THESIS WORK FOR DUAL MASTER'S DEGREE

KIT M.Sc. in Mechanical Engineering

ITBA Mag. in Energy and Environment

Machine Learning-based Analysis of Residential Electricity Consumption Behavior for Consumers and Prosumers

Tamo Werner
Mechanical Engineering - KIT

Tutor
M.Sc. Jiao Jiao

Examiners
apl. Prof. Dr.-Ing. Ralf Mikut, KIT/ IAI
Prof. Dr. Veit Hagenmeyer, KIT/ IAI

Karlsruhe
15/10/2021

The present work is identical to the thesis handed to the Institute for Automation and Applied Informatics on the 15th of October 2021. This copy is made to verify the double master degree of the ITBA-KIT-Cooperation. Only the cover page is modiefied in this version.

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

Mannheim, 15.10.2021

. . . .

(Tamo Werner)

Abstract

With the shift towards a more sustainable energy system, the need for a better understanding of the behavior development over time of consumers and prosumers arises. Despite the growing penetration of smart meter infrastructure, the availability of energy usage behavior data is still limited, due to privacy and security concerns. Thus, connecting and comparing existing datasets is the key to observe the user behavior shifts as well as enhancing the utility of the available data.

In the present work, a novel workflow for combined analysis on multiple smart meter datasets is proposed, which links datasets with different scopes, temporal origins and specifications. In general, there are 4 steps: data preprocessing, clustering, location dependency check and dataset linking. First, the meteorological seasons combined with weekdays and weekends are picked for data segmentation in the data preprocessing, followed by missing value validation and normalization based on the maximum and minimum consumption value of each household. Thereafter, K-means clustering algorithm is applied to group the user behaviors, which stands out by 0.8186 Silhouette coefficient (SIL) and 0.2884 Davies-Bouldin Index (DBI) among Fuzzy C-Means and hierarchical clustering approach. Subsequently, two validation approaches on the location dependency, cluster center correlation (0.8048) and location share among clusters (4.99 % variability), prove the minor impact of the household location on the electricity consumption behavior within Germany. Based on the location dependency check, ultimately, the combined analysis of the two datasets reveals the temporal development of the residential consumption behaviors. It shows that new technologies, especially Photovoltaics (PV), Electric Vehicles (EV) and heat pumps, have influence on the user behavior shift and the energy consumption level.

Zusammenfassung

Mit dem Übergang zu einem nachhaltigeren Energiesystem entsteht der Bedarf nach einem besseren Verständnis der zeitlichen Entwicklung des Verbraucherverhaltens von Konsumenten und Prosumenten. Trotz der zunehmenden Verbreitung von intelligenten Zählern, ist die Verfügbarkeit von Daten zum Energienutzungsverhalten aufgrund von Datenschutz- und Sicherheitsbedenken begrenzt. Daher ist das Verknüpfen und Vergleichen von verfügbaren Datensätzen der Schlüssel zur Beobachtung von Änderungen im Nutzerverhalten und steigert den Nutzen vorhandener Daten.

In der vorliegenden Arbeit wird ein neuartiger Arbeitsablauf für die kombinierte Analyse mehrerer Smart-Meter-Datensätze vorgeschlagen, der Datensätze mit unterschiedlichem Umfang, zeitlichem Ursprung und Spezifikationen miteinander verbindet. Im Wesentlichen werden 4 Schritte durchgeführt: Datenvorverarbeitung, Clusterbildung, Überprüfung der Standortabhängigkeit und Verknüpfung der Datensätze. Zunächst werden bei der Datenvorverarbeitung die meteorologischen Jahreszeiten in Kombination mit Wochentagen und Wochenenden für die Datensegmentierung ausgewählt, gefolgt von der Überprüfung fehlender Werte und der Normalisierung auf der Grundlage des maximalen und minimalen Verbrauchswerts der Haushalte. Anschließend wird der K-Means-Clusteralgorithmus, der mit einem Silhouette-Koeffizienten von 0,8186 und einem Davies-Bouldin-Index von 0,2884 unter den Fuzzy-C-Means- und hierarchischen Clustering-Ansätzen hervorsticht, angewandt, um die Haushalte nach ihrem Konsumverhalten zu gruppieren. Die beiden Validierungsansätze zur Standortabhängigkeit, eine Korrelation der Clusterzentren unterschiedlicher Standorte von 0,8048 und die Standortverteilung innerhalb der Cluster mit durchschnittlich 4,99 % Variabilität, belegen anschließend den geringen Einfluss des Haushaltsstandortes auf das Stromverbrauchsverhalten innerhalb Deutschlands. Nach der Überprüfung der Standortabhängigkeit zeigt die kombinierte Analyse der beiden Datensätze schließlich die zeitliche Entwicklung des Verbrauchsverhaltens der Haushalte auf, insbesondere dass neue Technologien wie Photovoltaik (PV), Elektrofahrzeuge (EV) und Wärmepumpen, Einfluss auf die Veränderung des Nutzerverhaltens und das Energieverbrauchsniveau haben.

Contents

List of Symbols and Abbreviations

Abbreviations	
CBT	Customer Behaviour Trials
CoSSMic	Collaborating Smart Solar-powered Micro-grids
DBI	Davies-Bouldin Index
DSM	demand side management
EU	European Union
EV	electrical vehicle
FCM	Fuzzy C-Means
IDEAL	Intelligent Domestic Energy Advice Loop
LCL	Low Carbon London
PV	photovoltaic
RES	Renewable Energy Sources
SIL	Silhouette Coefficient
SM	smart meter
AMI	advanced metering infrastructure
SSE	Sum of Squared Error
T&D	transportation and distribution
Symbols	
$\boldsymbol{c}$	clustering center
C	set of feature vectors
$d(p, q)$	distance between two points or vectors
i, j	index variables
k/K	index/number for clusters
n/N	index/number for households
$O()$	computational time complexity
Q	objective function
q	fuzziness parameter
r	Pearson correlation coefficient
R_i, j	similarity between clusters i and j
t/T	index/number for iterations
$\boldsymbol{y}$	feature vector of household
$\tilde{y}$	normalized value
μ	fuzzy membership degree

List of Figures

List of Tables

1. Introduction

In this chapter, first the motivation and background of the thesis are presented. Thereafter, the aim of the thesis is outlined by presenting main research questions. Subsequently, the contributions of the thesis are listed. In the end, the structure of the thesis is presented.

1.1. Motivation

The residential sector is one of the major contributors to energy demand, in particular electricity consumption. In Germany, the residential sector is accountable for about 26 % of the end electricity use in 2020 [1]. Rooftop photovoltaic (PV) systems for residential prosumers (consumers that generate energy locally), are considered an important technology on the pathway towards a more sustainable and decarbonised energy supply, to reach the carbon dioxide emission goals and reduce the impact of the ongoing climate change[2]. It is estimated that 680 TWh solar electricity could be generated annually in the European Union (EU) by rooftop PV systems, which is equal to 24,4 % of the end electricity consumption of the EU in 2016 [3]. Thus, transferred to the German share of the electricity end use, this potential is theoretically almost sufficient to satisfy the electricity consumption of residential households and enable a more sustainable and decentralised energy supply. However, there are several challenges to overcome to realise this potential. One important aspect is the temporal difference between PV energy generation and household consumption. Also, the shift from consumers towards prosumers, as more active stakeholders of the energy system, calls for a better understanding of this transformation, to enable a fact based basis for policy making and decision making among the energy system stakeholders. Therefore, understanding the temporal development of the consumption behavior of residential households, is crucial.

Unsupervised machine learning methods like clustering, have proven to be a suitable technique to find and analyse groups of customers with a similar consumption behavior, when analysing electricity consumption data[4]. Thus, in the past decades a variety of clustering and cluster validation methods have been developed [5]. Usually, these methods use the electricity consumption data as inputs for the algorithms. An effective way to collect the electricity consumption data is the advanced metering infrastructure (AMI), where the consumption data is collected and shared automatically. The penetration of smart meters throughout the EU is expected to increase from 44 % to 71 % from 2018 to 2023 [6]. With the increasing penetration of smart meters, the quality and quantity of electricity consumption has improved. Although, the amount of energy consumption data is increasing, the data availability for research and analysis purposes is still limited due to privacy and security concerns [7]. Therefore, it appears interesting, to combine existing datasets with different specifications. For example, the comparison between datasets of different temporal origins can support the understanding of how the shift of consumers towards prosumers impacts the energy system. This can lead to recommendations for customers to conserve energy and support decision making for investments in PV and storage systems. Furthermore, electricity suppliers could support with the grid design and have the possibility to reduce the peak load with methods like dynamic pricing and demand side management (DSM).

1.2. Aim of the Thesis

The central aim of this thesis is to identify how the ongoing technical shift influences the residential energy system with a machine learning approach. In order to solve the issue of limited data availability and increase the utility of available datasets, this thesis targets to sort out the ideal pipeline settings for analyzing datasets with different temporal origins and specifications.

