to finally simulate the potential profit. Furthermore, the authors of [1] state that their work shows that an interesting path for future research would be to identify household characteristics leading to a beneficial combination in an energy sharing economy model.

Following this recommendation for future research, this work will investigate the following three research questions:

- Which statistical electricity parameters can be constructed from high-dimensional yearly load profiles in order to characterize their various properties in a lower-dimensional, yet highly informative manner?
- Can these electricity parameters serve as input features for the development of Machine Learning models that accurately predict the suitability of certain household combinations for energy sharing communities?
- Which patterns of certain parameter values could qualitatively indicate beneficial combinations of households that harbor considerable potential for creating a profitable energy sharing community?

In order to answer the research questions above, we will proceed with the following methodology: After analyzing related literature in the remaining part of Chapter 1., the available load profile data as well as the simulation data from [1] will be preprocessed, characterized an analyzed in detail in Chapter 2.

Subsequently, aiming to answer the first research question, we will present a comprehensive collection of versatile electricity parameters that can be derived from load profiles in Chapter 3., accompanied by several refinement approaches, two community parameter calculation approaches we propose in this work, and a variety of feature selection methods. In Chapter 4.1., diverse variants of Machine Learning models are trained and evaluated to investigate the second research question. In Chapter 4.2., to address the third research question, an in-depth exploratory analysis of the load profiles and selected parameter values of low- and high-performing communities is supposed to yield qualitative hypotheses about beneficial properties.

All programming operations for this work are conducted with Python, in particular using the Python libraries pandas [3], statsmodels [4] and scikit-learn [5]. To enable the convenient re-using of the analysis approaches and calculations proposed in this work for further research, the corresponding code and other materials are accessible in a public repository under the following permanent link:

https://github.com/firschau/bachelorthesis

1.2. Related Work

A broad body of literature exists on various data mining and machine learning applications based on electricity load profiles. A wide-spread methodology, for example used in [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], and [18], is the following: At first features, i.e. electricity parameters, are constructed from the load profiles. Subsequently, clustering algorithms are applied in order to discover different consumer groups and extract their representative load profiles. These are then used, for instance, to better understand residential consumption behavior, to enable energy suppliers to tailor new flexible electricity tariff structures, or to classify new customers regarding their closest representative load profile and therefore as a member of certain customer group. Thereby, most of the reviewed literature focused on residential load profiles, while some analyzed commercial building load profiles, such as [14]. Other approaches try to forecast *electricity* parameters from *socioeconomic* parameters, such as [19], or try to infer socioeconomic parameters directly from the load profiles, such as [20]. Another approach, as presented in [21], is to leverage weather data to predict electricity parameters. The approaches stated above offer some promising propositions for electricity parameters that allow us to summarize the complex properties of high-dimensional load profiles in few meaningful key figures. Regarding energy sharing communities, there is a vast body of very recent literature, for instance [22], where several optimal energy sharing behaviors of prosumers were identified, among others also certain energy consumption, generation and storage behaviors, but not with detailed regard to residential load profiles. In general, despite the apparent research interest, to our best knowledge, no investigations focusing purely on the relationship between the load profile shapes of the participants and the potential performance of energy sharing communities were conducted up to now, indicating a research gap that should be filled with this work.

2. Data Preparation

2.1. Data Preprocessing

In the following, the same dataset as in the underlying paper [1] will be used. We will work with the real electricity load profiles of 100,000 households from the Chicago area. From these load profiles, the **independent variables** (or *features*, using Machine Learning jargon) analyzed in this work, i.e. the electricity parameters**,** will be calculated, as presented in detail in Chapter 3.

An electricity load profile is defined as the "pattern of electricity demand for a consumer, or a group of consumers, over a given period of time" [10]. Each load profile covers the corresponding household's energy consumption of one year, measured in 30-minute-intervals, and therefore contains 17,520 timesteps and consequently features.

The raw load profile dataset is transformed such that the electricity load profiles are available as cleanly structured, storage-efficient univariate time series. To ensure a consistent mathematical notation throughout this work, particularly with regard to the parameter calculations, we will use the following formal definition for our set of load profiles L, containing $I = 100{,}000$ electricity load profiles, using a two-layered index to differentiate the time series values by both, the days of the year $d = 1, \dots, D,\ D = 365$, and the 30-minute timesteps within each day, $t = 1, \dots, T,\ T = 47$:

$$L = \{E^i\}, \qquad E^i = \{E^i_{d,t}\}, \qquad i \in \{1, \dots, I\}, d \in \{1, \dots, D\}, t \in \{1, \dots, T\}.$$

In order to maintain as much information as possible, the load profiles are, as recommended in [13], not subjected to a z-score standardization, Instead, they are normalized with their respective total annual consumption, as recently proposed in [23]; By this, differences between households regarding their general level of electricity consumption, for example due to a higher number of inhabitants, are balanced out, while the behavioral consumption patterns of the households are still maintained:

$$\tilde{L} = \{\tilde{E}^i\}, \qquad \tilde{E}^i = \left\{\frac{E^i_{d,t}}{\sum_{d=1}^{D}\sum_{t=1}^{T} E^i_{d,t}}\right\}, \quad i \in \{1, \dots, I\}, d \in \{1, \dots, D\}, t \in \{1, \dots, T\}.$$

In order to maintain memory efficiency and accelerate the required computations in Chapter 2.2, we use a randomly drawn sample of size $n = 5000$ household load profiles for the data analysis in the following chapters.

As announced in Chapter 1.1., it is to be analyzed whether different parameters, calculated from household load profiles, can be used to draw conclusions about the suitability of households for energy sharing communities. These consist of five households each, selected from the dataset described above, of which two are households using a PV system *and* a battery storage system, one household using only a PV system, and two households that are solely consumers. The *suitability* analyzed in this work is quantified by means of two **dependent variables** (or *targets*, using Machine Learning jargon) that were simulated in [1] for 1,000 energy sharing communities:

1. The *Community Profit*, i.e. the cumulative profit of all households participating in an energy sharing community. That means, the higher this value, the better is the performance of the households operating as a community. The relationship between the Community Profit and the characteristics of the load profiles of all households in the corresponding community will be investigated.
2. The *Community Gain*, i.e. the difference between the community profit and the sum of the individual profits of the households in case they would not participate in the energy sharing community. That means, the higher this value, the more the households benefit from participating in the community in comparison to operating individually. The relationship between the Community Gain and the characteristics of the two consumer households of the corresponding community, i.e. that of consumer households that are coupled with prosumers using PV and partly battery storage systems, will be investigated.

Since the objective of this work is not to predict exact, quantitative values of Community Profit or Community Gain, but rather to enable qualitative statements about the general tendency of suitability of certain household combinations for participating in energy sharing communities, the targets will be discretized. Although discretization of variables entails some disadvantages, such as possible loss of information, it also harbors several advantages, for instance a higher memory efficiency and faster processing, as stated in [24]. Additionally, this will facilitate the explorative analysis of the parameter values for low-, medium- and high-performing energy communities. The discretization is performed by calculating the terciles of each target, setting them as the boundaries of three intervals ("Low", "Medium" and "High") and then mapping each value of the targets to its corresponding interval. As a result, for each target variable, we will obtain a series of discrete, uniformly distributed values, one for each community. The distributions of both variables including the terciles as interval boundaries, are depicted in Figure 1, revealing that the distribution of the Community Profit is slightly left-skewed (Skewness of -0.310) and the distribution of the Community Gain is slightly left-skewed (Skewness of +0.329), which indicates that there seem to exist communities whose general profit as a community is, in relative terms, lower than the increase in profit they obtain from participating in the community compared to operating individually, and vice versa.

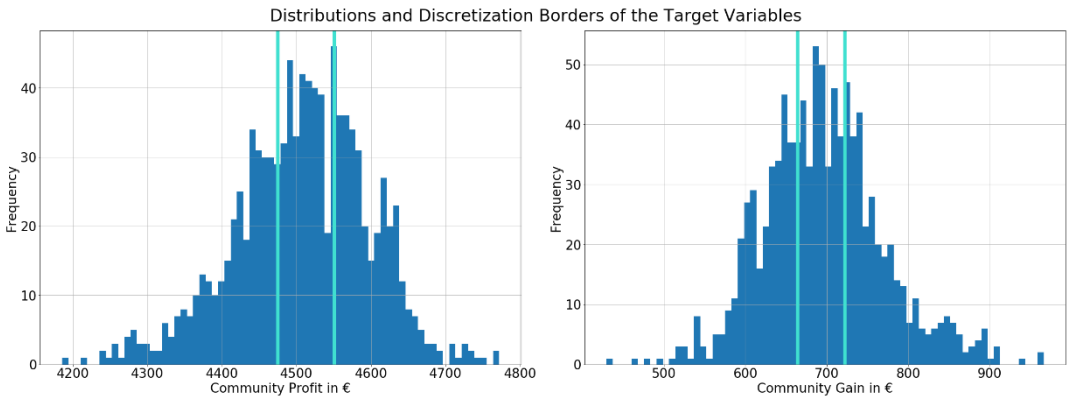

Figure 1: Distributions of Community Profit (left) and Community Gain (right) for 1000 energy sharing communities, with respective tercile values depicted as vertical lines

2.2. Load Profile Analysis

2.2.1. Visualization

In this chapter, a visualization of different load profiles should give a first insight into the properties of the available data.

At first, the load profiles as a whole are be visualized to analyze the load profile characteristics on *yearly level*. Therefore, a randomized sample of six yearly load profiles is selected and plotted in Figure 2.

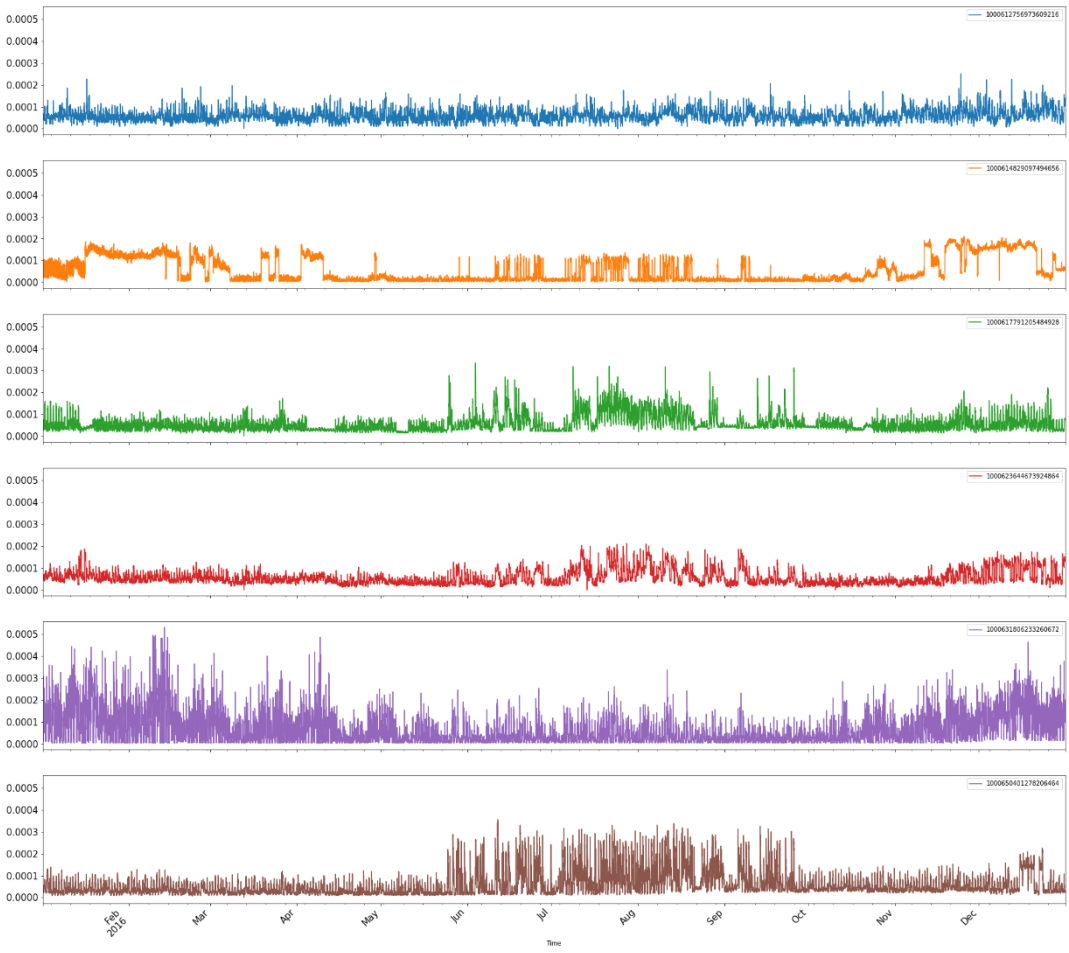

Figure 2: Visualization of six randomly selected yearly load profiles

The visualization suggests that there seem to be significant differences in the yearly load profile shapes regarding trends and seasonality, for instance, the fifth household features a significantly higher consumption during the winter months, while for the sixth household the exact opposite applies. To further explore these yearly differences, a detailed seasonality analysis will be conducted in Chapter 2.2.2.

In addition, the average daily load profiles of the same randomized sample are plotted in order to analyze exemplary load profile characteristics on *daily level*. It is presumed that the daily load profiles differ with regard to the different seasons due to climatic circumstances. For instance, households might use air conditioning systems in the summer months, or electrical heating systems in the winter months, or due to moderate temperatures none at all in spring or autumn. Furthermore, we suspect that the different times of sunrise and sunset cause a temporal shift in the electricity consumption behavior, for instance with regard to lighting, depending on the season. Hence, the exemplary plotting, shown in Figure 3, is conducted by comparing the average daily load profiles for three different seasons, which are defined and used throughout this work as follows, based on similar climatic conditions:

- *Summer*: June to September.
- *Winter*: December to March.
- *Spring and Autumn*: April, May, October and November.

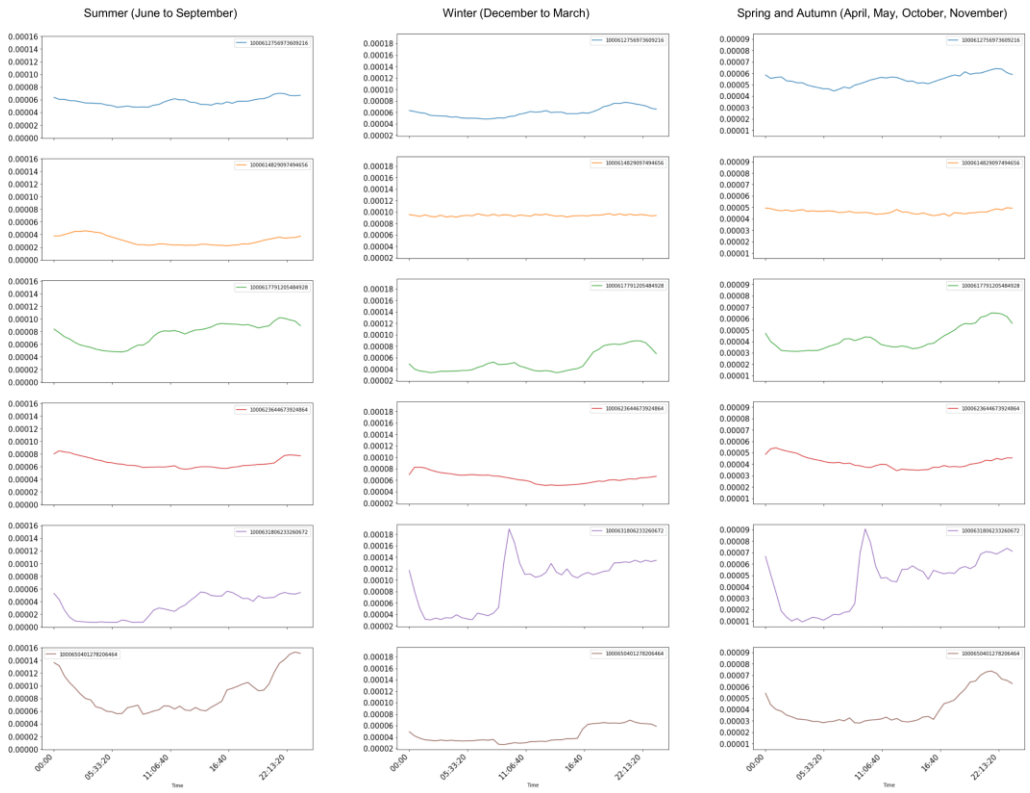

Figure 3: Visualization of six randomly drawn daily average load profiles, differentiated by three seasons

The visualization of these randomly selected average load profiles suggest that the daily load profile shapes vary considerably with respect to different seasons. For instance, the fifth load profile displays a lot steeper rise at forenoon on spring, autumn and winter days than on summer days, and the sixth load profile features a significantly flatter and earlier rise in consumption in the evening hours in winter than in the other seasons. In addition, the sixth load profile features a clearly lower base load in the winter months, which might be reasoned by this household using air conditioning systems in the other months. Due to these clear seasonal differences in load profile shapes indicated by the visualized sample, the seasonal variation in the daily load shapes will be considered for the parameter construction.

2.2.2. Clustering

The visualization of randomly selected load profiles on year-level in Chapter 2.2.1. implied that there are various yearly load shapes, with different types of seasonal trends. Differences in the yearly trend might be caused by seasonal behavioral patterns of the households, for example using air conditioning systems in the summer months, and/or electrical heating in the winter months. Consequently, the differences could have, for instance in connection with fluctuating PV energy generation throughout the year, an impact on the suitability that a combination of households provides for an energy sharing community. To analyze these differences in our households' yearly behavioral patterns, and to explore if we can condense them into a single, informative electricity parameter, Unsupervised Machine Learning is utilized, attempting to organize the raw annual load profiles into well-separated clusters that allow conclusions on annual behavioral patterns.

Prior to using clustering algorithms, we apply the "*Seasonal-Trend Decomposition Procedure based on Loess"* (STL), as originally presented in [25], using the Python implementation of statsmodels [4], to our annual load profiles. Thereby, we can extract the pure yearly trend component of the time series, which leads to significantly smoothed load shapes and therefore facilitates identifying differences in the yearly seasonal fluctuations clearly, as highlighted in the following Figure 4.

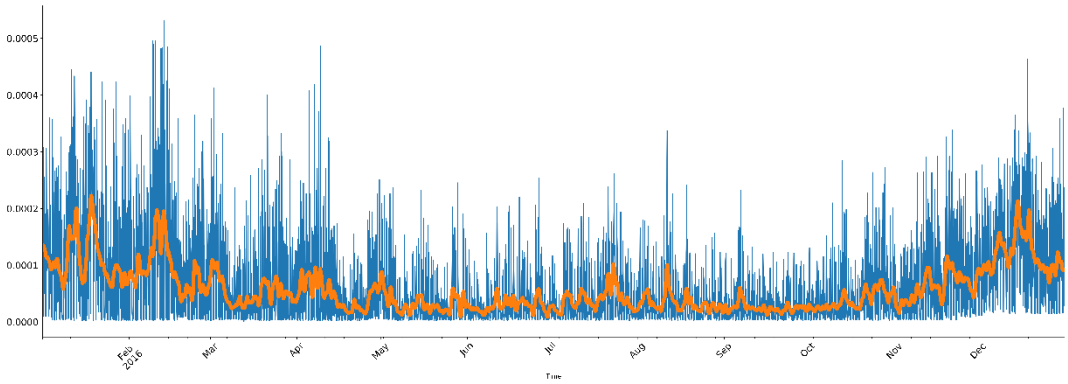

Figure 4: Comparison of a randomly selected yearly load profile, with (orange) and without (blue) the application of STL smoothing

Regarding the choice of clustering algorithm, the widely used, centroid-based k-Means-Algorithm is be applied, since it is, as stated in [23], amongst the most widely used clustering algorithms as it yields high versatility as well as scalability. In addition, it permits directly selecting the number of generated clusters, contrary to density-based clustering algorithms such as DBSCAN or SLINK, and a convenient heuristic to select the optimal number exists, as presented later in this chapter. The following summary of the algorithm is based on the detailed explanations in [26]: Given an integer k and a set of n samples X, The k-Means-Algorithm's goal is to choose k cluster centers C such that they minimize the objective function:

$$\Phi = \sum_{x \in X} \min_{C_i \in C} \|x - C_i\|^2.$$

The cluster centers are determined as follows:

1. The algorithm starts by randomly selecting initial cluster centers:
$$C = \{c_1, c_2, \dots, c_k\}.$$
2. The clusters C_i are generated by assigning each sample $x \in X$ to its closest cluster center:
$$C_i = \left\{ x \in X \mid \underset{i=1 \dots k}{\operatorname{argmin}} \|x - c_i\|^2 = \mathrm{i} \right\}, \forall\, i = 1 \dots K.$$
3. Each cluster center c_i is updated to be the arithmetic mean of the values of all samples that are currently assigned to it:
$$c_i = \frac{1}{|C_i|} \sum\nolimits_{x \in C_i} x\,, \forall\, i = 1 \dots K.$$
4. Steps 2. and 3. are repeated, until a pre-defined maximum number of iterations is reached, or the clusters C_i do not change anymore, i.e. the algorithm converges.

In order to discover a suitable number of clusters k, the heuristic *Elbow Method* will be applied, as explained for instance in [27]: we will run the k-Means-Algorithm multiple times with different values of k, and then plot the respective values for the *Within-Cluster-Sum-of-Squares*, which describes the intra-cluster variance and serves as the optimization criterion that k-Means tries to improve in every iteration. The scikit-learn implementation of k-Means [5] will be applied to compute the results in the following sections, using the following hyperparameters:

- k-Means++ augmentation to speed up convergence. Details can be found in [26].
- Fixed random number generation seed to ensure reproducibility of the results.
- Ten initializations with different centroid seeds.
- Maximum number of iterations per run of 1,000.

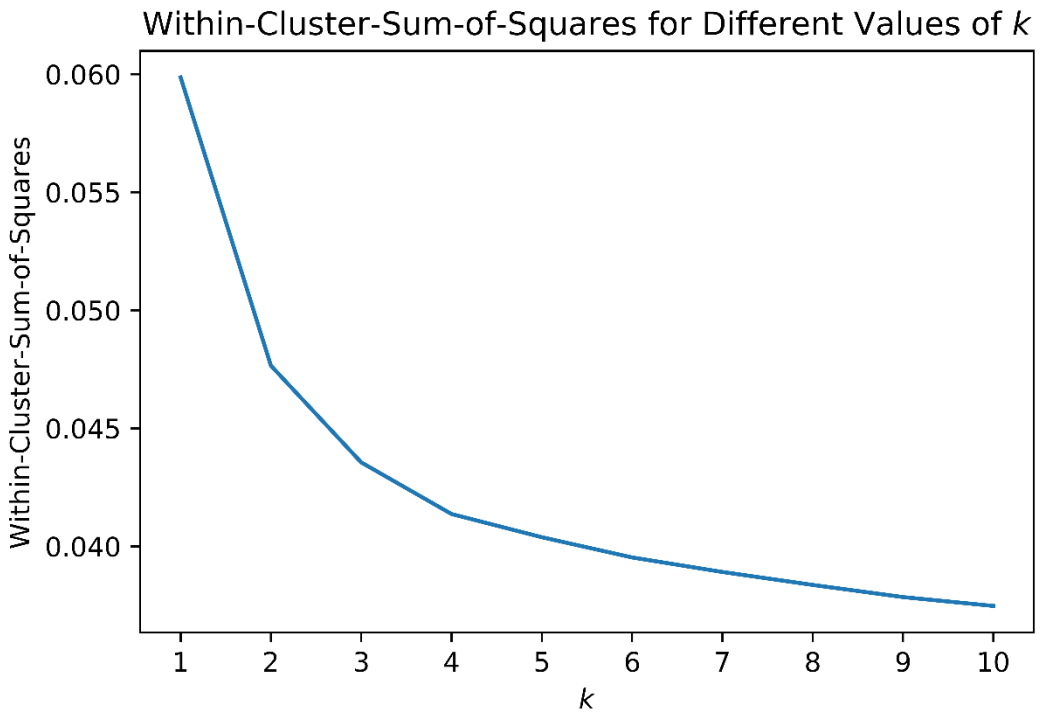

Figure 5: "Elbow Method" - Evolution of the Within-Cluster-Sum-of-Squares with increasing values of the number of clusters k

Figures 5 depicts the different values of WCSS for increasing k. We select 4 as our choice for k, since for values larger than $k = 4$, the WCSS decreases clearly less, i.e. an "elbow" can be recognized. In addition, $k = 4$ yields a concise and well interpretable number of clusters. In the following Figure 6, the resulting cluster centers c_0, c_1, c_2 and c_3, and their respective cluster sizes are visualized.

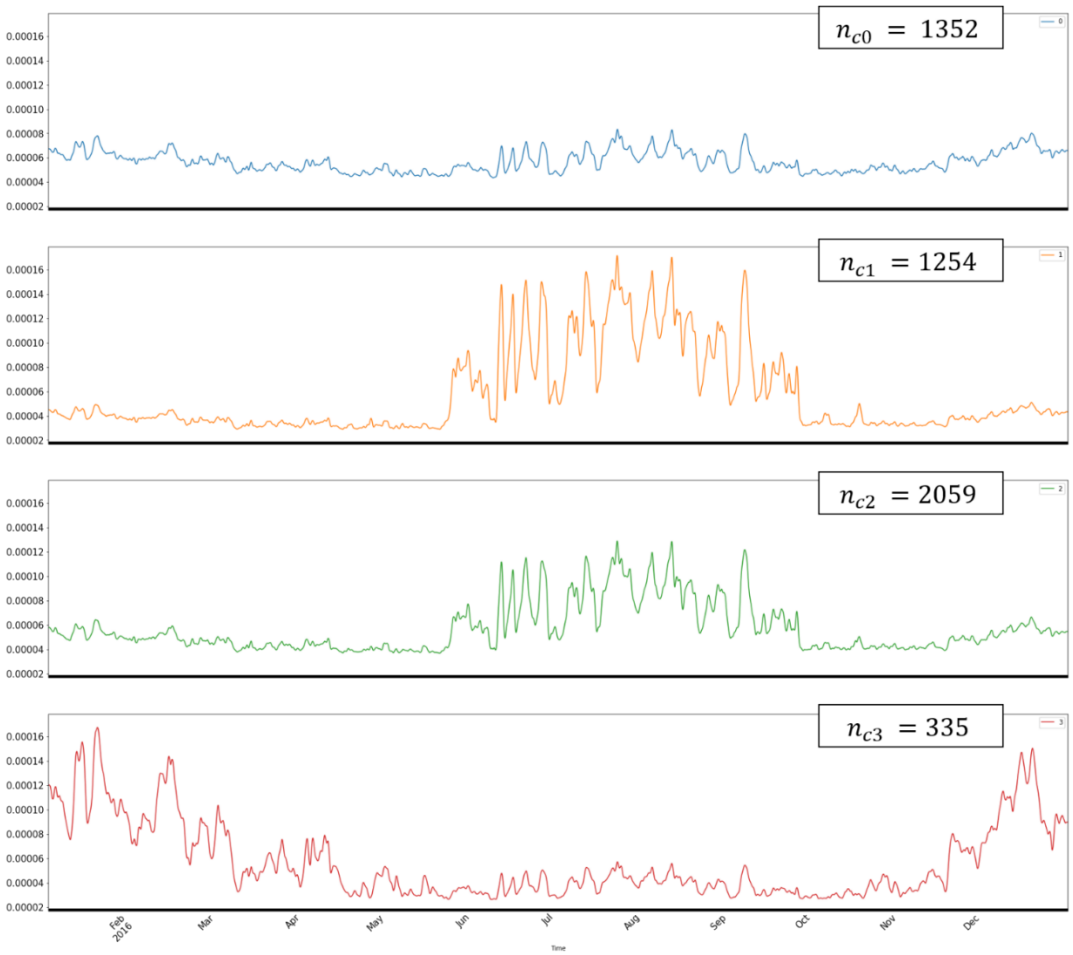

Figure 6: Visualization of the obtained cluster centers with corresponding cluster size

Observing the cluster centers in Figure 6 confirms that the observations made for the randomized household sample in Figure 2 seem to generalize for the analyzed sample of size $n = 5000$: The cluster centers of cluster 1 and 2 show significant peaks of varying strength in the summer months, approximately from June to September. This increase in consumption in the summer months might indicate that the households belonging to these clusters use air conditioning systems, that entail a high energy consumption. Contrary to this, it strikes that the cluster center of the relatively small cluster 3 displays sharp consumption peaks in the winter months, approximately between the end of November to late March. This could, for instance, be caused by the respective households using electrical heating systems. While the cluster centers of the clusters 1, 2 and 3 show clear annual trends, the cluster center of cluster 0 features a relatively flat load profile, i.e. the energy consumption is rather equally distributed across the year and there are no clear seasonal fluctuations. Following the hypotheses proposed above, this might be caused by the respective households using either no air conditioning or electrical heating systems at all, or using both, resulting in a fairly counterbalanced yearly load profile.

The latter seasonality analysis shows that there are significant differences in the yearly trend within the annual load profiles. In order to include these in our electricity parameter set and thereby increase its explanatory power, a new parameter expressing the ratio of electricity consumption in the summer months to the electricity consumption in the winter months, which we will label as the *Summer-Winter-Ratio,* is proposed in this work and added to the parameter collection in Chapter 3. With regard to seasonal fluctuations of the PV generation of prosumers, the yearly trend patterns, as described by the Summer-Winter-Ratio, could have a significant influence on the community performance on year-level.

3. Feature Engineering

3.1. Methodology

Machine Learning models rely heavily on the informative content of the features they use; as stated for instance in [28], it has been proven that using electricity parameters, that summarize complex load profile properties concisely, instead of the actual consumption data, is a very efficient approach regarding the classification of customers, which is supported by the vast amount of authors utilizing electricity parameters in their works on Data Mining based on load profiles. In order to find a combination of electricity parameters that characterize the properties of voluminous and high-dimensional residential load profiles in a concise and low-dimensional, yet precise manner, and therefore provide a high-quality set of input features for our evaluations regarding the suitability of households for energy sharing communities in Chapter 4., a broad collection of different parameters is compiled in Chapter 3.2. For this, extensive literature research was carried out on existing parameters, some of which were modified, and several, to our best knowledge new parameters were elaborated. Since many of the previous works make use of relatively small sets of parameters, and it is recommended to investigate more parameters in [12], this summarizing, cross-literature collection of parameters could provide a fundamental basis for further research on Data Mining based on load profiles. The complete parameter collection can be found in Table 1 in Chapter 3.2. Two considerations for the optimization of the parameters are discussed, and two approaches regarding their calculation are proposed in the following.

To increase the **robustness against outliers**, skewed distributions and anomalies within the load profiles, as proposed in [21], the parameter calculations are modified so that, where appropriate, the median is used instead of the arithmetic mean. Furthermore, another approach that could be applied is to modify the parameters such that “true” maximum and minimum values are replaced with “near-peak” and “near-base” values, for example the 97.5^{th} and 2.5^{th} percentile load. As stated in [21], for many buildings, these values are more stable as well as more relevant than the actual extreme values.

A similar approach, but with regard to commercial buildings, is also proposed in [14], where the 95th and 15th percentiles are used. Using the "near-peak" and "near-base" load vales can be particularly interesting if consumption behavior tendencies over very long timescales are to be analyzed, as stated in [21]; however, these techniques will not be applied in this work as we are interested in the extrema of consumption behavior and their implications.

To **increase** the **explanatory power**, as proposed in, among others [21], all parameters except those on yearly level are differentiated regarding the following aspects:

1. *Yearly differences: Seasons.* As the randomized sample analysis in Chapter 2.2.1. has indicated, it could be useful to compute the parameters separately for different seasons in order to reflect seasonal differences in the daily consumption patterns and mitigate the information loss on seasonal level caused by averaging. The cluster analysis in Chapter 2.2.2., showing strong differences in yearly trends, supports this. Furthermore, similar differentiations are also proposed in [14] and [21]. For further applications, we recommend that these seasons should always be adapted to the respective climatic conditions of the region under investigation.
2. *Weekly differences: Working- and weekend-days.* Many households, especially with a high share of professionals, could show significant differences in their electricity consumption behavior during the week and on weekends, which should also be reflected in the parameters, as proposed in [14] and [21].

There are several possibilities regarding the methodology of **parameter calculation**. As mentioned beforehand, the intention of analyzing electricity parameters instead of raw load profile data is the reduction of dimensionality; if all the parameters would be considered for every day of each household of every community, the effect of dimensionality reduction would vanish. Therefore, some form of averaging, regarding the load profiles and/or the parameters, is required.

With regard to **load profile averaging**, two exemplary approaches are, both presented in [12]:

1. *Time domain averaging (Longitudinal)*, which is averaging the load profiles across certain time periods, such as for instance day, week or year. This entails loss of information on intra-daily, and on seasonal components. The latter drawback is to be mitigated in this work using the seasonal differentiation mentioned above.
2. *Space Domain Averaging (Cross-sectional)*, which consists of averaging across individual households and entails loss of information on each individual household. Since we are interested in the characterization of individual communities, this approach is out of question regarding this work.

With regard to **parameter averaging**, one generic method, for instance used in [12], [21] and [23] with respectively varying specifications, is to calculate the parameter of interest separately for each day of a load profile.

Afterwards, a measure of central tendency of the daily parameter values is set as the load profile's final parameter value. In the case of [23], a discrete parameter is considered, wherefore the statistical mode is used as a measure, while in [12] and [21] the used parameters are continuous, wherefore the applied measure of central tendency is the non-parametric median. As the parameters considered here are partly continuous and partly discrete, the ideas of the latter works are combined.

As in this work, contrary to the existing approaches presented in Chapter 1.2., not single household load profiles, but communities consisting of *five* household load profiles each are considered, a new dimension of required averaging respectively aggregation arises: the **aggregation on community-level**. Considering and combining the various averaging methods presented above and expanding them to the context of energy sharing communities, the following two community parameter calculation approaches are designed and used throughout this work, with minor modifications regarding the two different classification tasks, community profit and gain, as defined in Chapter 2.1.:

1. *Parametrization First, Aggregation Second (PFAS)*: For this approach, in the first step, the daily parameter values are computed for each day of every household and, where appropriate, separately for the three seasons defined in Chapter 2.2.1. and for work- and weekend-days. Secondly, for each household, we compute the median (respectively mode) of the daily parameter values, in order to obtain a representative value for each household, which is robust against single, non-representative outliers. In the third step, we will sum up the median parameter values of all individual households, respectively the consumer households when considering the community gain, within each community in order to obtain the final aggregated parameter values for each community.
2. *Aggregation First, Parametrization Second (AFPS)*: For this approach, we start by calculating the pointwise sum of each community's five, respectively the two consumers' household load profiles to obtain summarized community load profiles. In the second step, the latter are used to calculate a representative daily average load profile for every community using time-domain averaging as presented above, whereby likewise a distinction is made between the three seasons defined in Chapter 2.2.1. and for work-and weekend-days, resulting in six representative daily load profiles for each community. Hereby, the median is used in the averaging process, since a brief analysis of the load profile sample has shown that the distributions of the load values at the individual time steps across all days are positively skewed for all households[1] and therefore not symmetrically distributed, which indicates the use of a non-parametric measure of central tendency.

[1] For each timestep and household, the skewness of the distribution of load values was computed. Then, for each timestep, the median value across all households was calculated. The median sample skewness of the load values for the different time steps moves in the range [0.628, 1.271], suggesting that the distributions are likely right-skewed.

Then, in the third step, the final community parameter values are calculated on each community's six representative daily load profiles.

The comparison of the algorithms of the two proposed calculation methods suggests that due to greater complexity a significantly higher runtime can be expected for PFAS than for AFPS: The latter starts with the (very efficient) point-wise summation and successive six-fold load profile averaging per community, and subsequently computing the respective parameter calculation function *once* on each of the six average load profiles, resulting in six calculations per parameter for *one community*. Contrary to that, PFAS requires computing the (partly considerably complex) parameter calculation functions for all 365 days of every load profile, resulting in 365 calculations per parameter solely for *one household*. For further theoretical or practical applications in contexts where large datasets are used or critical runtime requirements exist, we propose that this aspect should be considered carefully.

Both community parameter calculation approaches are visualized in the following Figures 7 and 8 below. These two community parameter calculation approaches, combined with the two target variables Community Profit, whose dependence on *all* household parameters within the communities is investigated, and Community Gain, whose dependence on *only* the two consumer parameters of each community is analyzed, result in four subsets of parameter value and target combinations (*PFAS – Profit, PFAS – Gain, AFPS – Profit, AFPS – Gain*), in the following referred to as *parameter value subsets*, to be analyzed in the evaluations in Chapter 4. This is supposed to cover the widest possible range of different analysis options.

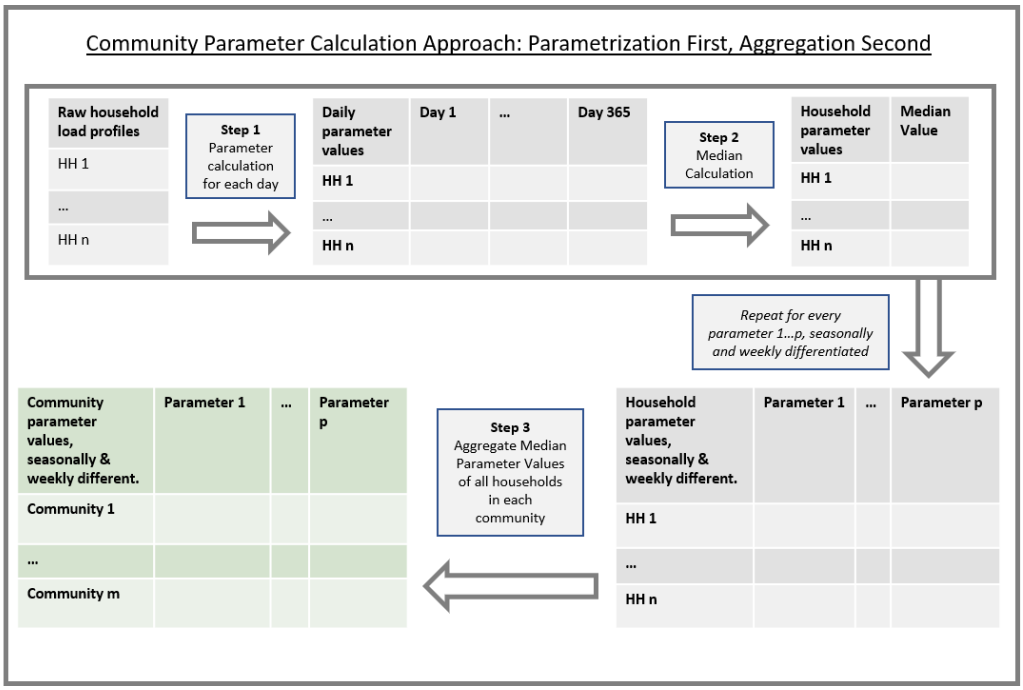

Figure 7: The process of the Parametrization First, Aggregation Second parameter calculation approach

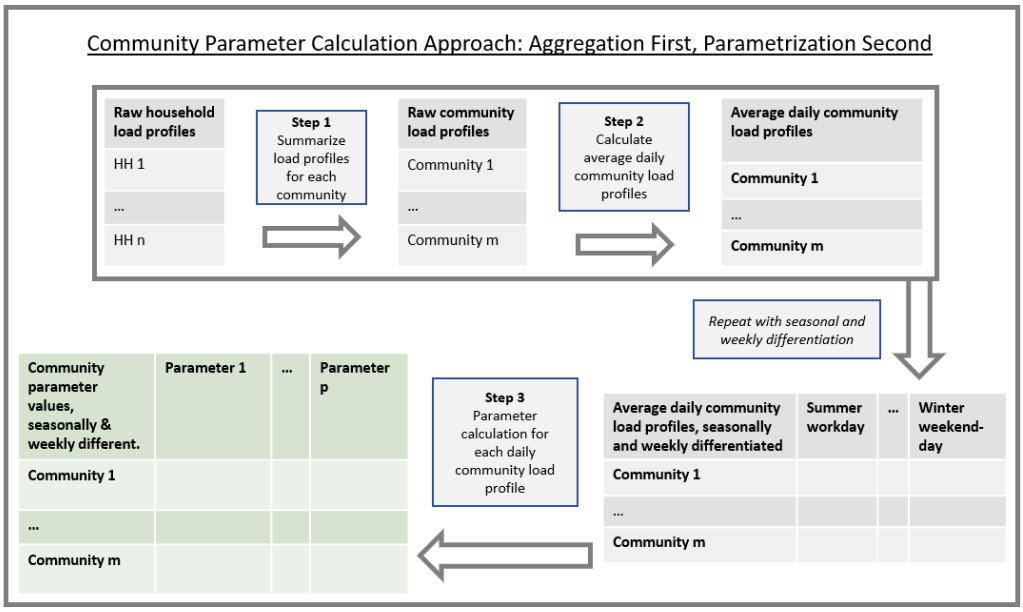

Figure 8: The process of the Aggregation First, Parametrization Second parameter calculation approach

3.2. Parameter Collection

In the following Table 1, the compiled parameters are presented. They are grouped by different approaches, mathematically defined and shortly described. As mentioned in Chapter 3.1., regarding the further usage of the parameters for future research, the definitions of the seasons should be adapted to the respective climatic conditions of the investigated scenario. In addition, other definitions, such as lunchtime or off-hours, should also be adapted to local socioeconomic circumstances. In order to enable convenient re-using of the presented parameters, the corresponding Python-code for the calculation is available in the thesis repository.