Specifically, for the sake of noise removal, this thesis proposes to figure out the most suitable group of preprocessing operations. Furthermore, in order to promote the clustering quality, our research intends to investigate the perfect clustering algorithms for smart meter data. Moreover, targeting at ensuring the dataset combination is reasonable, this thesis is designed to validate the location dependency within Germany and further explore the most fitting validation methods.

Tackling the main themes leads to the following sub-questions:

- Whether new energy technologies affect the residential user behaviors?

- How can different residential energy consumption datasets be analyzed?

- How the different data preprocessing techniques affect the quality of the clustering results?
- Which clustering approach is the best fitting for massive smart meter datasets?
- Which significance household locations within Germany have on the consumption behaviour of residential energy consumers?

1.3. Contributions

The key contributions of this thesis can be summarized in the following:

- Change of the consumption behavior with new technologies:

  With the combined visual analysis of the residential electricity consumption data, the impact of new technologies on the electricity behavior patterns is proven. And the different user behaviors between consumers and prosumers are confirmed by comparing in an intuitive way with representative consumption profiles.

- Expanding the utility of available data by the combined analysis of smart meter datasets:

  A workflow for a combined analysis of smart meter datasets with different temporal origins and specifications is proposed, which consists of data preprocessing, clustering, location dependency check and dataset linking. It is designed to support to draw a more complete picture of the development of the energy system.

- Selection on clustering techniques for smart meter data analysis:

  Based on the two cluster validation scores, Silhouette Coefficient (SIL) and Davies-Bouldin Index (DBI), K-Means algorithm is selected among existing clustering approaches for our massive smart meter dataset.

- Validation of the location dependency within Germany:

  Cluster center correlation and the variety of location share among clusters are proposed to validate the location dependency. Both methods indicate a low location dependency of the electricity consumption in German households.

- Analyzing the impact of normalization techniques on the cluster analysis:

> Two ways of smart meter data normalization are evaluated. The smart meter data normalized with the maximum and minimum value of the whole period show better clustering validation scores but the clustering membership is more dependent on the over all consumption level than normalized with the maximum and minimum value of each day.

1.4. Structure of the Thesis

The thesis is structured as follows: In chapter 2 the related literature is presented and reviewed. Thereafter, in chapter 3 an overview on the available datasets is given and the dataset selection as well as the structure of the two selected datasets are described. Chapter 4, explains the analysis methods used for the thesis in detail. In the following chapter 5, the performed analysis is presented and the analysis results are discussed. The last chapter 6 summarises the motivation, results and conclusions of the thesis and gives an outlook on the possibilities of further research.

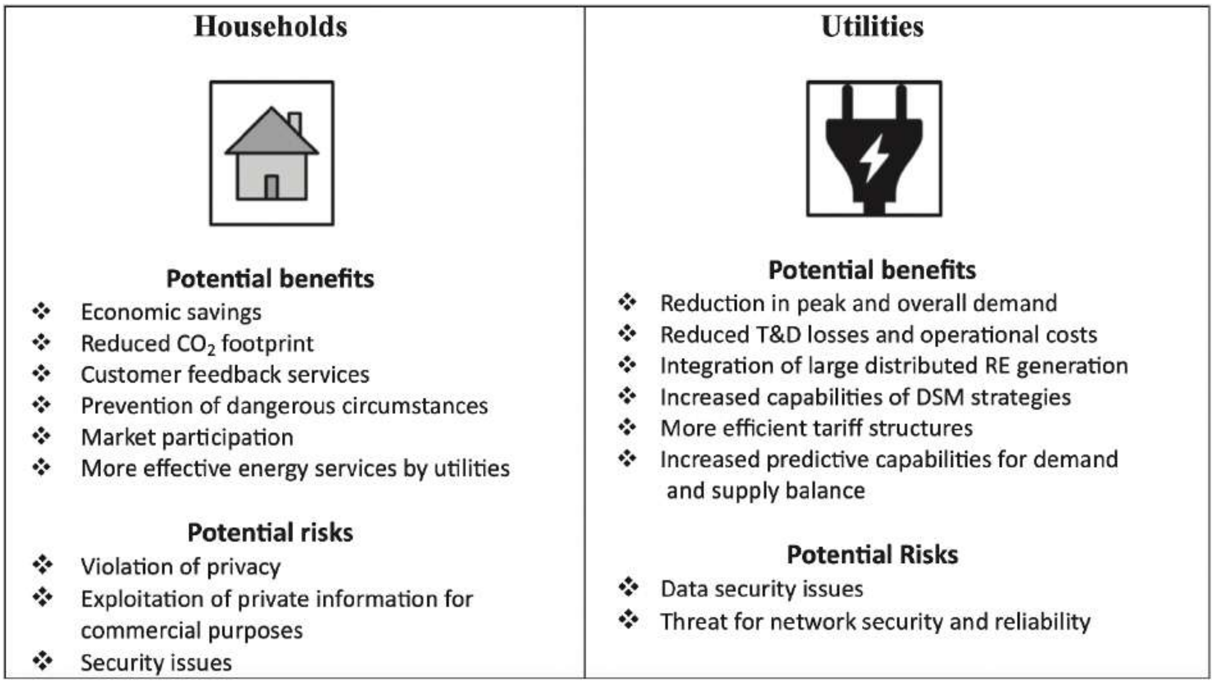
2. Literature Review

In this chapter the related literature is presented. First, the technology of SM for data collection is treated. Thereafter, the literature on residential electricity consumption analysis is discussed.

2.1. Smart Meter

Smart meters (SMs) are energy meters, which in contrast to conventional meters are able to exchange information with appliances, other meters and grid stakeholders. The meters are capable to measure the consumption of individual appliances and even control them. [8] With the enabled two-way communication, between the meter and the supplier, the term advanced metering infrastructure (AMI) is used. Regarding the frequency, the collected data is becoming finer with the development of the technology. Nowadays, SMs can collect the consumption data with different time steps ranging from daily measurements to detailed resolutions finer than one minute [9]. SMs are available for different fields of application such as electricity, gas, water or heating. In the following the term smart meter is used for electricity meters even though, the methods for analyzing are expected to be transferable to other sectors, to better understand the consumption behavior [10].

The penetration of SMs is increasing in many regions of the world. Until the end of 2016 there were 700 million SMs installed worldwide, with 350 million installed in China only [11]. The EU invested nearly 5 billion euros in smart meter related projects from 2002 to 2017 [12] and the share of SMs in the EU is expected to rise up to 71 % in 2023 [6]. In the United States over 94 million SM have been installed up to 2019, 83 million of those in the residential sector [13]. Despite the increasing worldwide penetration of SMs, the data availability for research and analysis purpose is still limited due to privacy and security concerns, like for example the abuse of the data by criminals or for governmental surveillance[4]. Figure 2.1 from [4] illustrates the concerns as potential risks in comparison with the prevailing possible benefits that are enabled by using and analysing SM data. For residential households feedback services can reduce the energy demand as shown in [14], a meta review which compared 118 studies



Households

Potential benefits

- Economic savings
- Reduced CO_2 footprint
- Customer feedback services
- Prevention of dangerous circumstances
- Market participation
- More effective energy services by utilities

Potential risks

- Violation of privacy
- Exploitation of private information for commercial purposes
- Security issues

Utilities

Potential benefits

- Reduction in peak and overall demand
- Reduced T&D losses and operational costs
- Integration of large distributed RE generation
- Increased capabilities of DSM strategies
- More efficient tariff structures
- Increased predictive capabilities for demand and supply balance

Potential Risks

- Data security issues
- Threat for network security and reliability

Figure 2.1.: Potential benefits and risks of using smart meter data adopted from [4]

related to feedback on the electricity consumption with the conclusion that this feedback can reduce a households energy consumption from 5 % up to 10 %. In [15] this results was questioned with an average energy saving of 4.5 %, reported among 12 surveyed studies.

For the energy utilities it enables demand side management (DSM), i.e., managing the peak load of their customers with techniques such as load shifting, energy conservation, valley filling and flexible loads [16]. Still, the possible concerns are justified and prevent that load datasets are published without extensive preprocessing like anonymization.

The development and diffusion of the smart meter technology is accompanied by increasing research and analysis on the consumption behavior, treated in the following section.

2.2. Electricity Consumption Analysis

In this section the relevant literature on analysis of electricity consumption is reviewed, with a focus on the residential sector. First, the research on the determinants of residential energy consumption is treated. Then, the relevant literature regarding cluster analysis in the residential energy sector is presented.

2.2.1. Determinants of Residential Electricity Consumption

The research to elaborate the determinants for residential energy demand is challenging as it depends on many influencing factors. In [4] it was pointed out that even if the same dataset and survey data is used, the attributes with significant impact on the consumption change between different studies. For example, in [17] and [18] the same dataset [19] is used with different clustering methods and aggregation of the data, resulting in different important attributes. Therefore, different determinants have been found to have an significant impact on the residential electricity demand in studies carried out in several countries, such as weather conditions, floor area, disposable household income, household location, head of household gender, dwelling type, household appliances and electricity prices [20][21][22][23]. In the following research on the household location and the impact of new technologies like PV, EV and heat pumps are presented in detail.

2.2.1.1. Location

Research on the location dependency of the energy consumption behavior within a country as an influencing factor of the consumption behavior is a field that has barely been discussed in the literature. Nevertheless, in [24] the authors examine the location dependency of different determinants, on the household energy consumption in the Netherlands. The study shows that some of the determinants like for example income and household size have a different impact on the consumption behavior in different locations within the Netherlands. For instance, the determinant household size even an opposite impact on the household energy consumption in different regions. The research on the location dependency of the consumption behavior can also support linking datasets from different locations within the same country to each other. With this connection of the datasets recommendations might be made in the future when only the household characteristics from survey data or only the demand characteristics from smart meter data are known.

In [21] the authors examine the influence of different determinants on the residential electricity consumption in the United States (US): The Result is that locality in form of a Zip Code is one of the determinants with the highest correlation to electricity consumption and explains 46 % of the variability of the residential electricity consumption. It is important to mention that the Zip code is connected to many other variables like climate conditions, building type and socio-economic factors and that the high penetration of air conditioning systems in the US results in a high impact of climate conditions on the electricity consumption.

2.2.1.2. Prosumers with PV

The impact of decentralized PV generation of prosumers on the residential electricity consumption is subject of debate in the current research. In [25], a study from Sydney, Australia, the authors conclude that households with PV installed have a significantly higher electricity consumption due to rebound effect eroding 21 % of the carbon mitigation. On the other hand, in [26] a negative rebound effect, of distributed energy resources including PV, led to increased energy savings.

In [27] is described, how prosumers are becoming more active stakeholders in the energy supply chain. They tend to develop new behaviors, have a good knowledge about their system and show higher interest in energy interventions like feedback on their consumption [28] or DSM. Moreover, the related literature highlights that the transition of the energy system towards prosumers with PV systems, as more active stakeholders of the energy system, together with new technologies like EVs and heat pumps, are changing the electricity consumption behaviour and therefore are an important factor to consider, when analysing residential electricity consumption behavior[7][4].

2.2.2. Load Profiling

The term load profiling refers to the classification of electricity consumers according to their consumption behavior [7]. There are different ways how these load profiles can be beneficial for consumers and energy supply stakeholder. Standard load profiles for example are representative load profiles composed from the average of consumers with a similar consumption behavior and can be used by energy suppliers for network planning with load forecasting and marketing balancing mechanisms like demand response programs [29]. For the consumers load profiles from smart meter data are helpful to understand their consumption behavior, compare it to other consumers and support decision making to reduce the energy demand.

Since the electricity consumption of residential customers is volatile among different days, households and dwelling types, a grouping into multiple load profiles is often sufficient to determine the consumption behaviour of a household. The unsupervised machine learning techniques of clustering, that discover patterns from the individual load profiles, are suitable for this grouping task and have been widely used in the related Literature. [10] Therefore, several clustering techniques used for load consumption analysis will be presented in the following section.

2.2.3. Cluster Analysis

A cluster analysis describes the task of grouping data into meaningful clusters. For the cluster analysis first the data is preprocessed, then clustered with an clustering algorithm and finally validated, with a suitable validation measure. The relevant literature for this steps is presented in the following sections.

2.2.3.1. Data Preparation

There are several possibilities of preparing the data for the clustering algorithms. The form of the input data is depending on the used dataset and the aim of the analysis. When dealing with energy consumption time series, that show different levels of absolute consumption, normalization is a common technique to set the focus on the consumption patterns. In [30] several data preprocessing methods have been tested by clustering electricity consumption data with the K-Means algorithm, with the result that data preprocessing methods like principal component analysis or wavelet transformation showed no improvement of the clustering results compared to a simple normalization of the data.

There are different ways to normalize the data. The data can be normalized using the daily peak consumption value as shown in [31] and [32] or by the consumption peak value of the household in the dataset, as performed in [18] and [33]. The used normalization technique can impact the results of an analysis. Still, the technique used for normalization is often insufficient discussed in the publications on load data analysis. For example in [34] normalization of the residential electricity consumption data was conducted, without describing how the normalization was performed.

2.2.3.2. Clustering

Clustering is considered as one of the most famous unsupervised machine learning technique for pattern recognition, with multiple areas of application like bioinformatics, image analysis, social science, text analysis or managing energy resources [35].

A distinction can be made between hard and soft clustering techniques. In the former, each data object belongs to one cluster exclusively, while in the latter, the data object can be assigned to multiple clusters with different probabilities. For soft clustering the term fuzzy clustering is used. [36] According to the authors of [7], clustering algorithms can be further categorized into direct and indirect clustering approaches. For direct clustering the original time series is clustered so every value represents a data point. Indirect clustering aims to reduce the dimensionality of the data through feature extraction.

Several clustering methods have been applied in the sector of energy consumption analysis. In [37] the authors compared 11 different clustering algorithms, regarding their suitability for residential load consumption clustering on daily profiles, resulting in the best performance of centroid-based and hierarchical clustering methods. Despite centroid based methods showing a good performance, the promising Fuzzy C-means clustering algorithm was not part of the comparison.

In [38] the K-means algorithm is used for clustering residential load data for customer baseline load estimation and demand response management. The authors of [39] use the K-means algorithm for clustering hourly smart meter data to extract representative standard load profiles. Fuzzy C-means clustering is used in [40] on daily load profiles to examine the influence of a household´s lifestyle on the electricity demand. In [31] the authors use the Fuzzy C-means algorithm, to analyze the effect of price signals on the consumption behavior of residential electricity customers.

In [41] hierarchical clustering is performed on daily smart meter data and combined with a door to door survey to define power consumption profiles of residential households. In [42] hierarchical clustering is used for clustering the energy consumption data of university buildings.

2.2.3.3. Cluster Validation

Since the true underlying structure of the data is unknown when using unsupervised learning, no natural quantification of the discrepancy between the model and the truth exists [43]. Therefore, validation criteria that evaluate the clustering results only with the parameters of the resulting clusters have been developed. This kind of evaluation is called internal evaluation and most of this internal validation indices are based on the cluster geometry. They often define a clustering result of high quality with the terms “compact” (the distance of elements within one cluster) and “distinct” (the distance of the clustering centers) [37]. As stated in [44] a high number of different validation indices that aim to evaluate the quality of clusters exist and the choice of index matters for the result of the clustering validation. Two common internal validation indices, that have proven their performance in the literature, are the Silhouette index (SIL) [18] and the Davies-Bouldin index (DBI) [38] [45].

A method that is used when the number of clusters is to be determined for the algorithm is the Elbow Method. It is a visual method that calculates the sum of squared errors within the clusters. It is often used in combination with the widely used K-Means algorithm, to estimate the number of clusters as parameter for the algorithm[46].

Further, multiple similarity and dissimilarity measures have been defined for time series. In [47] Pearson correlation coefficient is used as a similarity measure. According to [36] the Pearson correlation coefficient is a suitable method to measure the agreement of shapes between two patterns.

3. Data

In the beginning of this chapter an overview of the datasets available for the thesis is given. Thereafter, the selection of the datasets used for this thesis is described and at the end of the chapter the structure and specifications of the two selected datasets are presented in detail.

3.1. Data Availability

The privacy and security-related concerns, presented in section 2.1, that arise when analyzing smart meter data result in a limited availability of datasets available to the public. Nevertheless, several household level load datasets have been anonymized or semi-anonymized and are available for the present work [7]. The available datasets, together with a short description and information on the number of participants and trial duration, are presented in the following list:

- Ausgrid Resident [48]:

  Consumption and rooftop PV generation data from 300 Australian residents. Trial duration: 2010/7 - 2013/6; sample frequency: every 30 min.

- Customer Behaviour Trials (CBT)[19]:

  Irish load consumption datasets of the Commission for Energy Regulation (CER) of over 6000 households including survey data with demographics and socio-economic information. Trial duration: 2009/9 - 2011/1; sample frequency: every 30 min.

- Collaborating Smart Solar-powered Micro-grids (CoSSMic)[49]:

  Smart meter data from the EU project "Collaborating Smart Solar-powered Micro-grids" (CoSSMic). Consumption data from 12 German residential, industrial and public energy consumers. Availability of rooftop PV and appliance level data. Published by "Open Power System Data" [50]. Trial duration: 2013/5 - 2017/8; sample frequency: every minute.

- Intelligent Domestic Energy Advice Loop (IDEAL) [51]:

  Recently published dataset from two projects in the Edinburgh region. High frequency consumption data of 255 households of which 39 additionally have detailed appliance level data traced. Trial duration: 2014/12 - 2018/6; sample frequency: every second.

- Intelliekon [52]:

  Consumption data of different locations in Germany and Austria of over 2000 households. Survey data on soci-economic and household characteristics available. Available for the present thesis, not available for the public. Trial duration: 2009/5 - 2010/11, sample frequency: every hour.