Electricity Parameter Collection				
Approach	**Parameter**	**Mathematical Definition**	**Comment**	**References**
Overall Load Shape	Daily Load Factor	$P_{LF}^{i} = \left(\frac{\frac{1}{T}\sum_{t=1}^{T} E_{d,t}^{i}}{\max\limits_{1 \le t \le T} E_{d,t}^{i}} \right)$	Measures the "Peakiness" of a household's load profile, values closer to one indicate a more even consumption across the day. [9]	[6], [7], [8], [9], [10], [12], [13], [19]
	Daily Non-Uniformity Coefficient	$P_{NUC}^{i} = \left(\frac{\frac{1}{T}\sum_{t=1}^{T} E_{d,t}^{i}}{\max\limits_{1 \le t \le T} E_{d,t}^{i}} \right)$	Analogue to the Daily Load Factor but describing the variation towards the base load instead of the peak load.	[7], [8], [10]
	Daily Range	$P_{R}^{i} = \frac{\min\limits_{1 \le t \le T} E_{d,t}^{i}}{\max\limits_{1 \le t \le T} E_{d,t}^{i}}$	The range of the daily consumption values, expressed as a difference (standard range, absolute) or a ratio (used here, relative).	[10], [14], [16], [17]

Approach	Parameter	Mathematical Definition	Comment	References
Impact of time ranges[2]	Night Impact	$P_{NI}^{i} = \frac{1}{3} \frac{\frac{1}{N} \sum_{n \in N} E_{d,n}^{i}}{\frac{1}{T} \sum_{t=1}^{T} E_{d,t}^{i}}$	The impact of the nightly consumption on the total daily consumption. The night half-hours are defined as: $N = \{0 \dots 12\} \cup \{46,47\}$	[6], [7], [8], [10], [13]
	Lunch Impact	$P_{LI}^{i} = \frac{1}{8} \frac{\frac{1}{L} \sum_{l \in L} E_{d,l}^{i}}{\frac{1}{T} \sum_{t=1}^{T} E_{d,t}^{i}}$	The impact of the lunch hours on the total daily consumption. The lunch half-hours are defined as: $L = \{22 \dots 26\}$	[6], [8], [10], [13]
	End-of-Work-Impact	$P_{EoWI}^{i} = \frac{1}{6} \frac{\frac{1}{W} \sum_{e \in EoW} E_{d,e}^{i}}{\frac{1}{T} \sum_{t=1}^{T} E_{d,t}^{i}}$	The impact of the end-of-work hours on the total daily consumption. The end-of-work half-hours are defined as: $EoW = \{32 \dots 38\}$	*Proposed*
Slopes[3]	Morning Slope	$P_{MS}^{i} = E_{d,20}^{i} - E_{d,12}^{i}$	Difference between the load values at 10:00 am and 06:00 am.	[16], [17]
	Night Slope	$P_{NS}^{i} = E_{d,46}^{i} - E_{d,42}^{i}$	Difference between the load values at 11:00 pm and 09:00 pm.	[17]
Extrema	Daily Maximum Demand	$P_{Max}^{i} = \max_{1 \le t \le T} E_{d,t}^{i}$	Maximum of the daily load profile.	[12], [17], [19]
	Maximum Time of Use	$P_{Max\_ToU}^{i} = \operatorname{argmax}_{1 \le t \le T} E_{d,t}^{i}$	Daytime at which the daily maximum consumption occurs. Since this value is discrete, it will be aggregated with the mode instead of the median when using the PFAS approach.	[12], [17], [19]
	Daily Minimum Demand	$P_{Min}^{i} = \min_{1 \le t \le T} E_{d,t}^{i}$	Minimum of the daily load profile.	*Proposed*
	Minimum Time of Use	$P_{Min\_ToU}^{i} = \operatorname{argmin}_{1 \le t \le T} E_{d,t}^{i}$	Daytime at which the daily minimum consumption occurs. Since this value is discrete, it will be aggregated with the mode instead of the median when using the PFAS approach.	*Proposed*
Frequency-based	FFT Peak	$P_{FFTP}^{i} = \max_{1 \le f \le F} \left\lvert FFT_f\left(E_{d,t}^{i}\right) \right\rvert$	Maximum (used here) or sum) of the absolute values of the Fast Fourier Transformation	[9], [16], [17], [20]

[2] The shown parameters represent only a selection of time ranges; this parameter can also be constructed for arbitrary other time ranges, as presented for example in [8].

[3] Analogous to the impact of time ranges, this parameter could also be constructed as the slope between any two points in time.

Approach	Parameter	Mathematical Definition	Comment	References
Central Statistical Moments[4]	Variance	$P^i_{\sigma^2} = \frac{1}{T}\sum_{t=1}^{T}(E^i_{d,t} - \bar{E}^i_d)^2$	The sample standard deviation of the daily load profile.	[17]
	Skewness	$P^i_{\hat{\gamma}} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{(E^i_{d,t} - \bar{E}^i_d)}{\sqrt{\sigma^2}}\right)^3$	The sample skewness of the daily load profile.	[9]
	Kurtosis	$P^i_{\hat{\omega}} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{(E^i_{d,t} - \bar{E}^i_d)}{\sqrt{\sigma^2}}\right)^4$	The sample kurtosis of the daily load profile.	[17]
Similarity to Standard Profiles[5]	PV Correlation	$P^i_{\rho_{E^i_d, PV_d}} = \frac{cov(E^i_d, PV_d)}{\sigma_{E^i_d}\sigma_{PV_d}}$	The Pearson correlation coefficient with a standard load profile of a photovoltaic system PV_d.	*Proposed*
Weekly Differences	Workday-Weekend-Ratio	$P^i_{WWR} = \frac{\underset{w\in W}{\text{med}}(\sum_{t=1}^{T} E^i_{w,t})}{\underset{nw\in NW}{\text{med}}(\sum_{t=1}^{T} E^i_{nw,t})}$	The ratio of a typical daily consumption on workdays W, and on non-workdays NW. Originally proposed for commercial buildings load profiles, we transfer this parameter to residential load profiles as reasoned in 3.1. Higher values indicate a higher consumption on weekdays than on non-weekdays and vice versa. With AFPS, this parameter reduces to the ratio of total consumptions of the corresponding average daily community load profiles.	[14]
Seasonality	Summer-Winter-Ratio	$P^i_{SWR} = \frac{\sum_{s\in SM}\sum_{t=1}^{T} E^i_{s,t}}{\sum_{w\in WM}\sum_{t=1}^{T} E^i_{w,t}}$	The ratio of the total demand in summer months SM to the total demand in winter months WM, applying the findings of Chapter 2.2.2. Higher values indicate a higher consumption in the summer than in the winter months. With AFPS, this parameter reduces to the ratio of total consumptions of the corresponding average daily community load profiles.	*Proposed*

Table 1: Collected electricity parameters

[4] The central statistical moments are calculated with the corrected, unbiased sample estimators; in the definitions, the uncorrected formulas are depicted for the sake of conciseness.

[5] The used standard PV load profile can be found in the thesis repository. As a proposal for further research, the Person correlation could also be calculated for arbitrary further standard load profile to measure the similarity to certain load profile shapes.

3.3. Parameter Selection

Using the three seasonal and two weekly differentiation criteria as presented in Chapter 3.2. results in six combinations for the parameter calculation. Calculating all 19 parameters (except the annual parameters *Summer-Winter-Ratio* and *Workday-Weekend-Ratio*, which are, due to their definition, only differentiated regarding the weekly respectively seasonal criteria) using these subsets, we receive 107 community parameters for each of the four parameter value subsets (*PFAS – Profit, PFAS – Gain, AFPS – Profit, AFPS – Gain*). This already represents a significant dimensionality reduction considering that each of the original load profiles contain 17,520 timesteps and thus features each, only for one of the five households in each community. Nevertheless, further dimensionality reduction through *feature selection* methods, aiming to select a smaller subset of features that minimizes redundancy and maximizes relevance to the target, might improve the learning performance of models, reduce the computational complexity and decrease the required storage, as stated in [29]. To explore this potential benefit, it should be compared if the performance of Machine Learning models trained on all parameters is different from those trained solely on a selected subset of parameters. Since the relationships between features and target might differ within the four parameter value subsets, the feature selection methods are applied to each subset individually. That entails that the previous parameter value subsets, differentiated by parameter calculation approach and classification task, are extended by a third dimension of differentiation criteria, the *parameter subset*, i.e. all features or selected features, leading to in total eight parameter value subsets that will be used to evaluate the potential of Machine Learning models in Chapter 4.1. The following three feature selection techniques will be applied conjointly to estimate the most informative parameters with regard to the target variables:

1. $\boldsymbol{\chi^2}$ **Feature Selection.** The χ^2 independency test statistic between each feature and the target is calculated, and the desired number of features with the best χ^2 scores is selected. This test statistic measures the dependence between variables; therefore, this technique will eliminate the features that are likely to be independent of the class [30] and therefore irrelevant for classification tasks.
2. **ANOVA F-Value Feature Selection.** The procedure is analogous to that of the $\boldsymbol{\chi^2}$ feature selection. This test statistic evaluates if, when we group the features by our target, the means for each group are significantly different [31].
3. **Random Forest Feature Selection.** Random Forests are Machine Learning algorithms or, more specifically, *Ensemble Learning* algorithms that utilize an ensemble of Decision Tree estimators; a further explanation of Random Forests can be found in [32] and in Chapter 4.1.1. During their execution, Random Forests estimate the most important features by using metrics from information theory, such as the Gini impurity or information gain.

By fitting a simple Random Forest Classifier[6] and inspecting the "by-product", the used importance metric, an estimate of the feature importance can be obtained.

Separately for all four parameter value subsets, each selection technique is applied to discover the 20 best parameters, and then assign them with a score of 1, while all other parameters are assigned with a score of 0. Subsequently, these scores are aggregated across the techniques, and the overall 20 best parameters are selected. The aggregated results can be found in the following Table 2.

	PFAS Profit	**PFAS Gain**	**AFPS Profit**	**AFPS Gain**
1	Daily Maximum Demand (Winter/Weekend)	Maximum Time-of-Use (Summer/Workday)	Daily Maximum Demand (Spring-Autumn/Workday)	Variance (Summer/Weekend)
2	Night impact (Spring-Autumn/Weekend)	End-of-Work impact (Summer/Weekend)	Daily Range Factor (Winter/Weekend)	Kurtosis (Summer/Weekend)
3	FFT Peak (Summer/Workday)	Lunch Impact (Spring-Autumn/Weekend)	Daily Range Factor (Winter/Workday)	Daily Minimum Demand (Summer/Weekend)
4	Maximum Time-of-Use (Winter/Weekend)	Minimum Time-of-Use (Spring-Autumn/Workday)	Daily Non-Uniformity Coefficient (Winter/Workday)	Daily Minimum Demand (Summer/Workday)
5	Maximum Time-of-Use (Winter/Workday)	Kurtosis (Summer/Workday)	Daily Range Factor (Spring-Autumn/Workday)	End-of-Work impact (Summer/Workday)
6	Maximum Time-of-Use (Spring-Autumn/Weekend)	Kurtosis (Summer/Weekend)	Daily Load Factor (Spring-Autumn/Workday)	Daily Non-Uniformity Coefficient (Summer/Weekend)
7	Variance (Summer/Workday)	FFT Peak (Summer/Weekend)	Variance (Winter/Workday)	PV Correlation (Spring-Autumn/Weekend)
8	Variance (Summer/Weekend)	Morning slope (Summer/Workday)	Morning slope (Summer/Weekend)	Skewness (Summer/Weekend)
9	Variance (Winter/Workday)	End-of-Work impact (Summer/Workday)	Daily Maximum Demand (Summer/Workday)	Lunch Impact (Spring-Autumn/Weekend)
10	Variance (Winter/Weekend)	Daily Maximum Demand (Spring-Autumn/Workday)	Daily Maximum Demand (Winter/Workday)	Kurtosis (Summer/Workday)
11	End-of-Work impact (Spring-Autumn/Weekend)	Night impact (Winter/Weekend)	Maximum Time-of-Use (Spring-Autumn/Workday)	Morning slope (Spring-Autumn/Weekend)
12	Minimum Time-of-Use (Summer/Weekend)	Morning slope (Spring-Autumn/Workday)	Morning slope (Winter/Weekend)	Skewness (Winter/Workday)
13	Morning slope (Winter/Weekend)	FFT Peak (Summer/Workday)	Variance (Winter/Weekend)	Summer-Winter-Ratio (Weekend)

[6] The used classifier was fitted with $n_{estimators} = 150$ and a fixed random seed to ensure reproducibility. Apart from that, the other hyperparameters were left at the current default settings of the scikit-learn library.

	PFAS Profit	PFAS Gain	AFPS Profit	AFPS Gain
14	Night impact (Winter/Weekend)	Minimum Time-of-Use (Spring-Autumn/Weekend)	PV Correlation (Winter/Weekend)	Skewness (Summer/Workday)
15	Minimum Time-of-Use (Winter/Weekend)	Minimum Time-of-Use (Winter/Weekend)	Minimum Time-of-Use (Spring-Autumn/Workday)	Variance (Summer/Workday)
16	FFT Peak (Winter/Workday)	Minimum Time-of-Use (Winter/Workday)	Daily Non-Uniformity Coefficient (Winter/Weekend)	Maximum Time-of-Use (Winter/Workday)
17	Daily Load Factor (Winter/Weekend)	Daily Minimum Demand (Summer/Workday)	Variance (Spring-Autumn/Weekend)	Minimum Time-of-Use (Spring-Autumn/Workday)
18	FFT Peak (Winter/Weekend)	Variance (Spring-Autumn/Workday)	Daily Load Factor (Winter/Weekend)	Minimum Time-of-Use (Spring-Autumn/Weekend)
19	Daily Maximum Demand (Winter/Workday)	Night slope (Winter/Weekend)	Minimum Time-of-Use (Winter/Workday)	Minimum Time-of-Use (Winter/Weekend)

Table 2: Parameters for Machine Learning models selected by three feature selection techniques, individually for each parameter value subset

Analyzing the result reveals that, according to the selection techniques presented above, the most important parameters, counted over all datasets and independently of seasonal and weekly differentiation criteria, seem to be the Minimum Time of Use (11), proposed in this work, Variance (11), FFT Peak (6), Maximum Time of Use (6), Daily Maximum Demand (6), Morning Slope (6), Kurtosis (5) and End-of-Work Impact (4).

4. Analysis of the Suitability for Energy Sharing Communities

4.1. Predictive Analysis: Supervised Machine Learning

4.1.1. Background and Used Techniques

Based on the motivation presented in Chapter 1.1., we will now investigate whether the electrical parameters engineered in Chapter 3.2. can serve as features for the construction of Machine Learning models that predict the target variables *Community Profit* and *Community Gain*, and therefore allow forecasting the suitability of households for energy sharing communities, based on their parameter values.

Supervised Learning is a Machine Learning task, in which the learning algorithm is, contrary to Unsupervised Learning, provided with *labeled* data, which is split in *two* sets: a training set and a test set. The main idea is that the algorithm "learns" by deriving rules from the example objects in the training set, which are provided with correct labels, so that it can assign the labels of objects in the test set, and ideally for arbitrary objects in practical applications, with the highest possible accuracy [33]. If one or more continuous values are to be assigned to each object, this is called a *regression* problem, if one of two or more categorical or discrete values, i.e. the belonging to a "class", is to be assigned, this is called a *classification* problem.

Since our targets, *Community Profit* and *Community Gain,* were discretized into three classes in Chapter 2.1., and we want to predict the membership of objects, the communities, to a certain class, this poses a *classification* problem.

There is a variety of different Machine Learning techniques for solving classification tasks, from relatively simple approaches, such as the instance-based k-Nearest-Neighbor algorithm or the Naïve Bayes algorithm, via advanced methods, for instance classification using Logistic Regression or tree-based methods, such as Decision Trees or Random Forests, through to highly complex approaches such as classification with Support Vector Machines or Deep Neural Networks.

In this chapter, we will evaluate if forecasting the target based on the compiled parameters is possible by modelling, training and evaluating representatives of three different classes of Machine Learning algorithms, whose backgrounds are briefly explained in the following.

1. **Logistic Regression** is, as stated in [34] and [35], a standard probabilistic statistical classification model that has been widely used across several disciplines. It employs the so-called logit or log-odds function. Different from Linear Regression, the outcome of Logistic Regression on one object is the probability that the object is positive or negative, or in the case of multiclass-classification, the probability that the object belongs to one of at least three classes [34]. The class probabilities in multi-class penalized logistic regression are estimated via stochastic models, the corresponding extensive mathematical details can be found in [35]. To improve the performance, we use the scikit-learn implementation of an estimator that incrementally trains and updates regularized **Logistic Regression models with Stochastic Gradient Descent** (SGD) learning [36]. Logistic Regression offers the advantage of a high level of model transparency; as described in [37], it is widely used in cases where the goal is to understand the role of features in explaining the outcome; therefore, concerning to this work, this method could help to explain the impact of certain parameters in case of satisfactory classification results.
2. **Tree-based Classification** techniques rely on Decision Trees, a non-parametric supervised learning technique. Decision Tree classifiers are based on creating a tree-like model, recursively partitioning the feature space at each tree node in such way that the objects with identical labels are grouped as optimal as possible [38], where the optimization criterion is to minimize the impurity of the partitioning, measured for example by the Gini impurity, cross-entropy or the misclassification error [37]. As an augmentation to reduce overfitting and improve generalizability as well as robustness, for instance explained in [39], *Ensemble Methods* combine multiple “weak learners” into a single “strong learner”: *Averaging Ensemble Methods*, such as Random Forests, focus on the idea to construct a diverse set of randomized tree classifiers, and then obtain the final prediction as the averaged prediction of the individual classifiers; *Boosting Ensemble Methods* are based on sequentially fitting new tree classifiers on repeatedly modified versions of the data. [39]

To achieve the best possible performance, we utilize the **Extreme Gradient Boosting** system, or shortly XGBoost, an advanced gradient boosting algorithm integrating regularization, which is currently widely used by data scientists to achieve state-of-the-art results on several machine learning competitions [40]. The technical details on general tree-based methods can be found in [37], details specifically on *XGBoost* in [40]. Tree-based techniques also yield a relatively good level of model transparency, since they all involve some measure of feature importance, for example the Gini impurity, which could be utilized to construct feature importance maps.

3. **Neural Networks** are nonlinear statistical models for classification or regression, typically represented by network diagrams. Their name derives from the fact that they were originally inspired by the principals of the human brain: They consist of units or nodes that represent neurons, and signal-transmitting, weighted connections between these neurons that represent synapses [37]. Due to intense research in the last years, a huge variety of advanced types of Neuronal Networks, such as Convolutional Neural Networks, Recurrent Neural Networks or Long Short-Term Memory Neuronal Networks, has evolved. An exemplary, comparably simple type of Neural Network, a **Multi-Layer Perceptron** (MLP), is composed of three types of components, as explained in [41]:
    1. One *input layer*; It consists of neurons representing the input features.
    2. An arbitrary number of *hidden layers*; They consist of an arbitrary number of neurons, and each neuron in the hidden layers transforms the values from the previous layer with a weighted linear combination, followed by a non-linear activation function, for instance the hyperbolic tangent or the sigmoid function.
    3. One *output layer*; It receives the values of the last hidden layer and transforms these values into the final output values.

    Depending on the Machine Learning problem, MLP use different *loss functions*, in the case of classification for instance the *cross-entropy*. Firstly, the MLP is initialized with random model parameters, called weights. Then, the MLP minimizes the value of the loss function by repeatedly updating the weights using *backward propagation*: A backward pass propagates the loss from the output layer in each training iteration to the previous layers, where each weight is updated in such way that the loss is reduced as greatly as possible [41]. Consequently, a Multi-Layer Perceptron can “learn” to correctly classify objects by estimating loss-reducing model weights iteratively from the training data. A drawback of Neural Networks is that, in contrast to the other methods presented, their complex model structure makes it challenging to understand the functioning of the trained model and to draw conclusions about the importance of individual features.

4.1.2. Model Configuration

For the following Machine Learning implementations, the scikit-learn library will be utilized [5]; if not further specified, the scikit-learn default hyperparameters are used. All models were trained with a fixed random number generation seed to ensure the reproducibility of results. The available data, namely the community parameter values, differentiated in the eight subsets defined in Chapter 3.3., with the corresponding targets, the Community Profit respectively Community Gain, is split up into 70% training data and 30% holdout test data for the final model evaluation. In order to find suitable configurations for the implementation of the Machine Learning techniques proposed in the last chapter, the hyperparameter optimization technique *Grid Search* is applied. Grid Search is an extensive search over a pre-defined set of hyperparameter values [42]: the given model is initialized and trained separately for all possible combinations of the given hyperparameters in order to determine the best-performing set of hyperparameters, as measured by a pre-defined performance metric. Since our available data of $n = 1000$ energy sharing communities is relatively scarce and therefore requires economic usage of the data regarding the selection of training, validation and test set, the Grid Search will be further complemented by using K-Fold Cross Validation (CV), as described for instance in [37]: first, the training data is split in K equally sized partitions. For $k = 1, \dots, K$, the kth partition will be used as a validation set to evaluate a model fitted on the remaining $K - 1$ partitions. Lastly, the K evaluations will be averaged. For the following analysis we choose $K = 3$, since this seems to provide a good trade-off between a sufficiently extensive evaluation and runtime efficiency. After completing the 3-fold validated Grid Search, the models will be trained on the respectively best found combination of hyperparameters, and finally evaluated using the previously unseen holdout data set.

The following configurations are searched:

1. For the **Logistic Regression with SGD Learning** Model, we choose to search the hyperparameter grid depicted in Table 3, containing among others different regularization methods and strengths. This hyperparameter grid results in the evaluation of 4,860 different hyperparameter configurations:

Hyperparameter	Examined Values
Regularization	L1, L2, Elastic Net
Regularization strength term α	0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 20, 100
Regularization mixing term $L1_{ratio}$[7]	0.15, 0.3, 0.5
SGD learning rate schedule η	Constant, Inverse Scaling, Adaptive, Optimal (Heuristic approach, proposed in [43])
Initial SGD learning rate η_0 (Not used with *Optimal* learning rate schedule)	0.0001, 0.001, 0.01, 0.1, 1
Tolerance[8] tol	0.00001, 0.0001, 0.001

Table 3: Hyperparameter grid searched for Logistic Regression with SGD learning

[7] $L1_{ratio} = 1$ results in pure L1 regularization, $L1_{ratio} = 0$ results in pure L2 regularization.
[8] The convergence threshold. If the loss function does not improve by at least this value for 10 iterations, the training stops.