- Low Carbon London (LCL)[53]:

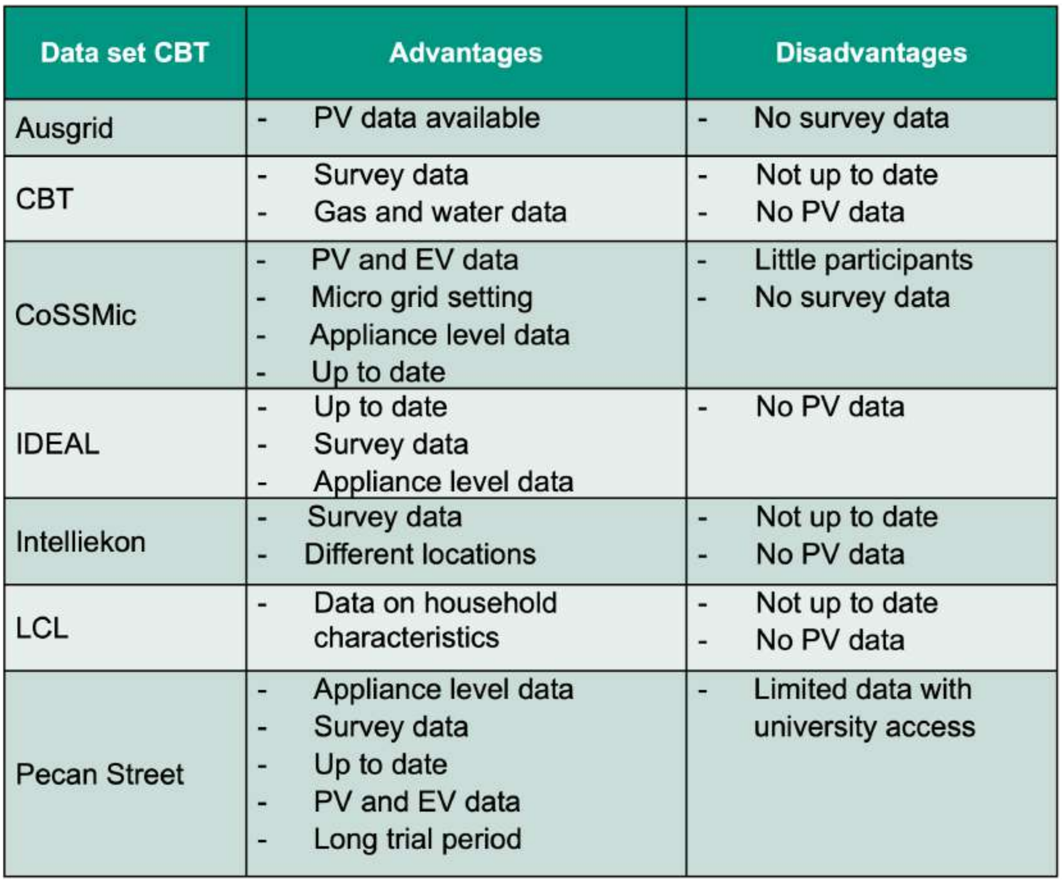
  London area load consumption and related pricing data. Similar to CBT with over 5000 households and survey data on socio-economic factors and appliance usage. Trial duration: 2013/1 - 2013/12; sample frequency: every 30 min.

- Pecan Street [54]:

  Detailed dataset from different locations in the USA of 500 households with a high frequency of up to 1 min sampling rate. Household level and appliance level consumption data available. EV and PV data available for some households. Only a part of the data data can be recived for research purpose as a university member. Trial duration: 2005/5 - Today; sample frequency: every minute.

Further, figure 3.1 provides an overview over the advantages and disadvantages of the datasets. The sample frequency is of subordinate importance for the present work but may be interesting for other As the data availability is limited, a combined analysis of datasets is a possible way to meet this limitation. Therefore, according to [4] a standardization of the trial design supports a inter comparison between different studies. Nevertheless, in the present work it is investigated if different study designs and characteristics can also complement each other.

In terms of the geographic aspect, the datasets are mostly collected in a specific area or the household locations are unavailable due to anonymization purposes. Exceptions are the Intelliekon and Pecan Street dataset where the data is collected from several different regions within a country. This aspect was one important factor for the dataset selection, which is explained in detail the next section.



Data set CBT	Advantages	Disadvantages
Ausgrid	- PV data available	- No survey data
CBT	- Survey data - Gas and water data	- Not up to date - No PV data
CoSSMic	- PV and EV data - Micro grid setting - Appliance level data - Up to date	- Little participants - No survey data
IDEAL	- Up to date - Survey data - Appliance level data	- No PV data
Intelliekon	- Survey data - Different locations	- Not up to date - No PV data
LCL	- Data on household characteristics	- Not up to date - No PV data
Pecan Street	- Appliance level data - Survey data - Up to date - PV and EV data - Long trial period	- Limited data with university access

Figure 3.1.: Advantages and disadvantages of the datasets

3.2. Dataset Selection

For the selection of the datasets several different aspects have been considered. First of all, the dataset has to be available. This appears to be obvious but some of the datasets listed in the previous section are available only under restrictions like a university access or other conditions must be met. Thus, the process of dataset selection is not only to find and select the best suitable datasets, but also investigating what is necessary to get access. Access conditions include request forms or personal contact with the person responsible for the dataset. As an overview, table 3.1 lists the access types for the different datasets.

The thesis aims to discover the opportunities to increase the utility of the available data, by a combined analysis of different datasets. While the selected datasets should contain different information for this combined analysis, some aspects should be similar to enable a reasonable linking and get informative results. Since for the present work, the temporal

Dataset	Access type
Ausgrid	Open access
CBT	Request form
CoSSMic	Open access
IDEAL	Contact via mail (Now, with the official release of the dataset: Open Access)
Intelliekon	Not published, access at Fraunhofer ISI inhouse database
LCL	Open access
Pecan Street	Sign up at "dataport"[55] with prove of institution for limited university access

Table 3.1.: Type of access for the datasets

development of the residential energy system was of considerable interest, it was decided to select datasets with temporal instead of geographical differences for the combination. Therefore, the geographical origin of the participating households is the second important aspect for the dataset selection. Accordingly, the datasets should be of the same country, in order to reduce country specific differences that affect the consumption behavior, like for example legislation and traditions. This led to the decision between the LCL and IDEAL household datasets, both collected in the UK and the Intelliekon and CoSSMic dataset of German households.

The third aspect impacting dataset selection are the data specifications. Here the UK datasets have the advantage of more trial participants, while the Intelliekon dataset offers more detailed survey data and the CoSSMic dataset, in contrast to the UK datasets, offers PV generation data. Because of the interesting possibility to compare two datasets with and without PV energy of different temporal origins, at the end the German datasets were chosen for the analysis and are presented in detail in the next section.

3.3. Data Structure

In this section the two selected datasets are presented. Therefore, general information and important aspects about the trials, where the datasets were collected, as well as the structure of the datasets are discussed.

3.3.1. Intelliekon Dataset

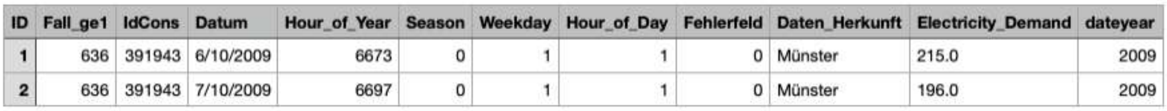
The Intelliekon dataset was created within a field study in Germany and Austria focusing on the effect of consumption information feedback for residential households. The households were covered with a AMI and some of the households received feedback on their electricity consumption while a randomly selected control group did not receive such feedback. The trial was carried out in cooperation with electricity utilities in eight German municipalities, namely Celle, Hassfurt, Kaiserslautern, Krefeld, Münster, Oelde, Schwerte and Ulm, from five different federal German states and one Austrian utility in Linz. The distribution of the households among the German locations is displayed in table 3.2.

The trial duration varies among the different locations. Nevertheless, from November of 2009 to November of 2010 the data collection was active for all of the households. Furthermore, the households were requested to take part on surveys regarding socio-economic factors and household characteristics before and after the trial. [52] In the present work only the data from the 600 German households was used, since the dataset combination and the location dependency of the consumption data within one country is to be explored.

Location	Number of households
Linz	1624
Münster	128
Kaiserslautern	126
Hassfurt	101
Krefeld	90
Ulm	62
Schwerte	61
Celle	49
Oelde	13

Table 3.2.: Number of participating households for the different locations of the Intelliekon trial

The Intelliekon dataset is composed out of multiple tables. The most relevant table contains the active power consumption data of the households. In figure 3.2 a screenshot of the first two rows of over 16 million rows in total of this table are displayed. All rows of this table represent the electricity consumption value of one hour for a specific household. Several identification numbers, temporal information such as date, hour of the day and year as well as the local origin of the data ("Daten_Herkunft") and the hourly electricity demand for this hour, are the columns of this table. The survey data is available in two tables for each survey, with one describing the questions asked and the other one with the quantified response from the households.



ID	Fall_ge1	IdCons	Datum	Hour_of_Year	Season	Weekday	Hour_of_Day	Fehlerfeld	Daten_Herkunft	Electricity_Demand	dateyear
1	636	391943	6/10/2009	6673	0	1	1	0	Münster	215.0	2009
2	636	391943	7/10/2009	6697	0	1	1	0	Münster	196.0	2009

Figure 3.2.: Example of consumption data from the Intelliekon dataset

The Intelliekon dataset is not available to the public and could be accessed for the present thesis because of the cooperation with the "Fraunhofer Institute for Systems and Innovation Research ISI", which was a co-director of the Intelliekon trial.

3.3.2. CoSSmic Dataset

The CoSSmic dataset was collected within the scope of the EU project "Collaborating Smart Solar-powered Micro-grids". It is composed out of 12 participants representing a neighbourhood with six residential, four industrial and two public (a school and a swimming pool) electricity consumers [56]. The trial collected the baseline consumption behavior of the electricity customers and simulated the automatic load shifting approach within the solar powered micro grid afterwards. The participants of the trial are located in the suburban area of Konstanz, a city in southern Germany. In the following the focus will be on the residential electricity consumers, as they are to be compared to the smart meter data from the Intelliekon dataset.

The collected data includes several parts. The total imported and exported electricity data of the households as well as appliance level consumption data, from appliances considered important for the consumption or suitable for automatic load shifting, such as fridge, freezer, dish washer, washing machine, heat pump and EV, was collected. Additionally, the electricity generation of the rooftop PV system was traced when available. Four of the six households where equipped with a PV system. The traced appliances for the different households are listed in table 3.3. [49]

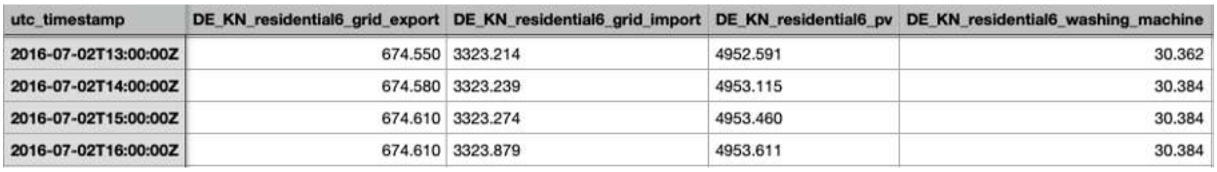
Originally the data is collected with a sampling frequency of 1 min. Nevertheless, the dataset is also published with a sampling frequency of 15 and 60 min of which the 60 min table was used in the present work to align the frequencies of the two datasets. Structure wise, every column of the table represents either the consumption of one of the appliances, the electricity import, export or the PV generation of one household. Figure 3.3 shows an example

Resident 1	Resident 2	Resident 3	Resident 4	Resident 5	Resident 6
10 kWp PV system Electrical heat pump Washing machine Freezer Dishwasher	No PV System Circulation pump Freezer Washing machine Dishwasher	5 kWp PV system Circulation pump Dishwasher Washing machine Dishwasher Refrigerator Freezer	10 kWp PV system EV Heat pump Washing machine Refrigerator Freezer	no PV System Washing machine Dishwasher Refrigerator	4 kWp + 5 kWp PV system Circulation pump Washing machine Dishwasher Freezer

Table 3.3.: Traced appliances of the residential households in the CoSSMic trial

screenshot of "Resident 6" with the UTC timestamp as index and the electricity export, import, PV generation and consumption of the washing machine as columns. As visible, in contrast to the Intellikon dataset, where the columns are the hourly values for one household, the table shows the meter readings and therefore, the values are given in absolute numbers.

The data from the trial is made available to the public by the project "Open Power System", a platform, promoting the open availability of data required for research on energy systems [50].



utc_timestamp	DE_KN_residential6_grid_export	DE_KN_residential6_grid_import	DE_KN_residential6_pv	DE_KN_residential6_washing_machine
2016-07-02T13:00:00Z	674.550	3323.214	4952.591	30.362
2016-07-02T14:00:00Z	674.580	3323.239	4953.115	30.384
2016-07-02T15:00:00Z	674.610	3323.274	4953.460	30.384
2016-07-02T16:00:00Z	674.610	3323.879	4953.611	30.384

Figure 3.3.: Example of consumption data from the CoSSMic dataset

4. Methodology

In this chapter the methods used in this thesis are presented and explained. Starting with the data preprocessing, the representation of the data and the normalization is described. Thereupon, the used clustering algorithms are described. At the end of the chapter the used clustering validation indices, for comparing clustering results of different algorithms or determining the number of clusters for algorithms are presented.

If a metric is necessary, in the present work, the euclidean distance defined as

$$d(q, p) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{4.1}$$

is used. Where p and q are points or vectors in the n-dimensional euclidean space $\mathbb{R}^n$.

4.1. Data Preprocessing

For the data preprocessing first the representation of the hourly consumption data by representative daily consumption profiles is described. Then the two normalization techniques used in this thesis are explained.

4.1.1. Data Representation

For an intuitive way of interpretation the electricity consumption data is transformed to representative daily load profiles. This representative profiles also represent the input and output values of the clustering algorithms. Therefore, each household is represented by a feature vector with the average of all hourly consumption values for each hour of the day as features. This can be formulated as:

$$\boldsymbol{y_n} = \frac{1}{D} * \sum_{i=1}^{D} \boldsymbol{y_{n,i}} \tag{4.2}$$

where $\boldsymbol{y_n}$ is the feature vector of the household n = (1, 2, ..., N), with N representing the total number of households used for the analysis. The feature vector contains 24 entries for each hour of the representative day $\boldsymbol{y_n} = (y_{n,1}, y_{n,2}, ..., y_{n,24})$. Further, $\boldsymbol{y_{n,d}}$ denotes the vector with one day of hourly consumption data and $d = (1, 2, ...D)$ represents the number of days D were the hourly consumption data is available for the specific household n.

4.1.2. Normalization

In this section, the two normalization techniques presented in section 2.2.3.1 are described. The first technique uses the maximum and minimum values of the household of the considered time period for the normalization. In contrast, the second technique normalizes the values with the daily maximum and minimum value. Both techniques, result in values between 0 and 1. The former sets the focus on the shape of the daily consumption patterns while the latter is retaining the information about the daily peak values [4]. The two approaches of normalization can be formulated as:

$$\tilde{y}_{t,n,1} = \frac{y_{t,n} - y_{n,min}}{y_{n,max} - y_{n,min}} \tag{4.3}$$

$$\tilde{y}_{t,n,2} = \frac{y_{t,n} - y_{n,dmin}}{y_{n,dmax} - y_{n,dmin}} \tag{4.4}$$

where $y_{t,n}$ is the hourly consumption value of the household $n = (1, 2, \dots, N)$ of the lag $t = (1, 2, \dots, T)$ where T indicates the number of hourly lags of each household. $\tilde{y}_{t,n,1}$ and $\tilde{y}_{t,n,2}$ denote the normalized hourly value calculated with the first and second approach. $y_{n,max}$ and $y_{n,min}$ represent the maximum and minimum consumption value of this household in the dataset, while $y_{n,dmax}$ and $y_{n,dmin}$ are the daily maximum and minimum consumption values.

4.2. Clustering algorithms

To find patterns in the load consumption behavior of residential electricity consumers, several clustering methods have shown to be useful in the past decades. In a structured literature review Tureczek and Nielsen [57] expelled the centroid based clustering techniques such as the K-means clustering algorithm, and derived algorithms like the fuzzy K-Means algorithm, as well as hierarchical clustering algorithms as the most popular clustering algorithms for smart meter data analysis. In this section the mentioned clustering algorithms are described, and their advantages and disadvantages are compared.

4.2.1. Centroid-based Clustering

Centroid based or partitional clustering algorithms, partition N observations to k clusters by iteratively assigning each observation to the closest centroid. The centroids are defined as the mean values of all households belonging to the cluster:

$$\boldsymbol{c_k} = \frac{1}{|C_k|} \sum_{i=1}^{n} \boldsymbol{y_i} \tag{4.5}$$

where $k = (1, 2, ..., K)$ denotes the cluster number and $\boldsymbol{c_k}$ describes the cluster center of cluster k. C_k is the set of feature vectors $\boldsymbol{y_i}$ of the households with the cluster membership of cluster k and $|C_k|$ denotes the cardinality of the set, in this case the number of feature vectors that belong to the set.

The centroid based algorithms aim to solve a expectation-maximisation problem that converge when the centroids do not change, or the change is smaller than a defined threshold. A drawback of this algorithms is that the number of clusters is to be determined a priori. In the following the K-Means algorithm and the Fuzzy C-Means algorithm are presented and discussed.

4.2.1.1. K-Means

This today widely used clustering algorithm was first mentioned in [58]. According to [59] it tends to minimize the within sum-of squares objective function of:

$$Q_K = \sum_{n=1}^{N} \sum_{k=1}^{K} I(\boldsymbol{y_n} \in C_k)(\boldsymbol{y_n} - c_k)(\boldsymbol{y_n} - \boldsymbol{c_k})^\top \tag{4.6}$$

where $I(\boldsymbol{y_n} \in C_k)$ is a binary variable that equals to 1 if the example $\boldsymbol{y_n} \in C_k$ and 0 otherwise. The algorithm can be described by the following steps:

1. Initialization: A random selection of k examples serve as initial the centroid seeds.

2. Clustering: For each iteration $t = (1, 2, ..., T)$ with T being the total number of iterations of the algorithm. Each of the examples is assigned to the closest cluster center $\boldsymbol{c_k}$.