2. **Extreme Gradient Boosting** offers a broad variety of hyperparameters to be tuned (for details on the hyperparameters, see [40]), wherefore varying all parameters would require too much computational effort. All models were trained using the softmax function for multiclass classification as objective function. The hyperparameter space displayed in Table 4 was searched, examining in total 9,360 different hyperparameter configurations:

Hyperparameter	Examined Values
L1 regularization strength term α	0, 0.001, 0.1, 1, 10, 100
L2 regularization strength term λ	0, 0.001, 0.1, 1, 10, 100
Boosting Learning Rate η_0	0, 0.01, 0.1, 0.5
Maximum tree depth td_{max}	3, 5, 7, 9
Number of boosting rounds $n_{estimators}$	10, 20, 50, 100, 200, 400, 600, 800, 1000, 2000

Table 4: Hyperparameter grid searched for Extreme Gradient Boosting

3. For **Multi-Layer Perceptrons**, even more hyperparameters are required, since the user has to specify, inter alia, not only learning rates and regularization parameters, but also the complete network topology. For a reasonable set of hyperparameters, some common values as depicted in Table 5 were searched, resulting in 3,000 different hyperparameter configurations to be investigated. All models were trained with the *rectified linear unit* as activation function, and a fixed maximum number of optimization iterations of 10^8 to enable convergence also for more complex network configurations. Contrary to the latter methods, MLP are *not* invariant to scale effects, wherefore, as proposed in [37], the parameters were z-score standardized to have mean 0 and variance 0 beforehand.

Hyperparameter	Examined Values
Number of neurons per hidden layer $n_{neurons}$	10, 30, 50, 70, 100
Number of hidden layers n_{hidden_layers}	1, 2, 3, 4, 5, 10, 30, 50, 70, 100
Weight optimization solver S	Adam (Stochastic gradient-based method, for details see [44]), L-FBGS (Quasi-Newton method, for details see [45])
L2 regularization strength term α	0.0001, 0.001, 0.01, 0.1, 1
Tolerance[9] tol	0.000001, 0.00001, 0.0001
Exponential decay rate for estimates of the first movement vector β_1 (Only used with *Adam* optimizer)	0.8, 0.9

Table 5: Hyperparameter grid searched for Multi-Layer Perceptron

[9] The convergence threshold. If the loss function does not improve by at least this value for 10 iterations, the training stops.

In order to analyze the possibility of using Machine Learning models to predict the suitability for energy sharing communities as wide-ranging as possible, the Grid Searches are conducted separately for all combinations of the differentiation criteria developed in the last chapters, as mentioned beforehand leading to in total eight different parameter value subsets, shortly recapitulated in the following:

- Both *classification tasks*, i.e. Community Profit vs. Community Gain, as explained in Chapter 2.1.
- Both *parameter calculation approaches*, i.e. PFAS vs. AFPS, as explained in Chapter 3.1.
- Both *parameter subsets*, i.e. the full set of parameters vs. solely the parameters selected in Chapter 3.3.

Performing the Grid Search, i.e. fitting and evaluating models with all possible hyperparameter combinations, for all three Machine Learning techniques on all of the eight parameter value subsets presented above, results in the total training and evaluation of 137,760 models, each of which is 3-fold cross-validated.

4.1.3. Model Evaluation

The multiclass accuracy classification score was selected as a metric for the evaluation, because it yields a straightforward interpretation, as it measures the overall percentage of correct predictions. As stated in [46], the accuracy score can be biased in case of imbalanced class distributions; since we work with uniformly distributed and therefore ideally balanced classes due to tercile-based discretization, this potential drawback does not pose a problem here. For C classes, the multiclass accuracy score is defined as:

$$accuracy = \sum_{c=1}^{C} \frac{TP_c + TN_c}{TP_c + TN_c + FP_c + FN_c}$$

where TP_c, TN_c, FP_c, FN_c are the true positives, true negatives, false positives, false negatives of class c [46].

Table 6 reports the results of the training and evaluation of the Machine Learning model configurations discussed in the last Chapter 4.1.2. for the classification of the Community Profit, Table 7 that of the classification of the Community Gain. They present the best found hyperparameter combinations[10] and the corresponding accuracy scores for all three Machine Learning techniques and all of the eight parameter value subsets, that were constructed as further extensions of the beforehand four parameter value subsets in Chapter 3.3.

[10] In case that multiple values for an individual hyperparameter led to the best found accuracy score during cross-validation, one value is selected and marked with [*] for the sake of clarity. The in-depth classification results, including all of the mentioned hyperparameter combinations, can be found in the thesis repository.

Classification of Community Profit Based on All Load Profile Parameters									
		PFAS Parameter Calculation Approach				AFPS Parameter Calculation Approach			
		All Features		Selected Features		All Features		Selected Features	
Model	Hyper-parameter	Values	Accuracy Score	Values	Accuracy Score	Values	Accuracy Score	Values	Accuracy Score
Logistic Regression with SGD Learning	Regularization Method	El. Net	0.38	L2	0.35	El. Net	0.35	El. Net	0.37
	Regularization strength term	1		10		0.1		0.1	
	Regularization mixing term	0.5		-		0.5		0.3	
	SGD Learning rate schedule	Adapt.		Adapt.		Adapt.		Adapt.	
	Initital learning rate	0.01		0.0001		1		0.01	
	Tolerance	0.00001		0.001		0.0001*		0.001	
Extreme Gradient Boosting	L1 strength term	0.1	0.34	0.001	0.36	0	0.37	0.1	0.31
	L2 strength term	100		1		0*		10	
	Boosting learning rate	0.01		0.1		0.5		0.01	
	Max. tree depth	3*		3*		3*		3*	
	Number of boosting roounds	100		13		10		50	
Multi-Layer Perceptron	Number of neurons	30	0.37	50	0.35	30	0.35	10	0.38
	Number of hidden layers	3		5		10		2	
	Weight optimization solver	LBFGS		LBFGS		LBFGS		LBFGS	
	L2 strength term	0.0001		0.1		0.01		0.01	
	Tolerance	0.000001		0.0001		0.0001		0.0001*	
	Exponential decay rate	-		-		-		-	

Table 6: Results for the classification of community profit

Classification of Community Gain Based on Consumer Load Profile Parameters									
		PFAS Parameter Calculation Approach				AFPS Parameter Calculation Approach			
		All Features		Selected Features		All Features		Selected Features	
Model	**Hyper-parameter**	**Values**	**Accuracy Score**	**Values**	**Accuracy Score**	**Values**	**Accuracy Score**	**Values**	**Accuracy Score**
Logistic Regression with SGD Learning	**Regularization Method**	L1	0.36	El. Net	0.36	L2	0.33	L2	0.34
	Regularization strength term	0.01		0.0001		20		0.1	
	Regularization mixing term	-		0.15		-		-	
	SGD Learning rate schedule	Adapt.		Adapt.		Opt.		Adapt.	
	Initital learning rate	0.01		0.1		1*		0.1	
	Tolerance	0.001		0.001*		0.0001*		0.00001	
Extreme Gradient Boosting	**L1 strength term**	0.1*	0.28	1	0.24	10	0.30	1	0.34
	L2 strength term	10		0		0.1		100	
	Boosting learning rate	0.01		0.01		0.5		0.5	
	Max. tree depth	3*		3*		3*		3*	
	Number of boosting roounds	13*		11		10		20	
Multi-Layer Perceptron	**Number of neurons**	70	0.31	30	0.33	100	0.27	10	0.30
	Number of hidden layers	4		5		10		5	
	Weight optimization solver	LBFGS		LBFGS		LBFGS		LBFGS	
	L2 strength term	1		0.01		0.001		0.01	
	Tolerance	0.0001*		0.00001		0.00001		0.0001*	
	Exponential decay rate	-		-		-		-	

Table 7: Results for the classification of community gain

Analyzing the classification results reveals that none of the investigated combinations of different Machine Learning techniques, classification tasks, parameter calculation approaches, and feature subsets yields satisfactory accuracy: Considering that the holdout test set is, analogous to the training set, uniformly distributed regarding the target classes due to stratified sampling, the expected value of randomly guessing the class would be $E = \frac{1}{3}$, and a trivial model constantly predicting one fixed arbitrary class would achieve the score $accuracy = \frac{1}{3}$ as well.

Therefore, as the best found accuracy score is 0.38, it can be concluded that, despite extensive hyperparameter optimization and the use of eight different subsets, none of the investigated model configurations was capable to predict the Community Profit respectively Community Gain class with more than 5% higher accuracy compared to randomly guessing or a trivial model.

Despite the unsatisfactory results, the table of results features some interesting observations: Although it represents the simplest of the examined techniques in terms of complexity, the Logistic Regression model with SGD learning outperforms on average the modern Extreme Gradient Boosting models as well as the complex Multi-Layer Perceptron models on the majority of all investigated classification tasks, parameter calculation methods and parameter subsets, while requiring a significantly lower computational effort at the same time.

We suspect that one possible explanation for the failure of the inspected Machine Learning models in predicting the suitability for energy sharing communities based on the proposed electricity parameters might be lack of information: it might be the case that the models were not able to explain the results of the complex simulation process designed in [1] based on the communities' load profiles alone, but need to be enriched with additional information that could have played a role during the simulation process, such as for instance the load profiles of the PV plants and battery storage systems, properties of the load flows between the pro- and consumers, or implications of the assumed pricing and operation strategies, in order to be able to predict the suitability precisely enough; extending the used Machine Learning models with suchlike additional information could be a topic of further research. This hypothesis possibly might be confirmed by the results of recent research presented in [23], stating that the load profile has a rather limited impact on the net present value of PV-coupled batteries. Another hypothesis would be that the part of the load profiles' characteristics expressed by the chosen collection of parameters is not correlated highly enough with the target, which would indicate future research using different or additional electricity parameters. In addition, a further hypothesis could be that the averaging and aggregation methods cause a too severe information loss: for instance, averaging the load profiles respectively parameter of a whole season into one single value could cause small, but for the community performance crucial observations on day or intra-day-level to vanish, and aggregating several households' load profiles respectively electricity parameters into a single one could fade out the complex interdependencies between the households within a community; this hypothesis would imply further research on extending the proposed or developing new approaches to aggregate the parameters on community level.

A further, but due to the extensive hyperparameter optimization and the use of different classes of Machine Learning algorithms, less likely explanation could be that the selected models and their configurations were not suitable for the underlying data, which would suggest investigating more sophisticated or differently configured Machine Learning algorithms on the available data; a possibility for that could be Support Vector Machines or the use of more complex Neural Networks with more sophisticated activation functions.

4.2. Descriptive Analysis: Data Exploration

As the Machine Learning Models investigated in the last chapter were not able to predict the profit and gain of the considered energy sharing communities with sufficient precision, in this chapter, the interdependency between certain characteristics of the communities' load profiles, respectively their parameter values, and their suitability for energy sharing communities will be examined with different qualitative explorative data analysis approaches.

4.2.1. Analysis of Daily Average Load Profiles

For the following visual analysis, the representative average daily load profile is calculated for each household in a first step, differentiated according to the seasonal and weekly criteria used previously to increase the accuracy and minimize information loss caused by averaging. Hereby, based on the same considerations as for the AFPS approach presented in chapter 3.1., the median is utilized. Subsequently, for each community the representative average load profiles of the included prosumers and consumers were plotted side by side in the following Figures 9-11 in order to compare the characteristics of the load profiles of the individual prosumers and consumers within the communities. This community load profile analysis on household-level was conducted for the highest performing (upper tercile) and lowest performing (lower tercile) households by Community Profit, in order to identify properties that characterize load profile combinations that yield particularly low and particularly high profit operating as a community. For each season, the corresponding load profiles for working and weekend days are analyzed conjointly.

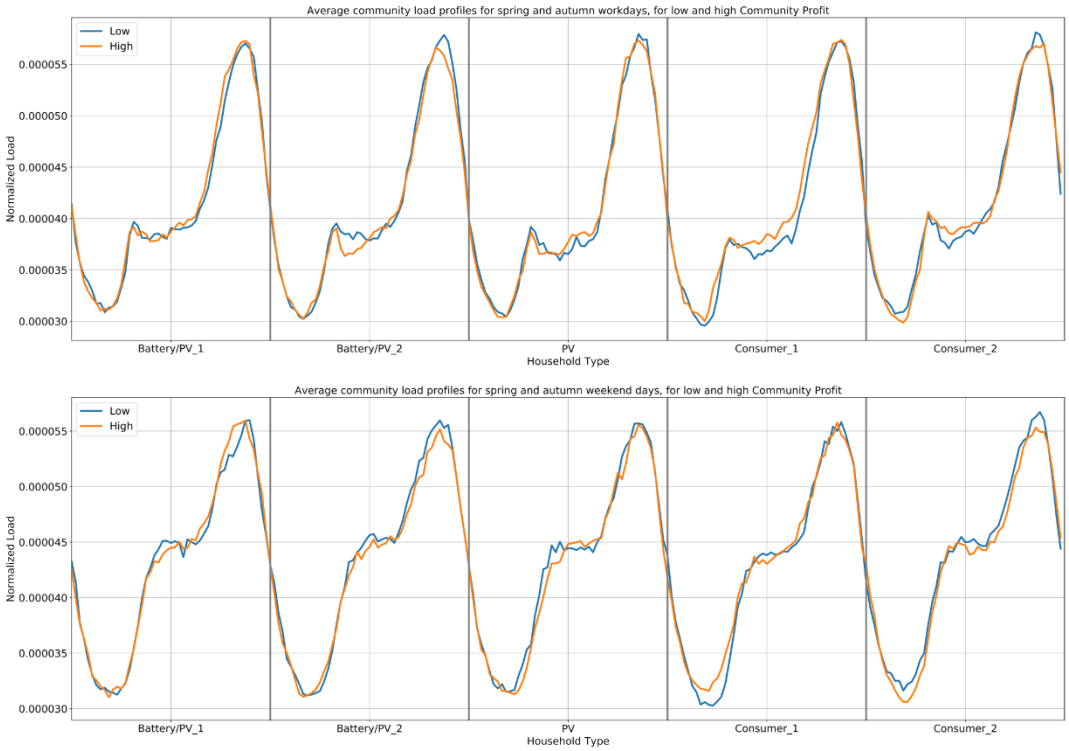

Figure 9: Visualization of the representative daily average load profiles on community-level for spring and autumn work- and weekend-days, differentiated by low- (blue) and high-performing (orange) communities

Figure 9 compares the average community load profiles for spring and autumn workdays and weekend days. The load profiles look nearly identically for the first Battery/PV household. The high-performing workday load profiles of the 2nd Battery/PV household and the PV household both display a steeper drop after the morning peak (around 7 a.m.) than the corresponding low-performing load profiles, while for the consumer households the exact opposite applies. One possible hypothesis to explain this might be that the load which was “saved” in this temporal area by the two prosumer households leads to excess PV generation, which in turn can be internally used to cover the higher demand of the consumers, leading to a more beneficial energy distribution and therefore Community Profit. On the weekend days, these findings seem not to apply: none of the households features a clear drop after the morning peak, which could be explained by fewer inhabitants leaving for work on the weekends.

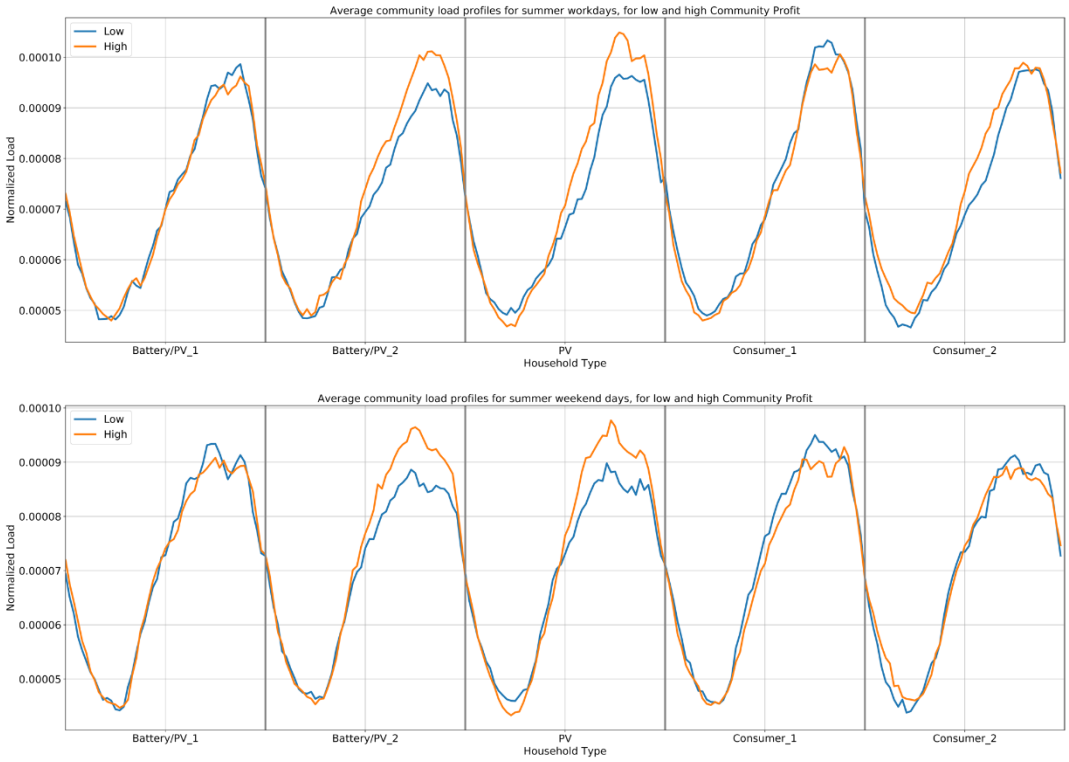

Figure 10: Visualization of the representative daily average load profiles on community-level for summer work- and weekend-days, differentiated by low- (blue) and high-performing (orange) communities

Figure 10 compares the average community load profiles for summer workdays and weekend days. Here can be observed that for both, workdays and weekend days, the high-performing average load profiles of the 2nd Battery/PV household and the PV household feature a drastically higher consumption level throughout the whole second half of the day than the low-performing load profiles, in particular the evening peaks are clearly more pronounced; in addition, on workdays the high-performing average load profile of the 2nd consumer household exhibits a slightly higher load consumption level than the low-performing, especially around the midday hours, when the highest PV generation through the consumers is to be expected. Regarding the consumer households, an overall slightly lower consumption level is observed for the high-performing communities than the low-performing, but due to the strong dominance in the consumption level of the prosumers, the overall consumption in the summer months of the high-performing communities' households appears to be higher considered as a whole. This is confirmed on parameter-level by the fact that the values of the Summer-Winter-Ratio and Daily Maximum Demand (Summer) parameters are, calculated with AFPS and grouped by the target Community Profit, on average on workdays and weekend days higher for the high-performing communities than for the low-performing[11]. A hypothetical reasoning for this relationship could be the following:

[11] Summer-Winter-Ratio Workdays/Weekend low (high) profit: 1.40/1.30, (1.44/1.33),
Daily Max. Demand Summer Workdays/Weekend low (high) profit: 0.00049/0.00046 (0.00051/0.00047)

in the summer months, a substantially higher solar radiation and thus also a significantly higher output of PV systems is to be assumed[12]. Consequently, it can be assumed that prosumer households that have a higher load demand in the summer months will also cover, in absolute terms, a higher amount of demand, either amongst each other or for the consumer households, from self-generated PV energy, resulting in higher energy flows between the community members and thereby a higher profit.

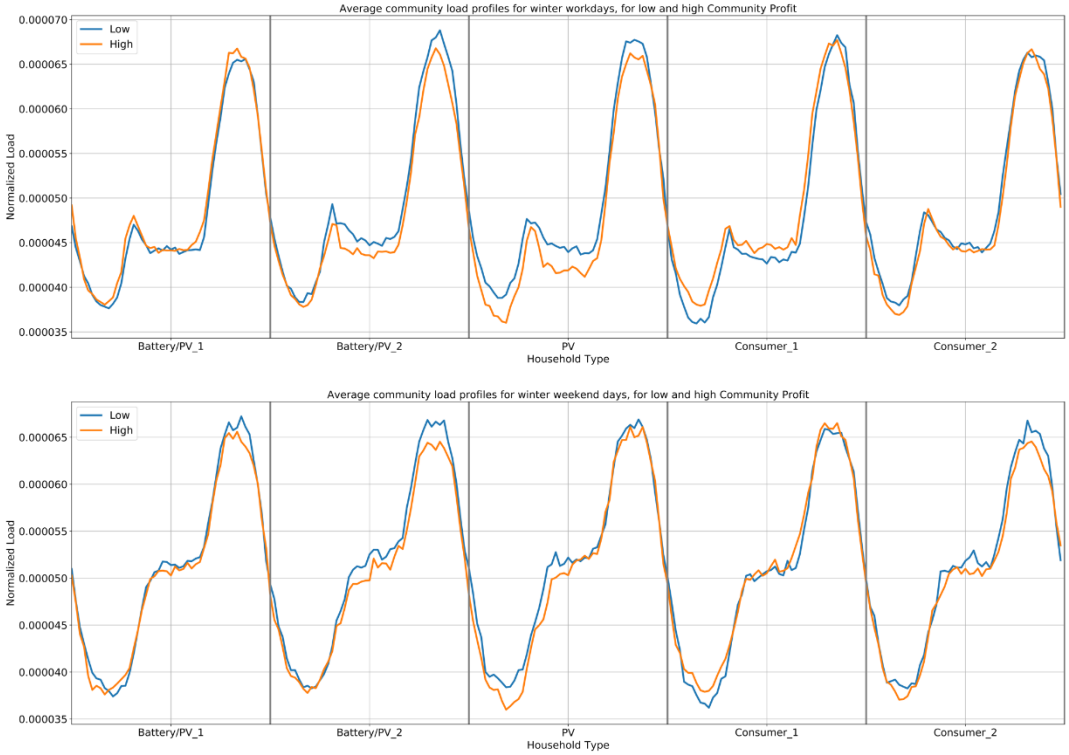

Figure 11: Visualization of the representative daily average load profiles on community-level for winter work- and weekend-days, differentiated by low- (blue) and high-performing (orange) communities

Figure 11 compares the average community load profiles for winter workdays and weekend days. With exception of the 1st Battery/PV household on workdays and the 1st consumer household on work- and weekend days, the load profiles of the low-performing households feature a slightly higher consumption level than the high-performing in general, including higher and steeper rising evening peaks. A possible hypothesis to explain this may be, for instance, that the difference is reasoned in the fact that the solar irradiation and therefore the PV generation is lower in the winter months, which could cause that the prosumers with a *lower* consumption level are despite the lower PV generation still able to cover a higher share of their demand, respectively that of the consumers, with self-generated PV energy, while the prosumers with a higher consumption level in the winter months can only cover a relatively smaller percentage of their demand with self-generated PV energy, forcing them to acquire more energy from the grid in addition, leading to higher overall cost for the energy sharing community.