3. Centroids update: Recalculation of the clustering centers $\boldsymbol{c_k}$ by calculating the mean of all assigned examples.

4. Termination: Termination of the algorithm, if either the defined number of maximum iterations is reached or the improvement of Q_K is lower than a pre-defined threshold ε with $Q_K(t) - Q_K(t+1) \leq \varepsilon$.

The algorithm has the drawback, that convergence in an global optimum, respectively global minimum of the sum of squares function, is not guaranteed, which is why the algorithm is run several times and the results with the best O_K is selected as the result. Advantages of the algorithm are the over all robustness and simplicity[7]. To improve the initialization process, in the present work the improved K-Means++ algorithm was implemented. It uses a heuristic to improve the selection of the centroid seeds and archives better time cost and better solutions than the original algorithm [60].

4.2.1.2. Fuzzy C-Means

The FCM algorithm is similar to the K-Means algorithm with the difference that an example can be assigned to more than one cluster, with the membership degree defined for the intervall $[0,1]$. Similar to K-Means the algorithm is based on the minimization of the following objective function

$$Q_q = \sum_{n=1}^{N} \sum_{k=1}^{K} \mu_{n,k}^{q} (\boldsymbol{y_n} - c_k)(\boldsymbol{y_n} - \boldsymbol{c_k})^\top \tag{4.7}$$

where $\mu_{n,k}$ is the cluster membership degree of household n regarding cluster k and q denotes the fuzziness parameter with $q \in (1, \infty)$. Thereby $q = 1$ equals hard clustering and high values for q indicate a high membership distribution for the examples. The main advantage of the algorithm is that it provides further information on the membership degree of the examples, enabling a better representation of the reality and supporting a finer interpretation of the clustering results[61].

4.2.2. Hierarchical Clustering

For hierarchical clustering algorithms the dataset is divided into sequences of nested partitions, instead of a single partition like the centroid based clustering techniques. Hierarchical clustering can be performed bottom-up starting with each object as a single cluster also called agglomerative hierarchical clustering or top-down, namely, divisive hierarchical clustering.

[36] The linking of the clusters is performed with different linkage methods described in the following:

- Ward: Minimizes the SSE within the clusters, similar to the K-Means approach combined with an agglomerative hierarchical approach.
- Complete linkage: Defines the maximum distance between examples of a pair of clusters as the cluster distance.
- Single linkage: The minimum distance between examples of a pair of clusters is defined as the cluster distance.
- Average linkage: The average distance between the examples of a pair of clusters is defined as the cluster distance.

An advantage of this algorithms is the descriptive interpretability with dendrograms that represent the clustering in a tree-like structure. A Drawback of the algorithm is that the clustering results are sensitive for the choice of similarity measures for the linkage of the clusters and that the data objects cannot be reallocated if grouped wrong in an early stage of the clustering process [36].

4.3. Cluster Validation

To evaluate the clustering results and chosen a suitable number of clusters in certain clustering algorithm, such as centroid based methods, several clustering validation measures have been developed. In the field of unsupervised learning, where the patterns in the data are not known before the analysis, cluster validation is considered as one of the most difficult issue in the clustering process [33]. A distinction is made between internal and external validation approaches. For internal validation the properties of the found clusters itself is examined, most internal indices describe the compactness within and distinction among the clusters. While, for external validation a part of the data with ground truth labels is used as a reference partition to verify the clustering results. According to [5] internal cluster validation indices can be used to compare clustering techniques among each other or evaluate clustering results when possible outliers are excluded or the number of clusters is changed. In the following several validation indices and similarity measures that are used for the analysis are presented.

4.3.1. Elbow Method

The Elbow Method is a commonly used visual method for determining the number of clusters for the K-Means algorithm [62]. For the method the sum of squared errors (SSE) is calculated for each example to the cluster center it belongs to. The mean SSE of all examples is then compared for different number of clusters. As the number of clusters increase, the SSE will become smaller the aim of the method is to determine the "Elbow Point" with the best trade of between SSE and manageable number of clusters. According to [63], the method can be described with the following steps with k representing the number of clusters in the K-Means algorithm:

1. Initialize the initial value of k; often k = 2 is used
2. Increase the value of k in defined steps up to a reasonable limit
3. Calculate the SSE results for each value of k
4. Plot the SSE results so that the number of k increase along the x-axis and the SSE decreases
5. Visually locate the elbow-shaped k value in the plot

The SSE is a statistical method to measure the discrepancy between the data and the estimation. It can be expressed as

$$SSE = \sum_{i=n}^{N_k} (d)^2 \tag{4.8}$$

where d is the distance between the example and the cluster center it is assigned to. The distance is calculated with a chosen distance measure.

4.3.2. Silhouette Coefficient

The SIL is a common way to combine cohesion within and separation among the clusters in a single measure [64]. It was first introduced by Peter J. Rousseeuw in 1987 [65]. The steps to calculate the SIL are calculated as defined in [66]:

1. For every example, the average distance $a(i)$ to all examples in the same cluster is computed:

$$a(i) = \frac{1}{|C_a|} \sum_{j \in C_a, i \neq j} d(i, j) \tag{4.9}$$

2. For every example, the minimum average distance between the example and all examples in each cluster not containing the analyzed example is calculated:

$$b(i) = \min_{C_b \neq C_a} \frac{1}{|C_b|} \sum_{j \in C_b} d(i, j) \tag{4.10}$$

3. For each example, the SIL is calculated by the following expression:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{4.11}$$

4. If necessary the global SIL is calculated as the average SIL of all examples:

$$SIL = \frac{1}{N} \sum_{i=1}^{N} s(i) \tag{4.12}$$

The SIL is defined in the interval $[-1, 1]$. The coefficient has the advantage to provide a simple framework for the qualification: positive values indicate a high separation between the cluster, negative values are an indication that the clusters are mixed with each other and values around zero indicate a uniformly distribution throughout the Euclidean space[67]. A drawback of the coefficient is the high computational time complexity of $O(dN^2$, making it difficult to scale to large datasets[68].

4.3.3. Davies-Bouldin Index

The DBI was first introduced in 1979 [69]. It defines the average similarity between each cluster C_i and its most similar one C_j, with $i, j = (1, 2, ..., K)$, the similarity is defined as:

$$R_{i,j} = \frac{\hat{d}(C_i) + \hat{d}(C_j)}{d(c_i, c_j)} \tag{4.13}$$

where $\hat{d}(C_i)$ is the average distance between each example of cluster i and the corresponding cluster center. Then the DBI is defined as:

$$DBI = \frac{1}{K} \sum_{i=1}^{K} \max_{i \neq j} R_{i,j} \tag{4.14}$$

The DBI has a lower computational time complexity of $O(d(K^2 + N))$ compared to the SIL and low values correspond to a better cluster configuration with 0 as the lowest possible value [68].

4.3.4. Pearson Correlation Coefficient

The Pearson correlation coefficient is a measure of linear correlation between two sets of data. It is defined as the ratio between the covariance of two variables and the product of their standart deviations. The Pearson correlation coefficient between two samples can be calculated as:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{4.15}$$

where n is the sample size, x_i, y_i denote the individual sample points and $\overline{x}, \overline{y}$ are the sample mean values. The Pearson correlation coefficient can be interpreted as a normalized covariance. Advantages of the coefficient is the interpretability of the degree of relationship [36]: positive values indicate the tendency that high values of the one example are associated to high values of the compared examples. Negative values indicate a reversed relation with high values of one example, leading to low values of the compared example. Values around 0 indicate no linear relationship among the examples. A drawback of the coefficient is that it is sensitive to outliers [70].

4.4. Software

In the present thesis the programming language Python version 3.8.5 [71] was used together with Jupyter Notebook version 6.1.4 [72] as the computational environment for the programming. For most of the cluster analysis the scikit-learn package version 0.24.2 was used[73].

5. Evaluation and Discussion

In this chapter, the workflow for the combined analysis is presented together with the results and discussion of the processing steps.In general, there are 4 steps performed: data preprocessing, clustering, location dependency check and dataset linking. Firstly, the data preprocessing including data segmentation, missing value investigation and normalization is treated. Thereafter, the cluster analysis is presented with the comparison of three clustering algorithms. Subsequently, two approaches on the location dependency of the clustering results are described and the results are discussed. Then the data processing of the CoSSMic data for the combined analysis is explained. Ultimately, the results from the combined analysis of the two data sets are discussed. An overview of the workflow is visualized in figure 5.1.

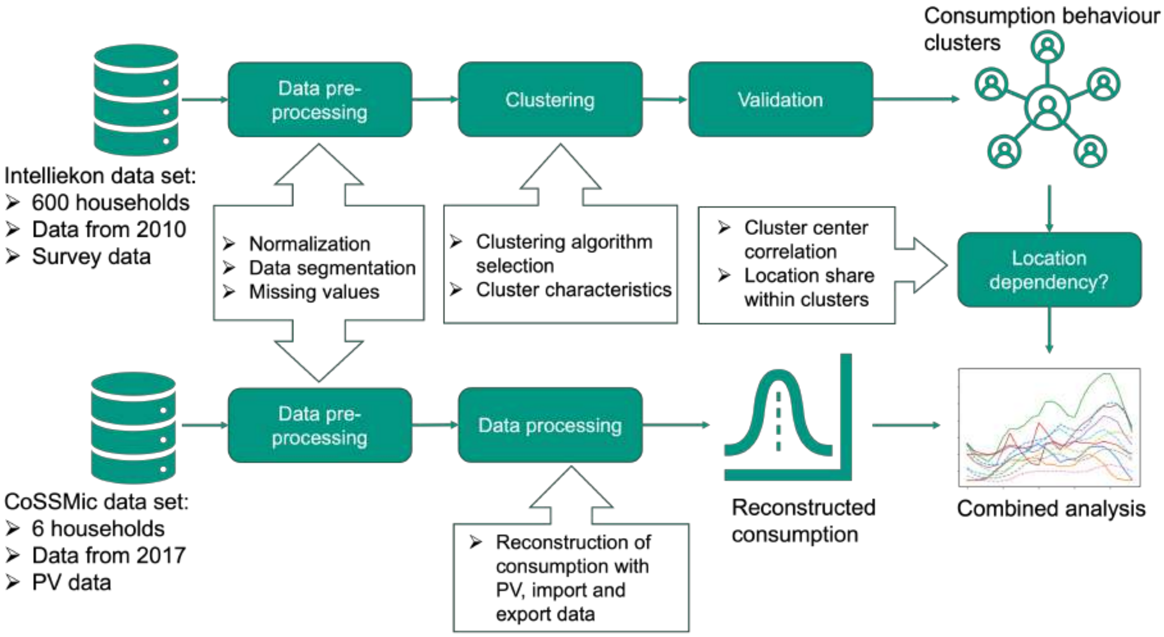


Figure 5.1.: Workflow of the data set combination

Type of season	Spring	Summer	Fall	Winter
Seasons from the literature	21st of March – 14th of May	15th of May – 14th of September	15th of September – 31st of October	1st of November – 20th of March
Meteorological seasons	1st of March – 31st of May	1st of June – 31st of August	1st of September – 30th of November	1st of December – 28th of February
Astronomical seasons	20th of March – 20th of June	21st of June - 14th of September	15th of September - 31st of October	21st of December – 19th of March

Table 5.1.: Time periods of different season types

5.1. Data Preprocessing

To prepare the data presented in the chapter 3.3 for the analysis and clustering, several preprocessing steps were conducted, including dividing the dataset into time periods to cover seasonality and differences between weekdays and weekend days, the search for missing values and the data normalization. For the normalization, the results of two normalization techniques are presented. Since the cluster analysis was performed on the Intelliekon dataset, our generalized data preprocessing was mainly applied to this dataset. This procedure is then applied to the CoSSMic data set as far as necessary, because no cluster analysis was carried out on this data set. Since the data of the CoSSMic data set is given as meter readings, and therefore absolute numbers of energy measured by the meter so far, the data was transformed to hourly consumption values by calculating the difference between the hourly time step as additional preprocessing step.

5.1.1. Time Periods

To cover seasonality, one year of data was used for the analysis. Further, to examine the differences in the consumption behavior the one year period data was divided into several time periods. Thereby, the challenge was to observe the differences of the time periods, while obtaining a clear and manageable amount of results. Therefore, possible time periods in terms of season and weekday were defined and the profiles of the mean hourly electricity consumption values of all households of the different time periods were compared for their differences.

5.1.1.1. Seasonality

The seasons that are compared to each other are the meteorological seasons, the astronomical seasons and seasons used in the related literature in [39] adopted from [74]. Table 5.1 shows the different starting and end dates for the different season types.

To compare the different type of seasons, representative load shapes were created by calculating the hourly mean normalized consumption of all households for each season. Thus, four representative load shapes representing the four seasons for each season type were calculated. The resulting representative load profiles for the different type of seasons are shown in figure 5.2. It is visible that the differences among the season types are rather small. To further evaluate the differences in between the season types, the mean Pearson correlation coefficient was used as similarity measure for the seasons within every season type. As the division of the data set into seasons aims to find seasons were the consumption difference is high in between the seasons, a low mean Pearson coefficient, representing diverse patterns, is desired. The similarity of the type of seasons is also expressed by the similarity of the coefficients with 0.97189 for the seasons from the literature, 0.97372 for the meteorological seasons and 0.97481 for the astronomical seasons. It is to mention that although, the Pearson correlation coefficient can be used as a similarity measure for time series, it shows the linear correlation among the seasons and not exactly how different the time series are from each other. Still, with the similar values for the season all of the seasons are reasonable choices. Therefore, the meteorological seasons were used for the analysis due to the following reasons:

- The seasons of the literature have a unequal length of the time periods with summer and winter season being longer than the transition seasons, resulting in less data for the calculated hourly mean values.

- While fitting well on the Intelliekon data set, for the CoSSMic data set two households finished the trial in February, therefore the meteorological seasons a more suitable for a combination of the data sets.

The use of the meteorological seasons combined with the availability of the data results in using the data from the 1st of December 2009 to the 30th of November 2010 for the Intelliekon data set and 1st of March 2016 to the 28th of Febuary 2017 for the CoSSMic data set. After the investigation of the seasons and time periods used for the two datasets, the differences between the days of a week are treated in the following section.

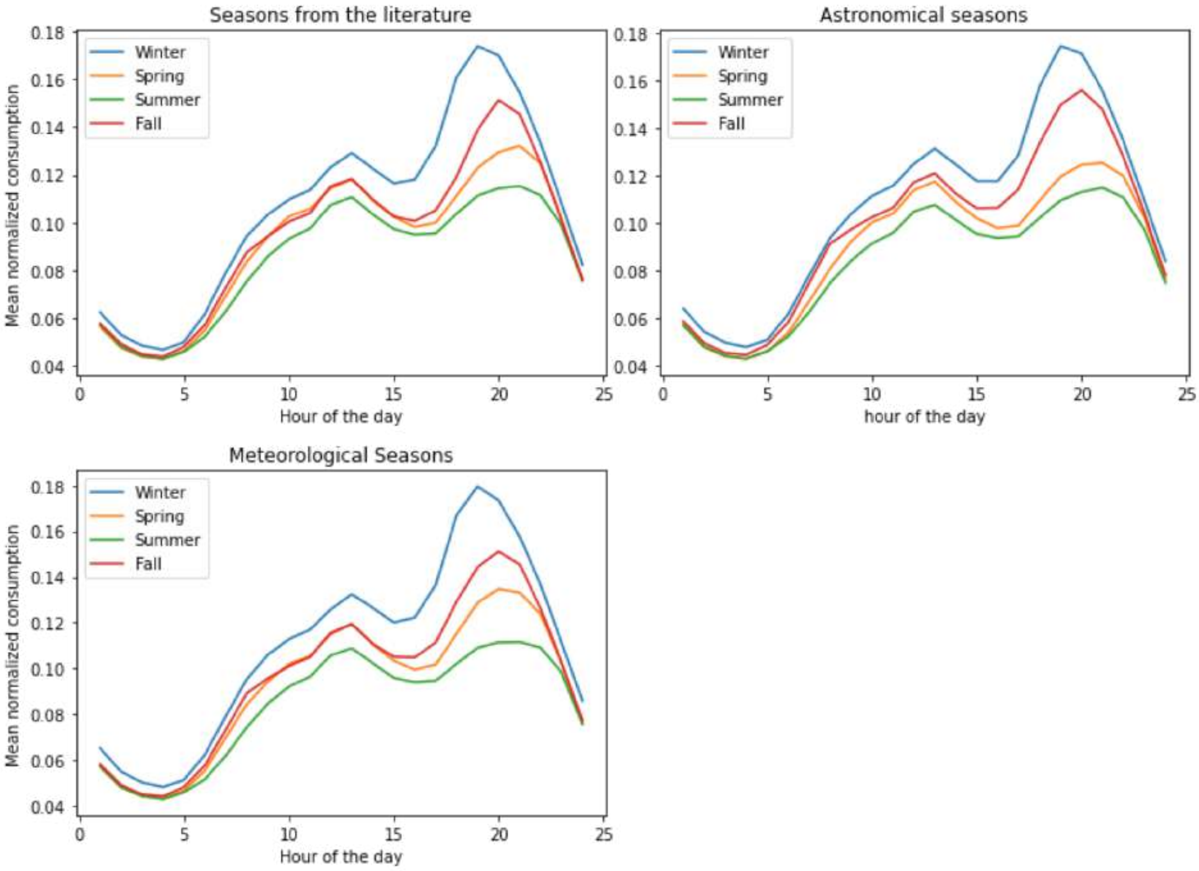


Figure 5.2.: Mean normalized consumption of the different seasons

5.1.1.2. Day of Week

To compare the different days of the week, the mean normalized consumption of all households was calculated for every day. The resulting load shapes are displayed in figure 5.3. It appears that the consumption behavior is similar during the weekdays and differs between weekdays and weekends. The distinction between Saturdays and Sundays is reasonable, as the peak consumption during the day is significantly higher for Sundays. Yet, in order to keep the number of generated results manageable and ensure enough data is available for all the time periods, Saturday and Sunday were treated together as weekend days. Therefore, in this thesis a distinction was made only between weekdays and weekends. The four seasons together with the distinction between weekdays and weekend days results in eight time periods covering one year of data. With this time periods distinguished, the next data preprocessing step was to examine the missing values for these two datasets,which is explained in the next section.