[12] As mentioned in chapter 3.1., statements like these should always be considered in the context of the climatological conditions of the region under investigation and should therefore be modified as necessary for further research.

Analogous to the analysis of the summer months above, this relationship is confirmed by the fact that the median values of the Summer-Winter-Ratio is on average higher for the high-performing communities than for the low-performing, and that the opposite applies for the median values of the Daily Maximum Demand (Winter) parameter.[13]

The above analysis shows notable differences between the average pro- and consumer load profiles of low- and high-performing communities. In summary, we propose the hypotheses that regarding parameters, a high Summer-Winter-Ratio, or, regarding load profile shapes, a summer load profile featuring high consumption levels starting from midday and pronounced evening peaks, and a winter load profile featuring a deep morning valley and a relatively low consumption level around midday, might be a favorable characteristic for prosumers regarding a successful energy sharing community, while simultaneously a lower Summer-Winter-Ratio seems to be favorable for consumers. As this hypothesis was formulated solely on the available sample of 1,000 simulated communities and therefore requires further revision, a topic of further research could be to test if the posed hypothesis generalizes on other datasets or not, or to further extend and refine it.

4.2.2. Analysis of Parameter Value Distributions

In this section, the empirical distributions of the parameter values of high- and low-performing communities are analyzed in order to investigate whether conclusions can be drawn about the relationship between certain characteristics of community parameter values and the Community Gain. As mentioned in Chapter 3.3., the differentiation by seasonal and weekly criteria leads to 107 parameters, which would make a manual analysis of the distributions of all these parameters highly complex. Therefore, in order to reduce the parameters to be analyzed to a feasible number, we propose the following heuristic selection approach:

For each parameter, we will select the parameter's community value distributions for high- and low-performing communities. Subsequently, the *Two-Sample Kolmogorov-Smirnov Test* (in the following *KS Test*) will be applied to the two value distributions of each parameter in order to discover which parameters show significant differences regarding the distributions of their values of low- and high-performing communities, and thus harbor promising potential to draw conclusions about the profit from characteristics of the community parameters. The Two-Sample KS Test is, as described in [47], a non-parametric statistical test used to determine if two one-dimensional populations have the same probability distributions or not; for two populations $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$ with unknown continuous cumulative distributions functions F_Y and G_Y, the test hypotheses are:

[13] Daily Max. Demand Winter Workdays/Weekend low (high) profit: 0.00034/0.00033 (0.00033/0.00032)

H_0: $F_X(x) = G_Y(x)$, X and Y are from the same probability distribution

H_1: $F_X(x) \neq G_Y(x)$, X and Y are not from the same probability distribution

The corresponding test statistic is, as described for instance in [47], the maximum absolute difference between the two corresponding empirical cumulative distribution functions $\hat{F}_m$ and $\hat{G}_n$:

$$\sqrt{\frac{mn}{m+n}} D_{mn} = \sqrt{\frac{mn}{m+n}} \sup_{x \in \mathbb{R}} \left| \hat{F}_m(x) - \hat{G}_n(x) \right|.$$

Since we do not aim to test for statistically significant test results but only to find a heuristic reference which parameters to subsequently analyze manually, we choose the significance level $\alpha = 0.15$, i.e. H_0 is rejected and two different distributions are suspected, if the KS test statistic indicates that the probability of the observed data being generated, given that H_0 (X and Y are from the same probability distribution) is true, i.e. the probability of a type I error, is located below 15%.

As the data subset for the analysis we choose the parameter values calculated with the AFPS approach and for the Community Gain classification task (AFPS – Gain). We suspect that this configuration might offer the best means of interpretability, since the AFPS approach parameters are based on one aggregated community load profile, which, due to the classification task, only requires the aggregation of the two consumer load profiles and therefore enables more fine-grained interpretations. The following Table 8 shows a selection of 20 parameters in the AFPS – Gain parameter value subset, for which H_0 was rejected at significance level $\alpha = 0.15$:

Parameter (AFPS-Gain)	KS Test p-Value
Skewness (Winter/Workday)	0,019236432
Kurtosis (Winter/Workday)	0,020588725
FFT Peak (Summer/Workday)	0,034057449
Minimum ToU (Winter/Weekend)	0,038826629
Night Slope (Winter/Weekend)	0,04236679
PV Correlation (Spring-Autumn/Workday)	0,055403387
Daily Non-Uniformity Coefficient (Winter/Weekend)	0,057500592
Summer-Winter-Ratio (Weekend)	0,0585978
Lunch Impact (Winter/Weekend)	0,060562683
Daily Minimum Demand (Summer/Workday)	0,065442305
Summer-Winter-Ratio (Workday)	0,088210028
Minimum ToU (Spring-Autumn/Workday)	0,093791142
FFT Peak (Summer/Weekend)	0,09463478
Night Impact (Winter/Workday)	0,100852555
FFT Peak (Spring-Autumn/Workday)	0,11496705
Daily Maximum Demand (Summer/Workday)	0,119894644
Daily Load Factor (Winter/Workday)	0,121061179
Kurtosis (Spring-Autumn/Workday)	0,140614317
Lunch Impact (Spring-Autumn/Weekend)	0,14110743
Night Slope (Spring-Autumn/Workday)	0,142523513

Table 8: Parameters with significant differences in their distributions for low- and high performing features, selected with Kolmogorov-Smirnoff tests

Regarding the visual analysis of the parameter value distributions, box-and-whisker plots and kernel density estimation (KDE) plots are used conjointly for each parameter. Box-and-whisker plots combine the most important non-parametric statistics, namely minimum, lower quartile, median, upper quartile, maximum, range and the interquartile range, a robust estimator of dispersion, in an informative, concise visualization; KDE plots might give us an indication of the actual distribution of the parameter values, estimating their probability density function by fitting a continuous, non-negative *kernel function*, such as the Gaussian Kernel, to the observed values. Complementary to the box-whisker plots, KDE plots also offer the capability to detect and analyze the properties of potential bi- or multimodal distributions. Details on KDE can be found, for instance, in [48]. Although this removes the actual maximum and minimum values from the box plots, extreme non-representative outliers (smaller than the 5^{th} percentile or larger than the 95^{th} percentile) were excluded for the sake of facilitating the visual analysis. The value distributions of the parameters selected above are displayed in the Figures 12-31 and analyzed in sequence in the following.

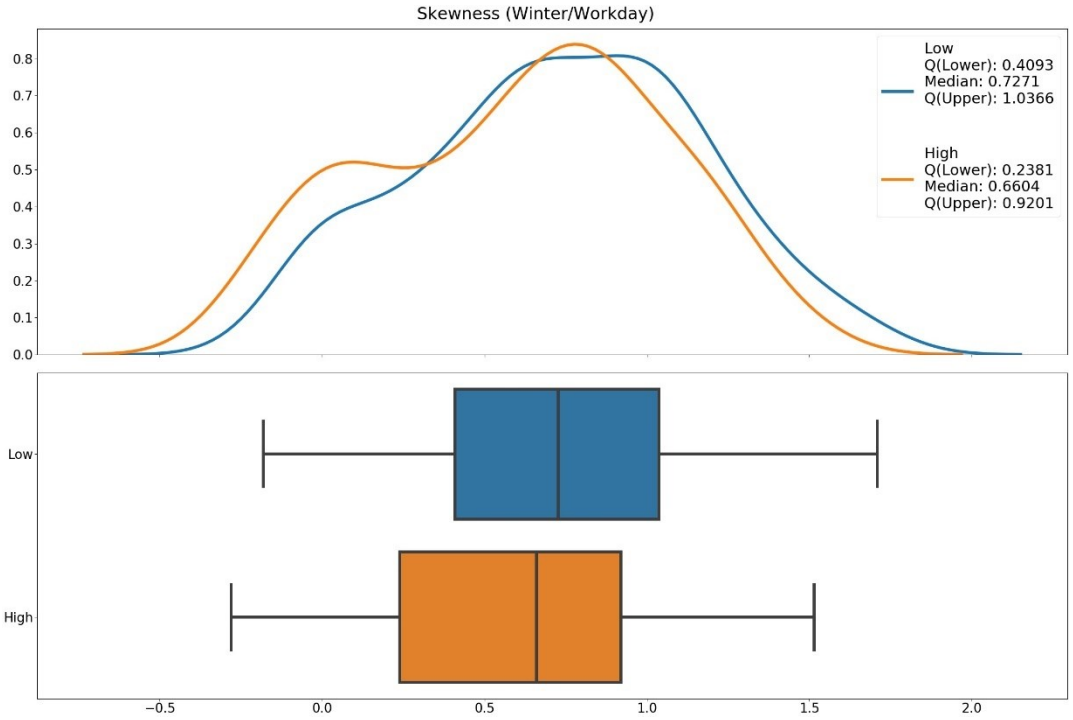

Figure 12: KDE plot and box-whisker plot of the value distributions of the parameter Skewness (Winter/Workday) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

Regarding the Skewness (Winter/Workday), we can observe in Figure 12 that the majority of both distribution's parameter values is positive. From the smaller median value and the further left located interquartile range, it appears that high-performing community load profiles tend to feature a lower (but still positive) Skewness on winter weekend days than the low-performing, i.e. that the distributions of their daily load values[14] tend to be *less* right-skewed, which in turn means, figuratively speaking, that very low load values occur less often, while moderate load values are more frequent. A possible explanation for that might be, that consumers with a higher frequency of moderate load demand values could be able to better cover these moderate load values from PV energy generated by the prosumers, while a higher relative frequency of low load demand leads to less opportunities where a higher amount of PV energy generated by the prosumers can be used by the prosumers, leading to a less favorable load distribution across the community.

[14] Annotation: Please note that all central statistical moment parameters, such as variance, skewness and kurtosis are not to be interpreted for the load profile itself, which represents a time series and therefore underlies a fixed temporal structure, but for the *distribution of the different load values* of a load profile. For instance, a positive skewness value does *not* mean that the shape of the load profile is right-skewed, i.e. that its consumption peak is located in the morning hours, but that the *distribution* of its values is right-skewed, i.e. that it features values that are smaller than the arithmetic mean more frequently.

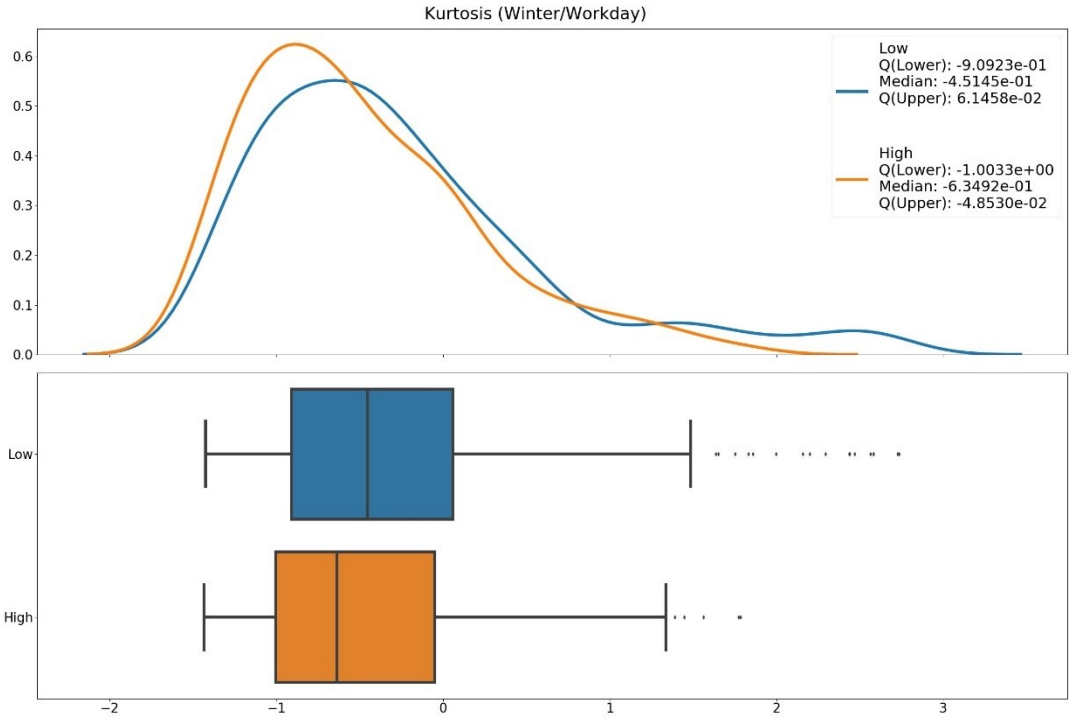

Figure 13: KDE plot and box-whisker plot of the value distributions of the parameter Kurtosis (Winter/Workday) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

For the parameter Kurtosis (Winter/Workday), whose distributions are depicted in Figure 13, almost only negative values can be observed; while at least 75% of the high-performing load profiles feature solely negative values, the low-performing display more positive values and, despite the 5% outlier removal, still some very high positive outliers. Similar proportions as for the latter parameter can be observed: the interquartile range of the high-performing community load profiles is located more to the left than that of the low-performing, and the median is also located clearly lower. Figuratively interpreted, the value distributions of a load profile with a positive Kurtosis are rather steep-peaked, i.e. the different strengths of load values are less evenly distributed.

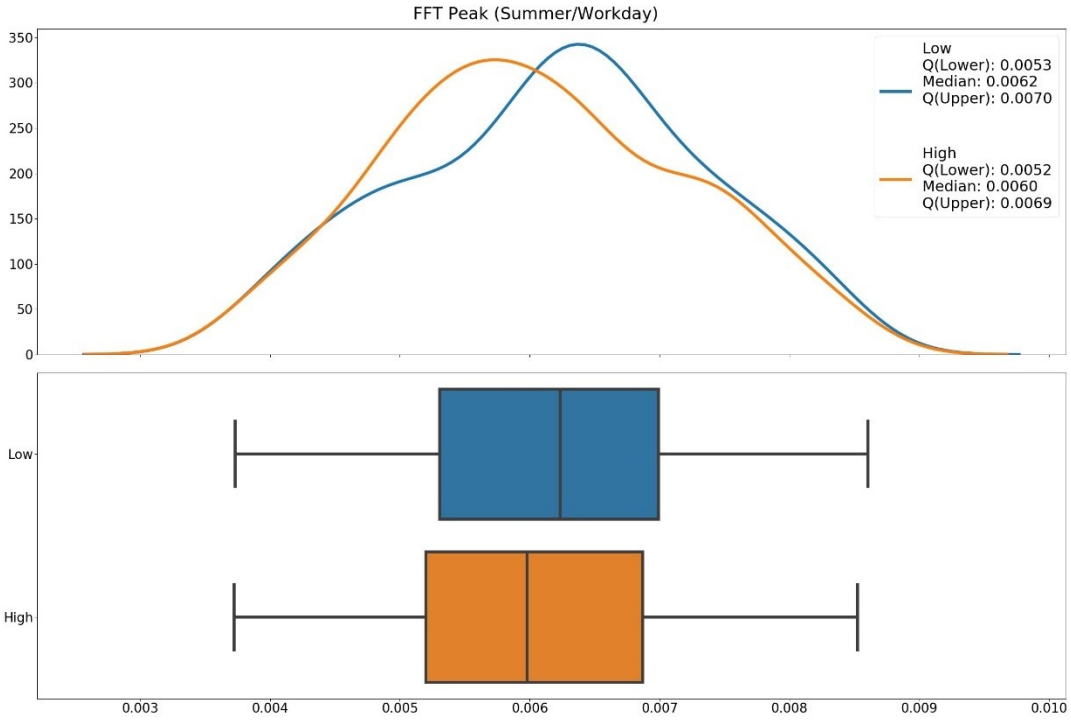

Figure 14: KDE plot and box-whisker plot of the value distributions of the parameter FFT Peak (Summer/Workday) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

Regarding the parameter FFT Peak (Summer/Workday), i.e. the maximum of the absolute values of the Fast Fourier Transform of the aggregated community load profile on summer workdays, it can be observed in Figure 14 that although the interquartile ranges are fairly similar, the median value of FFT Peaks is clearly lower for high-performing community load profiles than for low-performing, and the distribution seems slightly right-skewed compared to the rather left-skewed distribution of the low-performing.

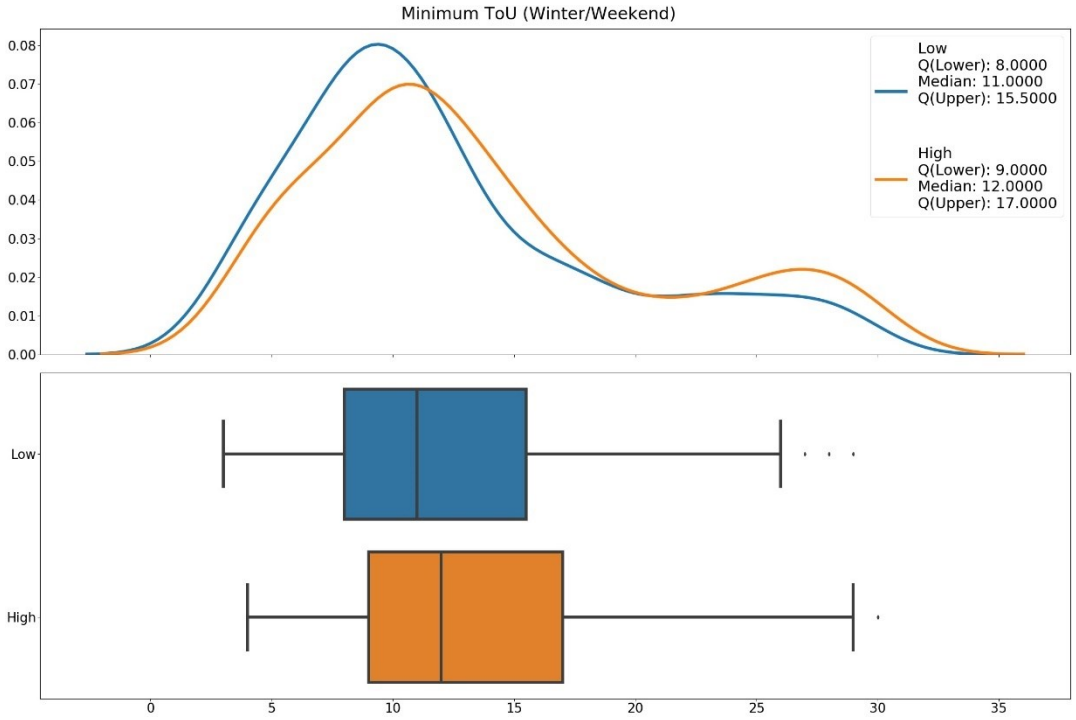

Figure 15: KDE plot and box-whisker plot of the value distributions of the parameter Minimum Time-of-Use (Winter/Weekend) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

Observing the KDE plot of the parameter Minimum Time-of-Use (Winter/Weekend) in Figure 15, it strikes that the distribution of the high-performing load profiles appears to be bimodal. The interquartile range for the daytime of the minimum load consumption of the low-performing community load profiles extends from 4:00 a.m. to 7:45 a.m., while that of the high-performing moves in the range from 4:30 a.m. to 8:30 a.m. In addition, several high-performing load profiles seem to have their minimum load consumption from approx. 1 p.m. to 2 p.m., i.e. in the post-lunch hours, indicated by the second, smaller peak around that time range and the long-reaching right whisker. Generally speaking, the hypotheses arises that a *later* minimum time-of-use on winter weekend days seem to be a favorable characteristic of consumer load profiles for more profitable energy sharing communities. This might be, for instance, reasoned in the fact that after the minimum consumption in the morning hours, normally the morning rise in consumption begins; shifting this to a slightly later time of day, when the sun has already risen even in the winter months, might enable the consumer households to be able to already use some of the PV energy generated by the prosumers instead of energy from the grid, resulting in better profits compared to operating individually.

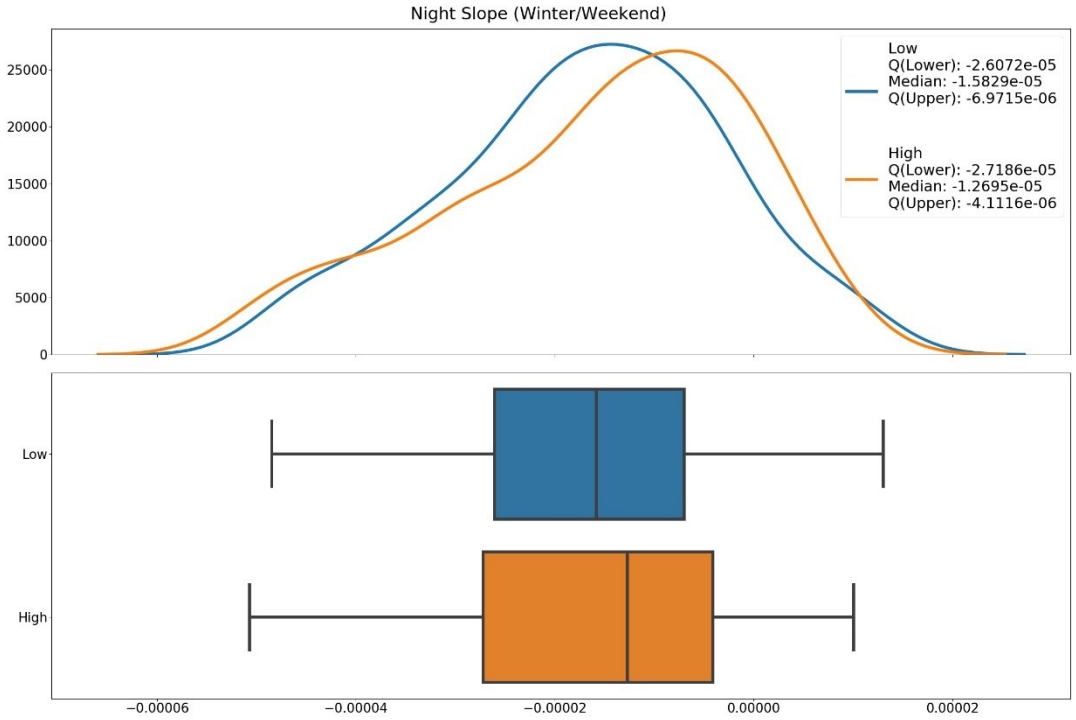

Figure 16: KDE plot and box-whisker plot of the value distributions of the parameter Night Slope (Winter/Workday) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

Regarding the parameter Night Slope (Winter/Weekend) it can be observed in Figure 16 that the majority of values are negative. As the Night Slope measures the difference between the load values at 11 p.m. and 9 p.m., a negative value means that the consumption level drops as the evening progresses, a more negative value implies a steeper drop and vice versa. The high-performing consumer load profiles show a significantly less negative median Night Slope on winter weekend days than the low-performing, which is complemented by the corresponding KDE plot displaying a clearly less left-skewed distribution of the high-performing load profiles' values. This suggests the hypothesis that a less steep drop in the consumer households' consumption levels from 9 p.m. to 11 p.m. on winter weekend days could be a characteristic that leads to a profit-increasing constellation; Figure 11 in chapter 4.2.1., especially the 2nd consumer's average representative load profile further confirms this visually on the absolute Community Profit level.

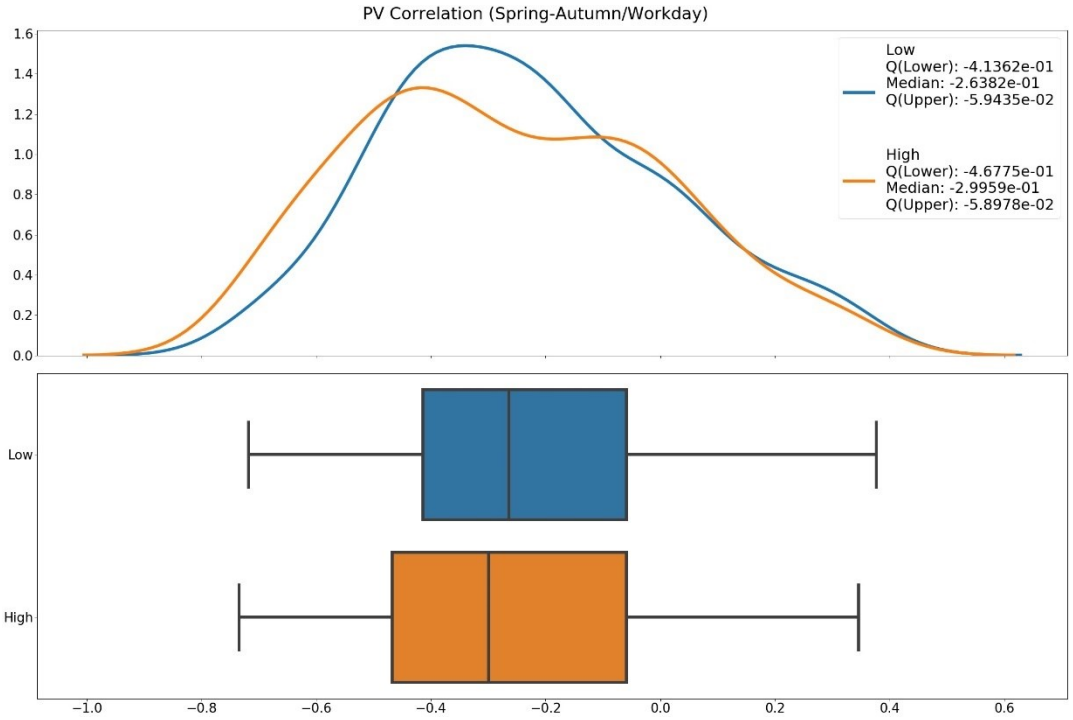

Figure 17: KDE plot and box-whisker plot of the value distributions of the parameter PV Correlation (Spring-Autumn/Workday) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

Observing the distributions of the PV Correlation (Spring-Autumns/Workday) in Figure 17 it can be stated that the high-performing consumer load profiles' median correlation with the used standard PV load profile is clearly, by approx. 0.036, more negative than for the low-performing, and also the interquartile range extends further into the negative range than for the low-performing consumer load profiles, suggesting that tendentially lower values might lead to a higher increase in profit through operating in a community.

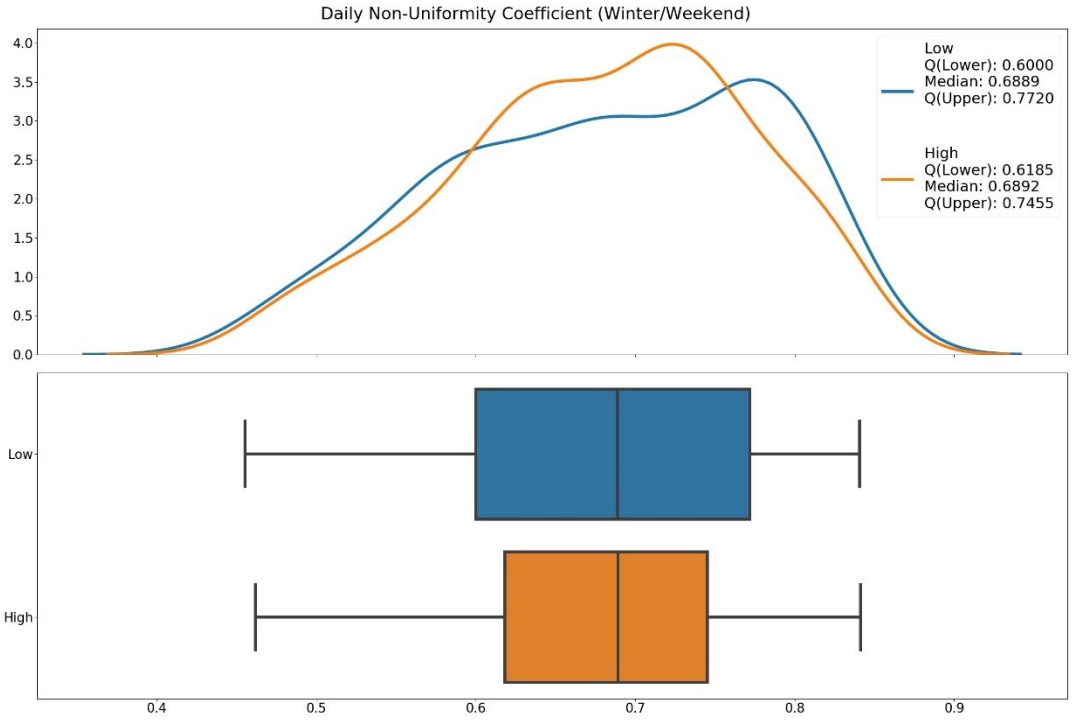

Figure 18: KDE plot and box-whisker plot of the value distributions of the parameter Daily Non-Uniformity Coefficient (Winter/Weekend) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

For the parameter Daily Non-Uniformity Coefficient (Winter/Weekend), vividly put the base load relative to the average load, we observe in Figure 18 that although the values are unequally distributed, no unambiguous tendency can be identified, and the median values are almost identical. Looking at the KDE plot, it seems that rather low (0.5-0.6) and rather high (0.75-0.9) values appear more often for low-performing consumer load profiles than for high performing, whose values are more often moderate (0.6-0.75), leading to the supposition that extremely low and extremely high values regarding the consumption valleys, relative to the consumption mean, on winter weekend days could lead to less beneficial community constellation. This extends the temporal hypotheses drawn from the parameter Minimum Time-of-Use (Winter/Weekend) above by a consumption level dimension.

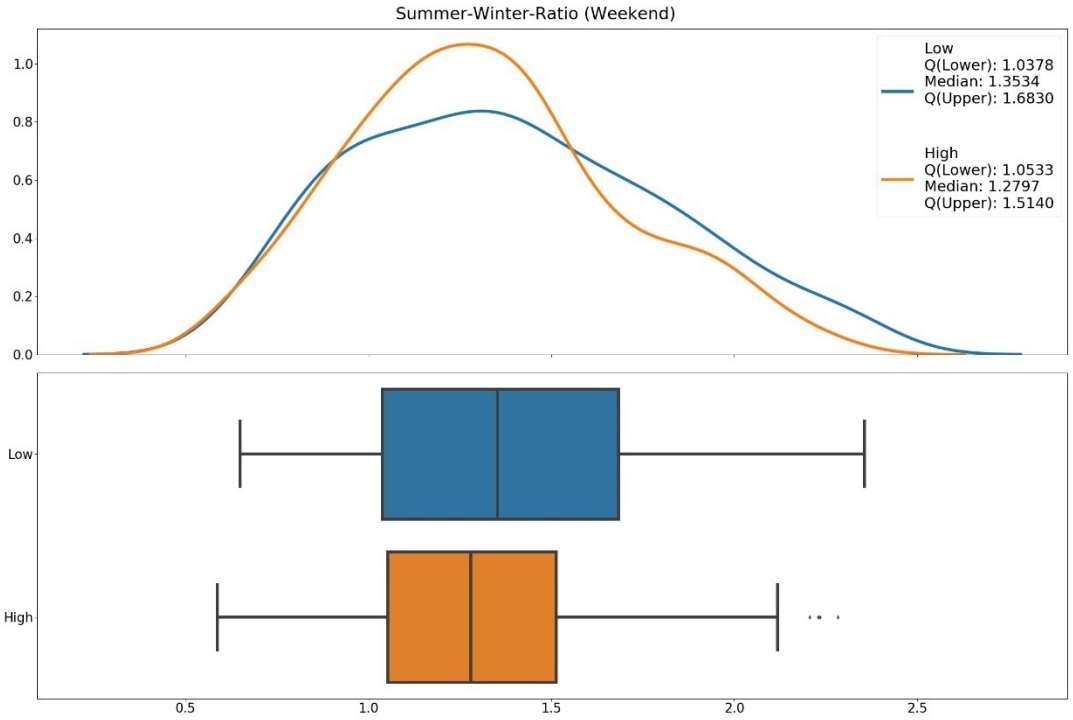

Figure 19: KDE plot and box-whisker plot of the value distributions of the parameter Summer-Winter-Ratio (Weekend) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

Figure 19 shows the distributions for the parameter Summer-Winter-Ratio (Weekend), proposed in this work. It describes the ratio of the cumulated consumption in summer months to that in winter months. It can be observed that the interquartile range of the consumer load profiles of high-performing communities is much more condensed, comparing the upper quartile borders and the whisker lengths suggests that at least 12.5% of the parameter values of the low-performing are greater than 75% of the parameter values of the high-performing. From the significantly differing median values (more than 7% smaller for high-performing) we can conclude that on summer weekend days, high-performing communities' consumer load profiles have a lower consumption compared to their respective winter weekend days than that of low-performing. This is observation is consistent with the visualization of the consumer load profiles in Figure 10 and 11, and the corresponding concluding hypothesis of Chapter 4.2.1., which was drawn in the context of all five community load profiles.

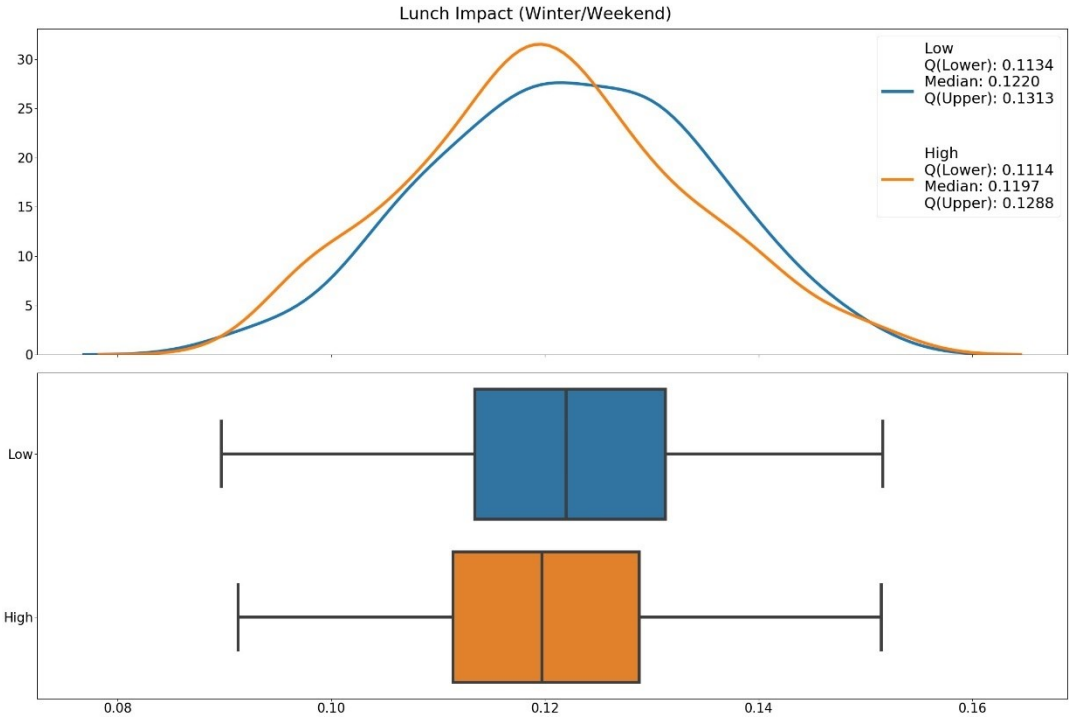

Figure 20: KDE plot and box-whisker plot of the value distributions of the parameter Lunch Impact (Winter/Weekend) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

The parameter Lunch Impact (Winter/Weekend), whose distributions are given in Figure 20, describes the scaled ratio from the average consumption during lunch hours (here: 11 a.m. to 1 p.m.) to the daily average consumption, i.e. higher values indicate a higher average consumption during lunch hours compared to the overall average consumption and vice versa. Looking at Figure 20 reveals that the interquartile range of the Lunch Impact values of the high-performing communities' consumer load profiles is visibly shifted to the left compared to the low-performing, indicating that a smaller consumer consumption during lunch hours on winter weekend days could be favorable for a stronger increase in community performance. Inspecting the average load profiles in the context of general Community Profit in chapter 4.2.1. confirms this hypothesis, especially for the 2nd high-performing consumer load profile, a clearly lower consumption around 12 p.m. can be observed, confirming this assumption.

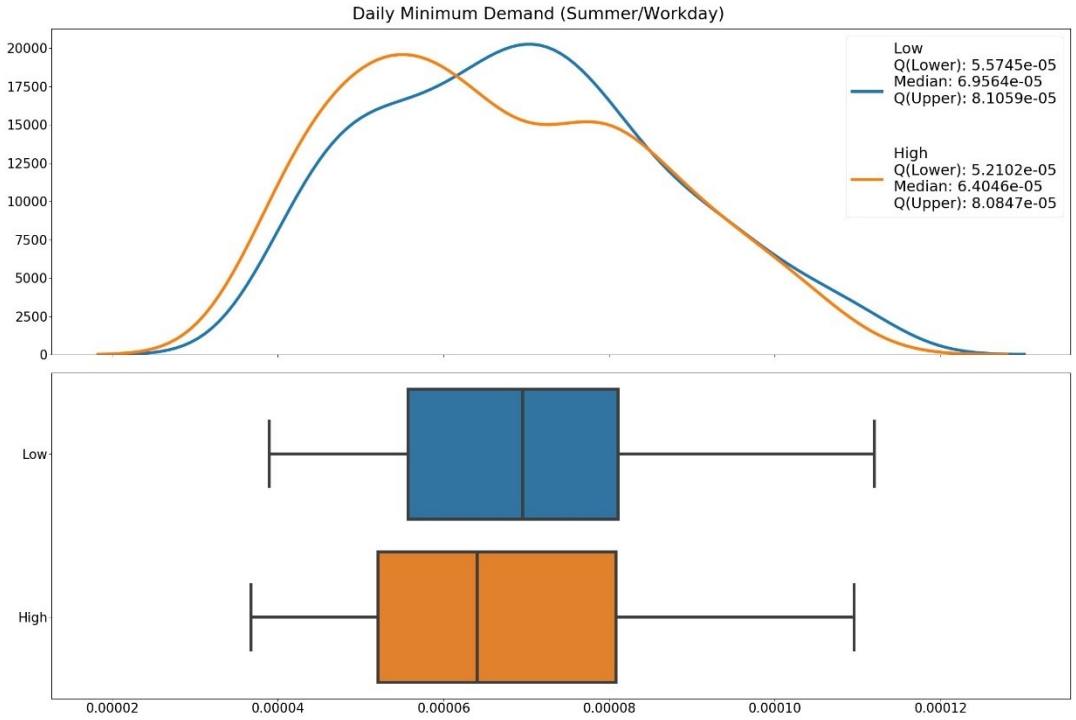

Figure 21: KDE plot and box-whisker plot of the value distributions of the parameter Daily Minimum Demand (Summer/Workday) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

Figure 21 depicts the distributions of the parameter Daily Minimum Demand (Summer/Workday). An analysis of the median values reveals that the Daily Minimum Demand for the high-performing communities' consumer load profiles is approx. 8% lower than for the low-performing, and also the KDEs and the interquartile range are clearly shifted to the left, suggesting that consumers' profiles with lower base consumption values on summer workdays seem to be more favorable for an increased gain in profit by participating in a community.

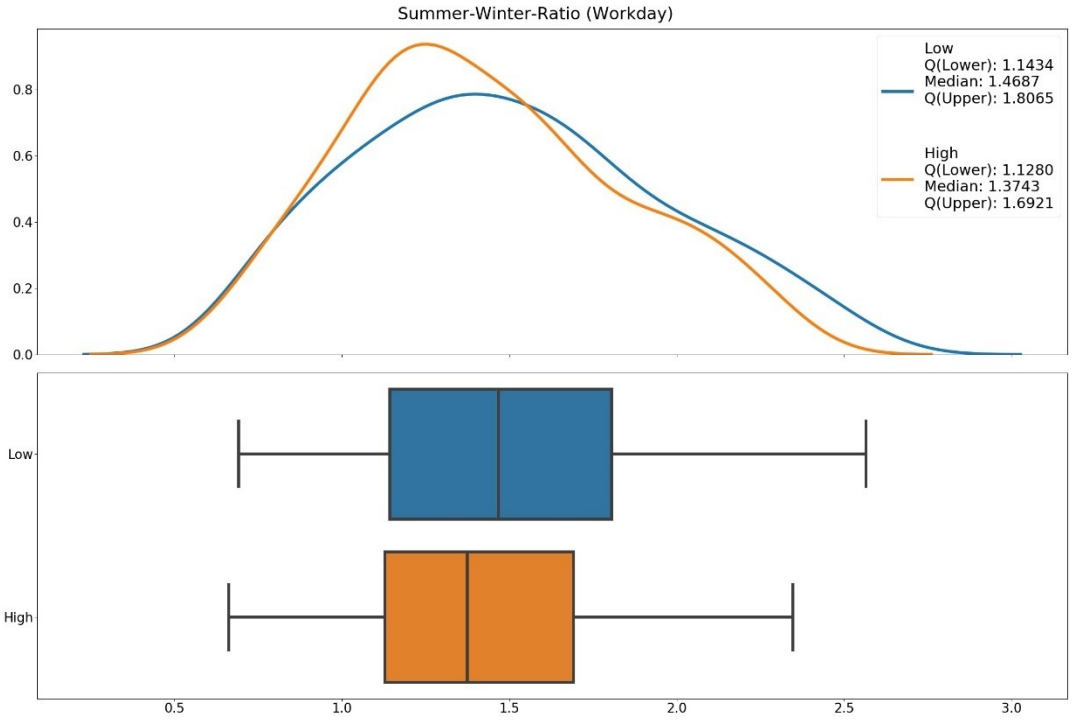

Figure 22: KDE plot and box-whisker plot of the value distributions of the parameter Summer-Winter-Ratio (Workday) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

For the Summer-Winter-Ratio on workdays, the difference in the median values of low- and high-performing communities' consumer load profiles shown in Figure 22 is even greater than that for weekend days, namely more than 11% difference; clearly more high-performing consumer load profiles' values fall into the range from 1.0 to 1.5 than for low-performing. This is confirmed by inspecting the average representative winter and summer load profiles in the context of overall community profit in chapter 4.2.1., and is consistent with the concluding hypotheses proposed there: for low community profit, the 1st consumer load profile shows higher values in summer than the high-performing and vice versa for winter months, resulting in a relatively higher Summer-Winter-Ratio. These proportions seem not to apply for the 2nd consumer load profile, yet the differences seem to be relatively smaller than for the 1st consumer, which suggests that aggregating both consumer load profiles still results in a higher parameter value for low-performing communities.

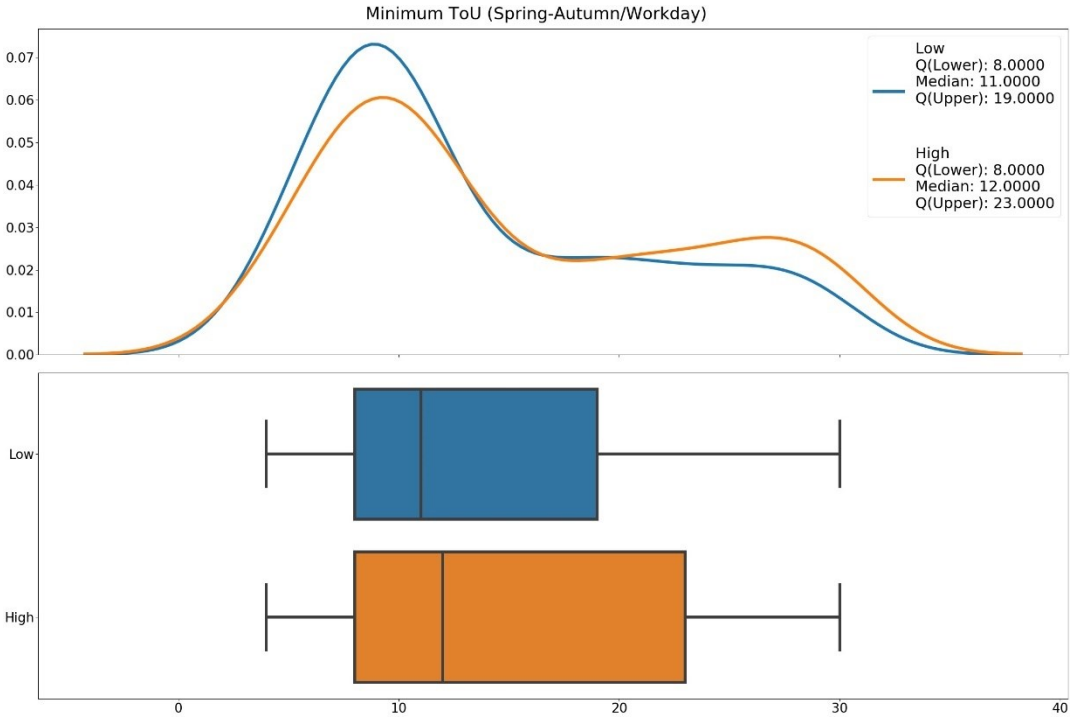

Figure 23: KDE plot and box-whisker plot of the value distributions of the parameter Minimum Time-of-Use (Spring-Autumn/Workday) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

Observing the distributions of the parameter Minimum Time-of-Use (Spring-Autumn/Workday) in Figure 23 displays a similar distribution pattern to that of the Minimum Time-of-Use (Winter/Weekend) analyzed above: again, the high-performing load profiles feature a bimodal distribution with a second, small mode at the late lunch hours, around 2 p.m. Although the medians are identical to that on winter weekend days, both interquartile ranges stretch out far more to the right, in particular that of the consumer load profiles causing a higher increase in community profit. The upper quartile shows that 25% of the high-performing consumer load profile's minimum time of use is later than 11:30 a.m., 25% of the low-performing is later than only 9:30 a.m., which confirms the hypothesis that a *later* minimum time-of-use seems to be a favorable characteristic of consumer load profiles for more profitable energy sharing communities, not only for winter weekend days, but also for spring and autumn workdays.

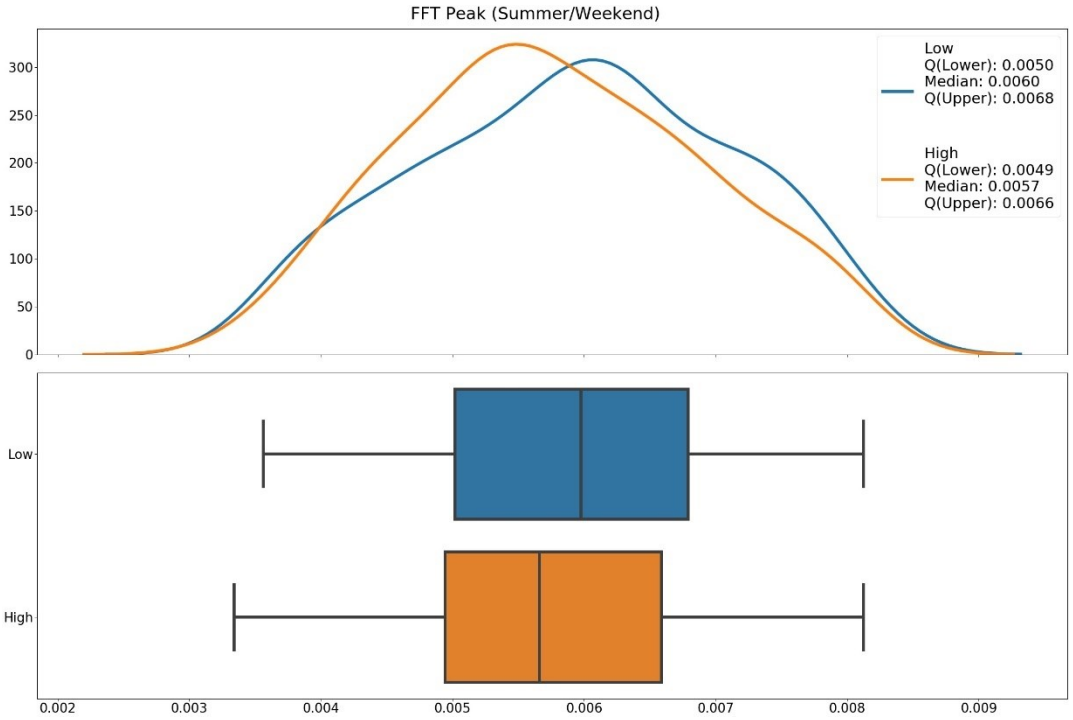

Figure 24: KDE plot and box-whisker plot of the value distributions of the parameter FFT Peak (Summer/Weekend) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

The distribution shapes and interquartile ranges for the parameter FFT Peak (Summer/Weekend) in Figure 24 imply that the distribution of low-performing communities' consumer load profiles' values are shifted slightly to the right, and a comparison of the median indicates that lower maxima of the absolute values of the Fast Fourier Transform of the aggregated consumers' load profiles for summer weekend days seem to be more advantageous than higher values, extending the hypotheses proposed for the FFT Peak on summer workdays above also to summer weekend days.

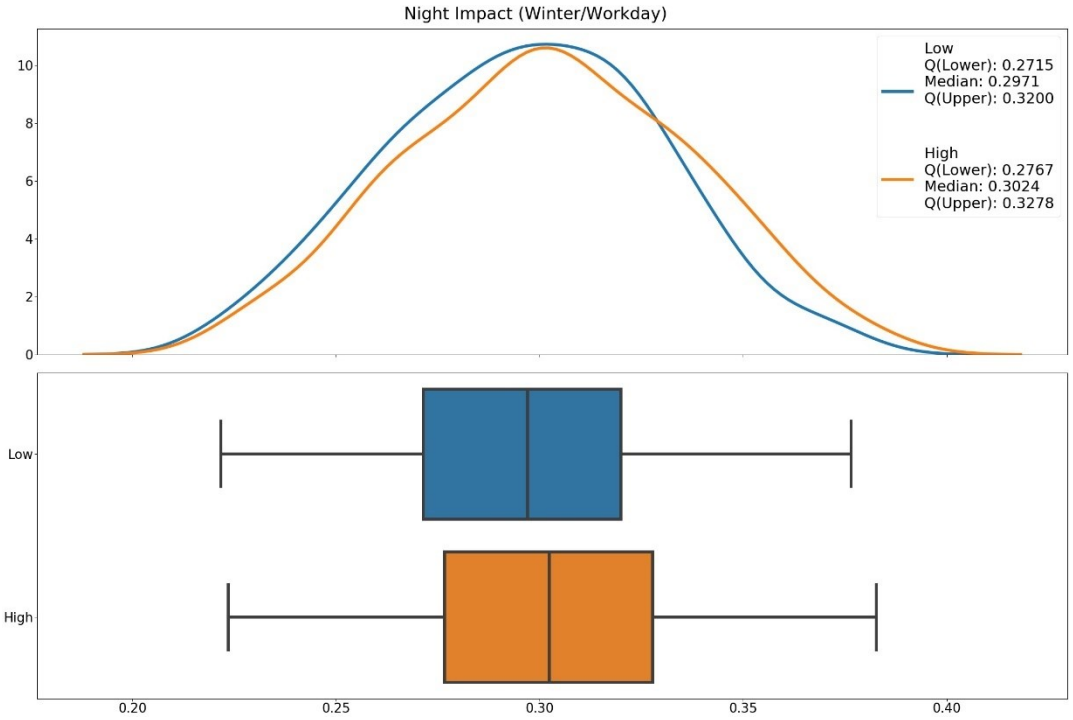

Figure 25: KDE plot and box-whisker plot of the value distributions of the parameter Night Slope (Winter/Workday) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

The Night Impact, analogous to the Lunch Impact, describes the scaled ratio from the average consumption during night hours (here: 11 p.m. to 6 p.m.) to the overall daily average consumption, i.e. higher values indicate a higher consumption during the night compared to the average daily consumption and vice versa. Although there are no strong trends visible in the corresponding Figure 25, the KDE plot shows that vales up to approx. 0.33 seem to appear more often for consumer load profiles that cause a lower increase in community profit, while values greater than 0.33 appear more often for high-performing load profiles; this is confirmed by the sightly right-shifted interquartile ranges and median of high-performing consumer load profiles, and is also backed by the visual analysis of winter workday load profiles in the overall Community Profit context in Figure 11 in Chapter 4.2.1.; this figure also suggests that one possible reasoning for this might be the following: analyzing the corresponding high-performing *prosumer* load profiles reveals that the latter feature a lower consumption level throughout the midday than the low-performing, which might result in more generated excess PV energy to be stored in the battery systems, which in turn could later be used to cover the relatively high nighttime consumption of the *consumers*, resulting in a better energy distribution, a better self-consumption rate and therefore better overall community profit.

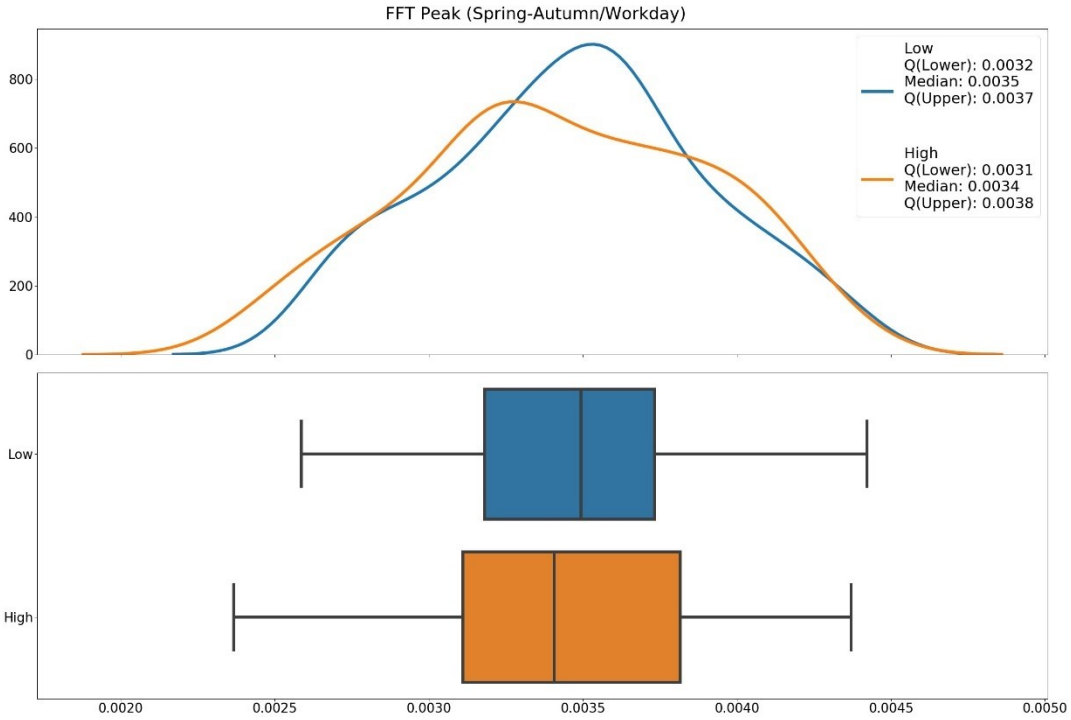

Figure 26: KDE plot and box-whisker plot of the value distributions of the parameter FFT Peak (Spring-Autumn/Workday) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

The distribution pattern of the FFT Peak on spring and autumn workdays depicted in Figure 26 appears to be less unambiguous than that of the FFT Peak on summer workdays and weekend days analyzed above. Nevertheless, the slightly lower median indicates once more that lower FFT peaks seem to be a favorable consumer load profile characteristic.

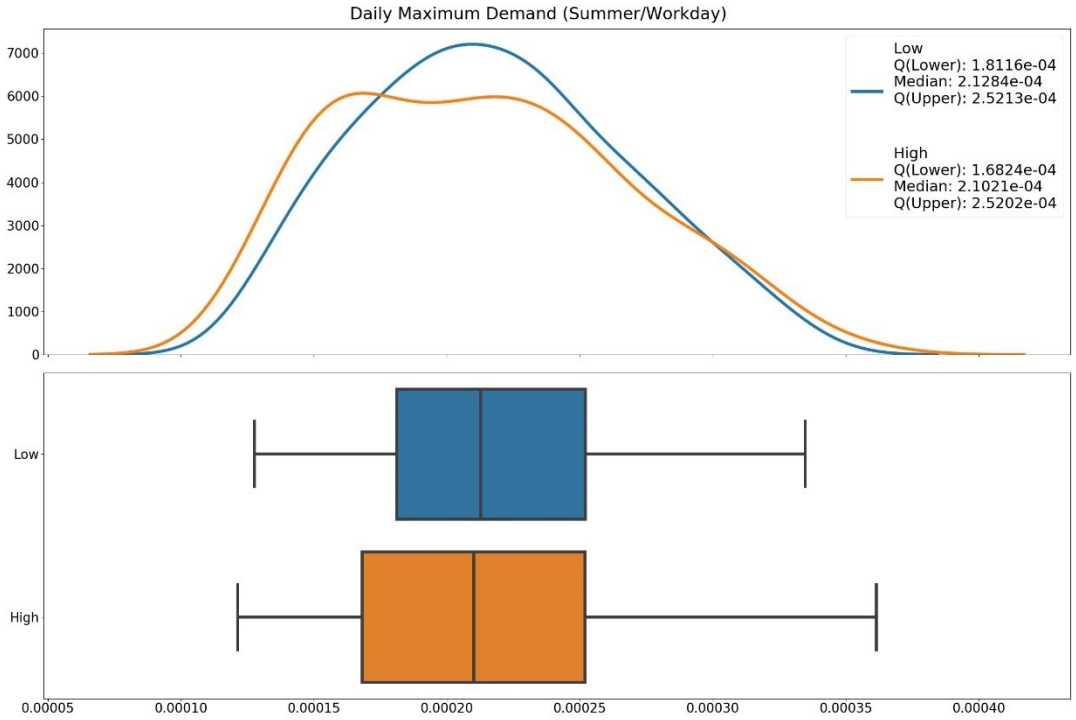

Figure 27: KDE plot and box-whisker plot of the value distributions of the parameter Daily Maximum Demand (Summer/Workday) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

Figure 27 shows the distributions of the consumer households' Daily Maximum Demand parameter values. Although the differences are rather small here, it can be stated that consumer load profiles that cause a higher profit in a community appear to feature lower daily peaks. In particular, observing the quantiles reveals that the smallest 25% of values of the high-performing consumers is clearly lower located than the smallest 25% of values of the low-performing; this is confirmed by the KDE plot, which indicates that normalized load values up to approx. 0.00017 are observed more often for high-performing consumer load profiles. In general, this suggests the assumption that smaller daily peaks on summer workdays might be an advantageous consumer load profile characteristic, which seems to be once more consistent with the concluding hypotheses from Chapter 4.2.1. and the above observations on the Summer-Winter-Ratio. One possible complementing explanation for this, could be, that extraordinary high energy demand outliers of the consumers, as they can be found in steep-peaked load profiles, can be covered less likely by the energy-generating prosumers due to their extremity, leading to the need to purchase energy from the grid, resulting in turn in a lower self-consumption rate and therefore lower community profit.

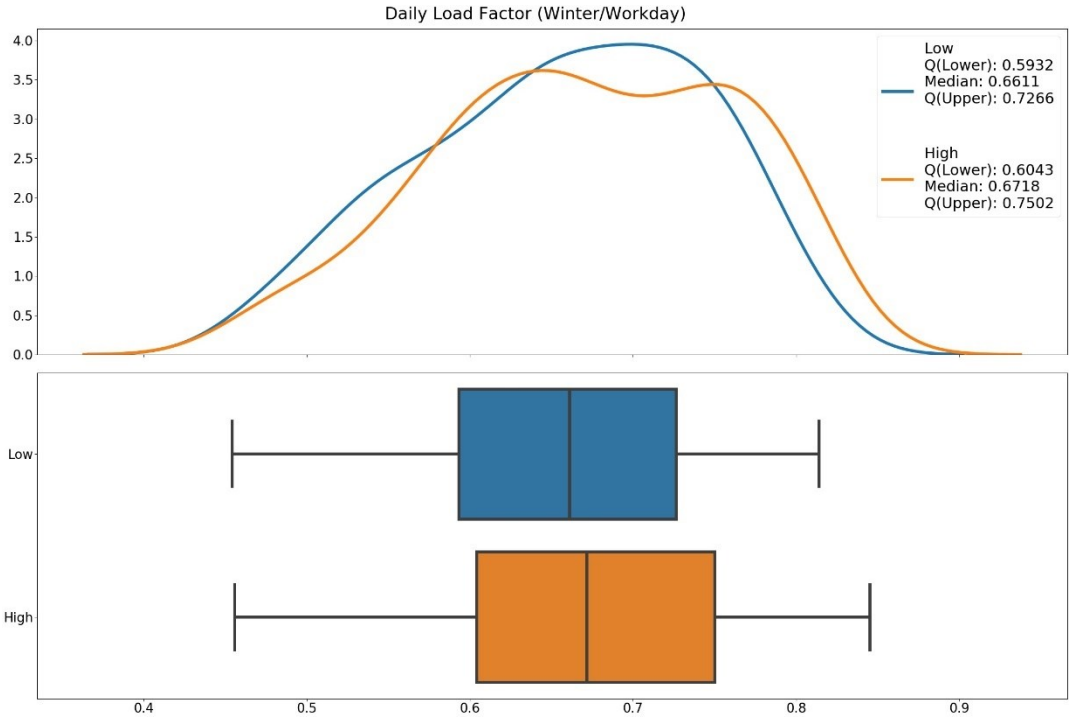

Figure 28: KDE plot and box-whisker plot of the value distributions of the parameter Daily Load Factor (Winter/Workday) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

The parameter Daily Load Factor (Winter/Workday), whose distributions are depicted in Figure 28, describes the ratio from the daily average to the daily maximum consumption. That means that a peak, which is more extreme in comparison to the daily average, results in a lower parameter value and vice versa, or to say it with the illustrative formulation from [12]: the Daily Load Factor measures the "peakiness" of a load profile. Observing the distributions reveals a slightly higher interquartile range and median for the high-performing communities' consumer load profiles on winter workdays. Also, the KDE plot makes clear that values greater than approx. 0.75, i.e. load profiles whose peak is at most one-third of the daily average higher than the daily average itself, appear more often for high-performing consumer load profiles. A short complementary analysis shows that while the Daily Load Factor values on spring and autumn workdays are almost identic for high- and low performing consumer load profiles, the latter assumption also holds for summer workdays. Additionally, inspecting Figure 11 in Chapter 4.2.1. shows that the high-performing communities' 1st consumer on summer workdays features a higher consumption in the non-peak time areas than the low-performing, indicating that the higher Daily Load Factor derives not from higher peaks but from a higher daily average and therefore overall higher consumption level on winter days, confirming the above hypotheses on the Summer-Winter-Ratio once more.

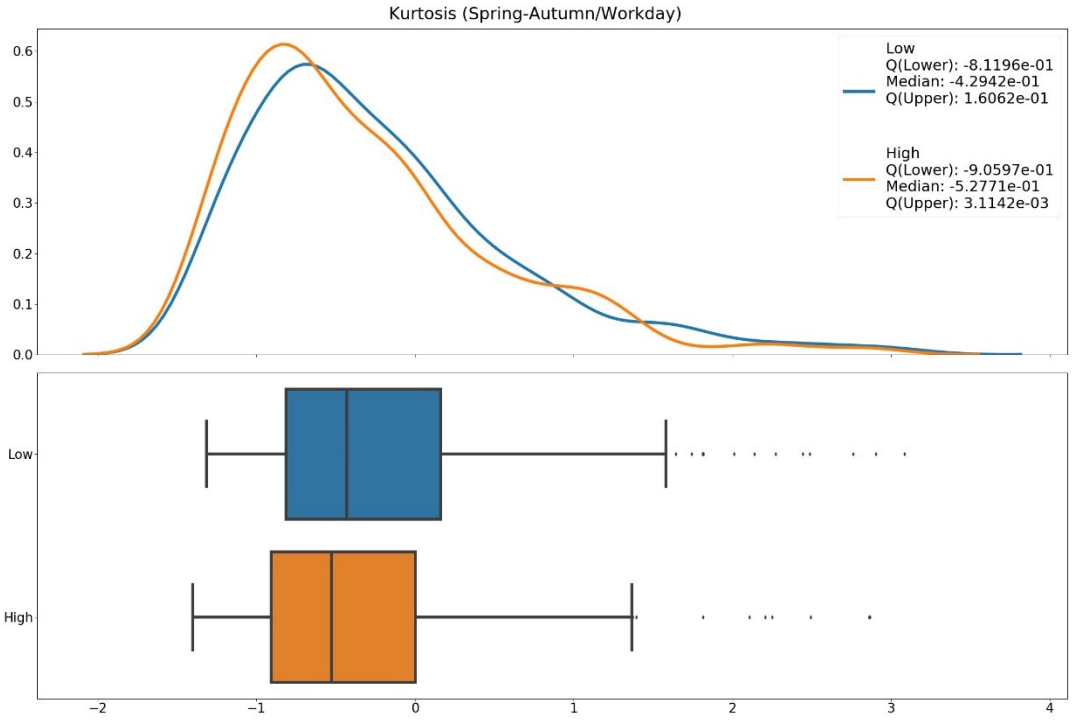

Figure 29: KDE plot and box-whisker plot of the value distributions of the parameter Kurtosis (Spring-Autumn/Workday) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

Figure 29 depicts the distributions of the parameter Kurtosis (Spring-Autumn/Workday). It strikes that the KDE, the interquartile range and the median are all shifted towards lower values for consumer load profiles of high-performing communities. This extends the hypothesis from above, that more negative Kurtosis values of the daily load value distributions of consumer households might be an indicator for an increased profit through energy communities, from winter workdays to spring and autumn workdays. A short complementary analysis of the Kurtosis values for summer workdays shows that this assumption seems to be valid only for spring, autumn and weekend workdays, whereas for summer workdays the opposite seems to apply.

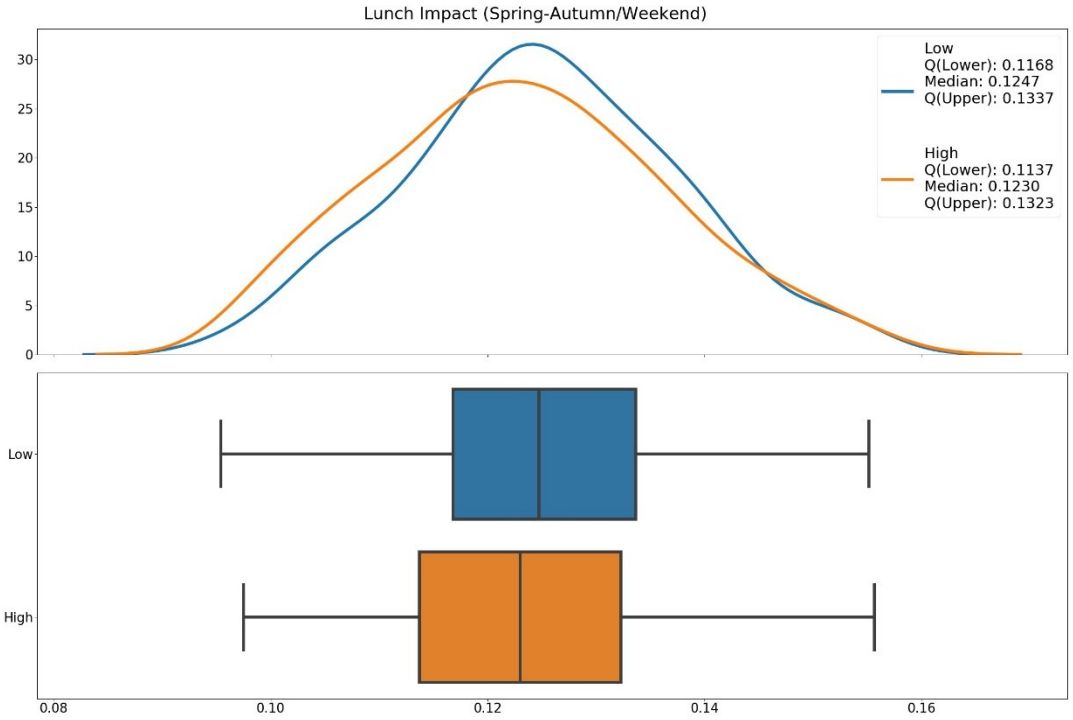

Figure 30: KDE plot and box-whisker plot of the value distributions of the parameter Lunch Impact (Spring-Autumn/Weekend) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

Although to a lesser extent, the observations on the distributions of the Lunch Impact (Spring-Autumn/Weekend) parameter values in Figure 30 are almost identical to the observations on the Lunch Impact for winter weekend days above, further confirming the assumption that lower relative consumption around the lunch hours on weekends might be a favorable characteristic for a communities' consumer load profiles.

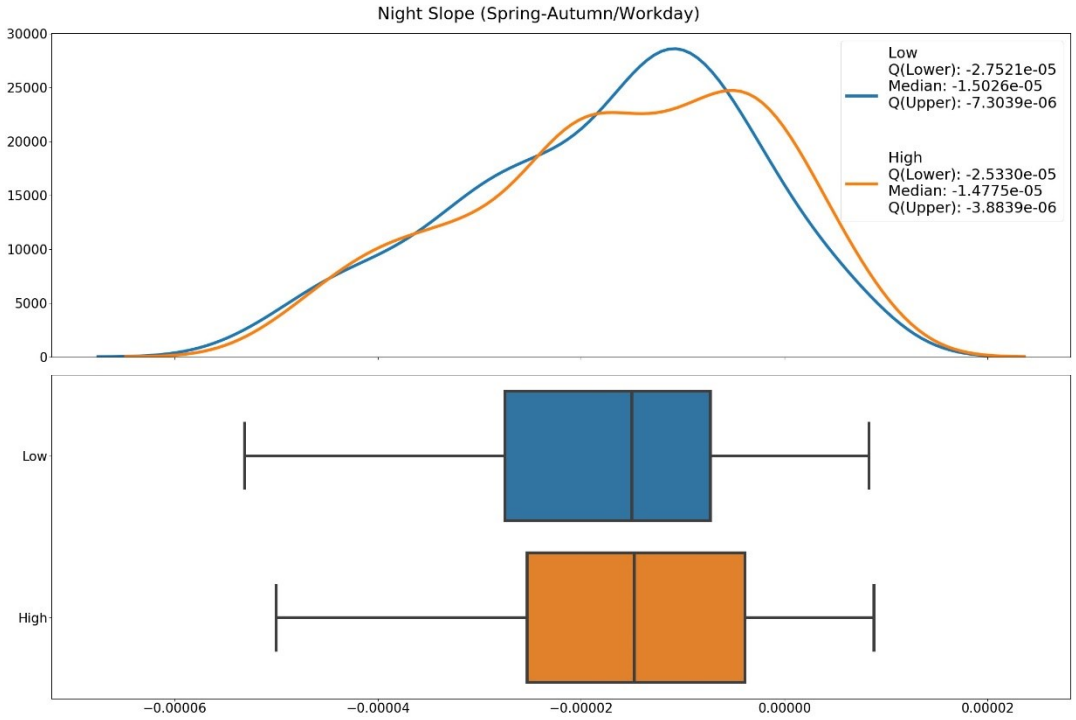

Figure 31: KDE plot and box-whisker plot of the value distributions of the parameter Night Slope (Spring-Autumn/Workday) of low- (blue) and high-performing (orange) communities' aggregated and averaged consumer load profiles

Figure 31 shows the distributions of the Night Slope, this time analyzed for spring and autumn workdays. As explained above, a more negative Night Slope value indicates a steeper drop in consumption from the early to the later evening hours (9 p.m. to 11 p.m.). The interquartile range and the median of the high-performing communities' consumer Night Slope values is shifted clearly towards more positive values. The KDE plot reveals: while the distributions seem to oscillate and there is no clear trend for strongly negative values, it seems that relatively negative values (approx. difference in normalized load from -0.000025 to -0.000005), i.e. moderate evening consumption drops, are clearly more frequently observed for low-performing consumer load profiles, and slightly negative to even positive values (approx. larger than -0.000005), i.e. very soft consumption drops in the evening hours or even increases, seem to appear more often for high-performing households. That means, the hypotheses that less steep drops are a favorable characteristic from the analysis of the Night Slope on winter weekend days also seems to apply for spring and autumn workdays.

5. Conclusion & Outlook

This work investigated the relationship between properties of residential load profiles and their suitability for energy sharing communities, consisting of three prosumer- and two consumer households each as defined in [1]. Hereby, the *suitability* was quantified by means of two community simulation results presented in [1], the Community Profit and Community Gain.

To answer the first research question, a comprehensive set of electricity parameters characterizing the properties of residential load profiles was compiled by combining an extensive cross-literature research on existing parameters with the construction of new parameters, partly by modifying and extending existing concepts, partly from observations made during the analysis of the available data. The parameter compilation was further refined by adding seasonal and weekly differentiation criteria. In addition, as the characterization of load profiles on community-level requires some form of aggregation, we propose two, to our best knowledge new, approaches for parameter calculation on community-level in this work, namely *Parametrization-First-Aggregation-Second* and *Aggregation-First-Parametrization-Second*, that complement our parameter compilation and could serve as a basis for potential future research on energy sharing communities. While most of the observed papers used only a relatively small selection of parameters each, this work offers a wide-ranging collection; nonetheless, there is still potential to extend the collection even further in future research: For instance, in the doctoral thesis [12], more time-series-specific approaches, such as Autoregressive Markov-Chains, Fourier Transforms, Gaussian Processes and Wavelets are used to characterize the temporal characteristics of load profiles. Another interesting and very recent approach, that could also be applied to energy sharing community research, is proposed in [23]: here, all days of all households are clustered by four parameters, the average consumption for three time periods and the daily variance. Subsequently, each day of every household is assigned to its corresponding closest cluster ID, and the household is finally characterized by the statistical mode of all of its daily cluster IDs. A further, interesting possibility for future research on parameters we propose here is the usage of the Python package "*tsfresh (Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests)*, an open-source library for the automatic extraction of up to 794 time series features, presented in [49]. In general, the parameter collection we propose offers potential to serve as a basis for and contribution to arbitrary types of research on electricity load profiles, enabling for instance classification, clustering, load forecasting, or pattern mining approaches.

In order to investigate the second research question, multiple Machine Learning techniques for classification problems, namely Logistic Regression, Extreme Gradient Boosting and Multi-Layer Perceptron, were reviewed, and a 3-fold cross-validated grid search was applied for hyperparameter optimization.

Hereby, the models were not only trained on different hyperparameter combinations, but also on different parameter calculation approaches (PFAS and FPAS), different classification tasks (Classification of Community Profit and Community Gain), and on different input configurations (training on all features and on a heuristic selection of the most important features), resulting in the 3-fold cross-validated fitting and evaluation of 137,760 different model configurations. Despite this substantial research, none of the models yielded a satisfactory level of accuracy in predicting if a constellation of potential community households will lead to a low, medium, or high profit based on the proposed electricity parameters. We proposed several possible hypotheses to explain this failure; among others one reasoning might be that Machine Learning models are not able to predict the result of the complex community simulation process solely based on load profile parameters, and need additional, more broadly diversified input parameters representing other properties influencing the simulation process in order to make accurate predictions, suggesting the extension of the models by such parameters, for instance socioeconomical or climatic parameter, in further research. This hypothesis might be correlated with the results presented in [23]. Another consideration proposed here is that averaging on a temporal level could obfuscate seemingly small, but for the simulation crucial characteristics and events on daily or intra-daily level, or that aggregation on community level could blur the complex interdependencies between households within a community. This hypothesis would in turn suggest further research on different averaging and aggregation with regard to energy sharing communities. Furthermore, we proposed utilizing more complex Machine Learning models or working with further electricity parameters as possible future research paths aiming to enable the accurate prediction of households' suitability for energy sharing communities.

With regard to the third research question, two different explorative data analysis approaches were conducted and combined. Firstly, representative daily community load profiles, differentiated by seasonal and weekly criteria, were constructed and visually analyzed, which resulted in several indications of which qualitative properties of pro- and consumer load profiles might be favorable for a high energy sharing community performance. Secondly, we proposed a heuristic selection approach, in which the Two-Sample Kolmogorov-Smirnov test was used to identify parameters whose value distributions differ considerably for low- and high performing communities. This selection of 20 parameters included 6 of the parameters proposed in this work, indicating that these parameters could serve as an effective distinction criterion regarding household characterization. Subsequently, the selected value distributions and their non-parametric statistical properties were visualized with kernel density estimations and box-whisker plots. As a result of detailed investigations of the distributions, interconnected with the observations from the daily average load profile analysis, in Chapter 4.2.2. we propose a set of qualitative hypotheses about certain parameter value characteristics that might be advantageous properties for building beneficial energy sharing communities and provide them with possible explanatory approaches.

Amongst others, one exemplary hypothesis is, that regarding winter days, consumers with a later Minimum Time-of-Use might be more favorable for a high increase in profit by participating in a community. Another proposed hypothesis, which was strongly suggested by both, the load profile analyses *and* several parameter value distributions, is that the average prosumer load profiles of generally high-performing communities seem to feature a significantly higher consumption level and steeper peaks on summer days compared to their respective winter days, while for corresponding consumer load profiles the opposite applies, which is also visible in the parameter value distributions of the parameter Daily Maximum Demand and the proposed parameter Summer-Winter-Ratio. As the proposed hypotheses were built solely on the available data and therefore need further revision, an interesting subject of future research could be, to investigate if these hypotheses can be confirmed or rejected for different datasets, and if applicable to refine and further extend them. Another possible path for potential research could be to apply dimensionality reduction techniques, such as the widely used *Principal Component Analysis (PCA)*, described in detail for instance in [50], or advanced techniques as the *t-Distributed Stochastic Neighbor Embedding (t-SNE)*, a Machine Learning algorithm for visualization based on minimizing the Kullback-Leibler divergence between high- and a low-dimensional probability distributions and originally presented in [51]. Thereby, new beneficial load profile characteristics could be identified, and a novel class of input parameters for the research on further Machine Learning models using residential load profiles to predict the suitability for energy sharing communities could be provided.

6. Declaration about the Thesis

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 05.10.2020 _____

Kai Firschau

References

1. Henni S. A Platform-based Sharing Economy Model for Residential Solar Generation and Storage Capacities. Forty-First International Conference on Information Systems, India 2020. 2020.
2. Roberts J, Frieden D, d'Herbemont S. Energy Community Definitions. Deliverable Developed Under the Scope of the COMPILE Project: Integrating Community Power in Energy Islands (H2020-824424). 2019.
3. McKinney W. Data Structures for Statistical Computing in Python. In: Python in Science Conference; June 28 - July 3 2010; Austin, Texas: SciPy; 2010. p. 56–61. doi:10.25080/Majora-92bf1922-00a.
4. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. In: : Austin, TX. p. 61.
5. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825–30.
6. Cembranel SS, Lezama F, Soares J, Ramos S, Gomes A, Vale Z. A Short Review on Data Mining Techniques for Electricity Customers Characterization. In: : IEEE. p. 194–199.
7. Chicco G, Napoli R, Postolache P, Scutariu M, Toader C. Customer characterization options for improving the tariff offer. IEEE Transactions on Power Systems. 2003;18:381–7.
8. Milton M-A, Pedro C-O, Xavier S-G, Guillermo E-E. Characterization and Classification of Daily Electricity Consumption Profiles: Shape Factors and k-Means Clustering Technique. In: : EDP Sciences. p. 8004.
9. Rajabi A, Li L, Zhang J, Zhu J, Ghavidel S, Ghadi MJ. A review on clustering of residential electricity customers and its applications. In: : IEEE. p. 1–6.
10. Ramos S, Vale Z. Data mining techniques application in power distribution utilities. In: : IEEE. p. 1–8.
11. Abreu JM, Pereira FC, Ferrão P. Using pattern recognition to identify habitual behavior in residential electricity consumption. Energy and buildings. 2012;49:479–87.
12. McLoughlin F. Characterising domestic electricity demand for customer load profile segmentation. 2013.
13. Bicego M, Farinelli A, Grosso E, Paolini D, Ramchurn SD. On the distinctiveness of the electricity load profile. Pattern Recognition. 2018;74:317–25.
14. Luo X, Hong T, Chen Y, Piette MA. Electric load shape benchmarking for small-and medium-sized commercial buildings. Applied energy. 2017;204:715–25.
15. Al-Otaibi R, Jin N, Wilcox T, Flach P. Feature construction and calibration for clustering daily load curves from smart-meter data. IEEE Transactions on industrial informatics. 2016;12:645–54.

16. Ferraro P, Crisostomi E, Tucci M, Raugi M. Comparison and clustering analysis of the daily electrical load in eight European countries. Electric Power Systems Research. 2016;141:114–23.
17. Tanwar AK, Crisostomi E, Ferraro P, Raugi M, Tucci M, Giunta G. Clustering analysis of the electrical load in european countries. In: : IEEE. p. 1–8.
18. Chicco G. Overview and performance assessment of the clustering methods for electrical load pattern grouping. Energy. 2012;42:68–80.
19. McLoughlin F, Duffy A, Conlon M. Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study. Energy and buildings. 2012;48:240–8.
20. Fahim M, Sillitti A. Analyzing load profiles of energy consumption to infer household characteristics using smart meters. Energies. 2019;12:773.
21. Price P. Methods for analyzing electric load shape and its variability.
22. Rathnayaka AD, Potdar VM, Hussain O, Dillon T. Identifying prosumer's energy sharing behaviours for forming optimal prosumer-communities. In: : IEEE. p. 199–206.
23. Pena-Bello A, Barbour E, Gonzalez MC, Yilmaz S, Patel MK, Parra D. How Does the Electricity Demand Profile Impact the Attractiveness of PV-Coupled Battery Systems Combining Applications? Energies. 2020;13:4038.
24. Félix LCM, Rezende SO, Monard MC, Caulkins CW. Transforming a regression problem into a classification problem using hybrid discretization. Computación y Sistemas. 2000;4:44–52.
25. Cleveland RB, Cleveland WS, McRae JE, Terpenning I. STL: A seasonal-trend decomposition. Journal of official statistics. 1990;6:3–73.
26. Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding.
27. Kodinariya TM, Makwana PR. Review on determining number of Cluster in K-Means Clustering. International Journal. 2013;1:90–5.
28. Piao M, Ryu KH. Local characterization-based load shape factor definition for electricity customer classification. IEEJ Transactions on Electrical and Electronic Engineering. 2017;12:S110-S116.
29. Tang J, Alelyani S, Liu H. Feature selection for classification: A review. Data classification: Algorithms and applications. 2014:37.
30. scikit-learn developers. Chi-Squared Feature Selection. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html. Accessed 19 Sep 2020.
31. Albon C. ANOVA F-value For Feature Selection. 2017. https://chrisalbon.com/machine_learning/feature_selection/anova_f-value_for_feature_selection/. Accessed 20 Sep 2020.
32. Ho TK. Random decision forests. In: : IEEE. p. 278–282.
33. Learned-Miller EG. Introduction to supervised learning. I: Department of Computer Science, University of Massachusetts. 2014.

34. Feng J, Xu H, Mannor S, Yan S. Robust logistic regression and classification. In: p. 253–261.
35. Karsmakers P, Pelckmans K, Suykens JAK. Multi-class kernel logistic regression: a fixed-size implementation. In: : IEEE. p. 1756–1761.
36. scikit-learn developers. SGD Classifier. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html. Accessed 21 Sep 2020.
37. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction: Springer Science & Business Media; 2009.
38. Ray PK, Mohanty SR, Kishor N, Catalão JPS. Optimal feature and decision tree-based classification of power quality disturbances in distributed generation systems. IEEE Transactions on Sustainable Energy. 2013;5:200–8.
39. scikit-learn developers. Ensemble Methods. https://scikit-learn.org/stable/modules/ensemble.html. Accessed 21 Sep 2020.
40. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: p. 785–794.
41. scikit-learn developers. Neural Network Models (Supervised). https://scikit-learn.org/stable/modules/neural_networks_supervised.html#neural-networks-supervised. Accessed 7 Sep 2020.
42. scikit-learn developers. GridSearchCV. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html. Accessed 24 Sep 2020.
43. Bottou L. Stochastic gradient descent tricks. In: Neural networks: Tricks of the trade: Springer; 2012. p. 421–436.
44. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014.
45. Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Mathematical programming. 1989;45:503–28.
46. Hossin M, Sulaiman MN. A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process. 2015;5:1.
47. Hesamian G, Chachi J. Two-sample Kolmogorov–Smirnov fuzzy test for fuzzy random variables. Statistical Papers. 2015;56:61–82.
48. Zanin Zambom A, Dias R. A Review of Kernel Density Estimation with Applications to Econometrics. arXiv. 2012:arXiv-1212.
49. Christ M, Braun N, Neuffer J, Kempa-Liehr AW. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). Neurocomputing. 2018;307:72–7. doi:10.1016/j.neucom.2018.03.067.
50. c Abdi H, Williams LJ. Principal component analysis. vol. 2010;2:27.
51. van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008;9:2579–605.