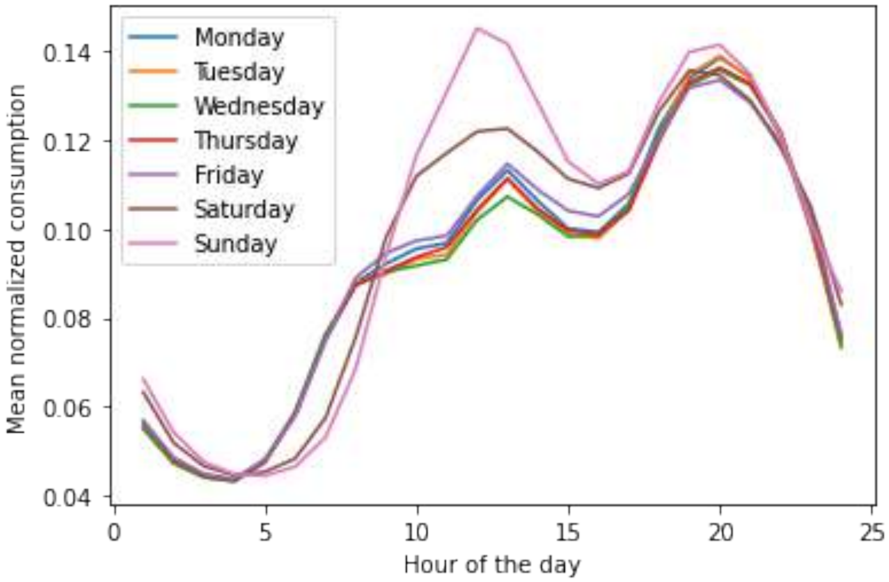


Figure 5.3.: Mean normalized consumption of the different days of a week

5.1.2. Missing Values

Missing values happen to appear in smart meter data sets because of reasons like failure of data collection, communication problems of the smart mete devices, unplanned events and maintenance [7].

In the Intelliekon data set, smart meter data is structured with the hourly consumption values in rows, as shown in section 3.3.1. All of the existing rows contain values, thus no null values appear in the data set. The missing values are therefore calculated by comparing the number of present data rows for each household ID with a complete dummy series of 8760 hourly values of the considered period of one year. This results in an overall rate of missing values of 23.66 % for the one year period. Since this rate of missing values is relatively high, it needs to be further examined.

With the technique of calculating the representative consumption profile with the mean values for every hour of the day, missing values are of minor importance if they are evenly distributed among the time periods and households. Therefore, the number of values for each hour of the day was examined. This number is equal for each hour of the day, hence, it can be concluded that only complete days are missing in the dataset, which also explains the high rate of missing values. In addition, the number of available values over the one year period was investigated. Figure 5.4 shows the distribution of available values for each day over the observed period for all households. Although, there are some fluctuations between 8000 and 12000, available values are still sufficiently equally distributed among the time periods.

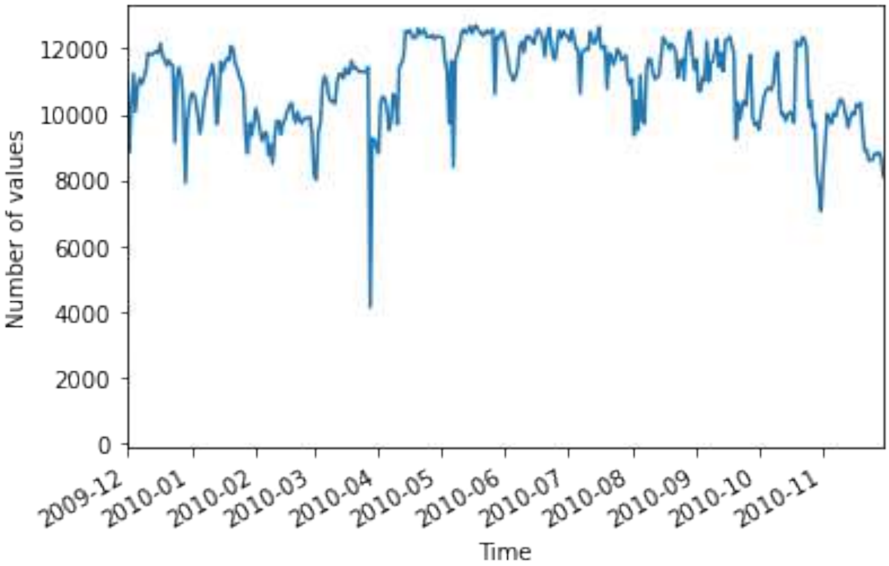


Figure 5.4.: Number of values available for each day over the observed year

The distribution of the missing values among the different households is less equally distributed. Figure 5.5 shows the rate of missing values for the German households with 0 indicating no missing values for a household and a rate of 1 a household with only missing values, calculated with the dummy variable explained earlier in this section. The households with a high rate of missing values are a problem because they can influence the clustering results with mean values as features of only a small amount of days. To exclude households with a large share of missing values, while maintaining a high number of households for the comparison of locations, the cut-of-value was set to a share of missing values of 0.7 (marked with a dashed line in 5.5). This value was selected because of the high increase of the missing values rate visible in the chart and led to 37 households being excluded from the data set. These 37 households are from two locations only, 26 households are from Kaiserslautern and 11 from Münster. The overall rate of missing values, without the 37 households and 563 households remaining, is 19.32 %.

The CoSSMic data set does not contain missing values in the data with the exception of "Resident 2" that finished the trial in February of 2017 resulting in one month of missing data. The high completeness of the data set is due to the interpolation of the missing values that was already performed on the published dataset[50].

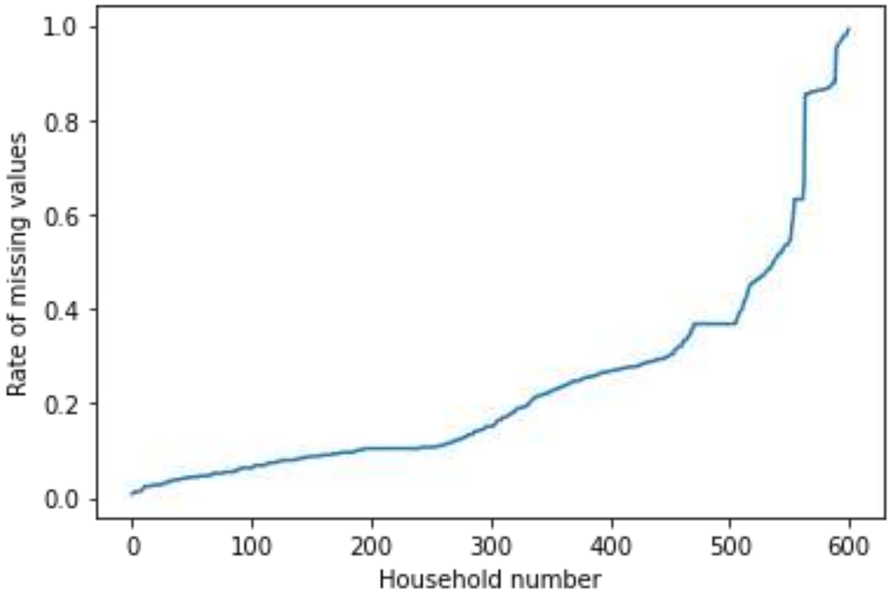


Figure 5.5.: Rate of missing values for the 600 German households

After the examination of missing values in the datasets, they have to be scaled for a better comparison between the households and as input for the clustering algorithm. Therefore, a normalization of the data was performed and explained in the following section.

5.1.3. Normalization

Normalization of the data is used to compare the consumption behavior of households with different absolute levels of consumption. In this work, the two techniques of normalization, described in section 4.1.2, are applied to the dataset. The first technique calculates the normalized value with the maximum value of the household as reference value while the second technique uses the daily maximum value of the household for the normalization. To visualize the difference between the two normalization techniques, the normalized mean values of all households for meteorological seasons are displayed in figure 5.6. It is noticeable that the y-axis, where the normalized consumption is displayed, has a different range of values due to the different values used for the normalization. Another important aspect is that the different levels of the consumption among the seasons are more remarkable in the where the normalization is performed with the maximum value of the whole data set. The differences

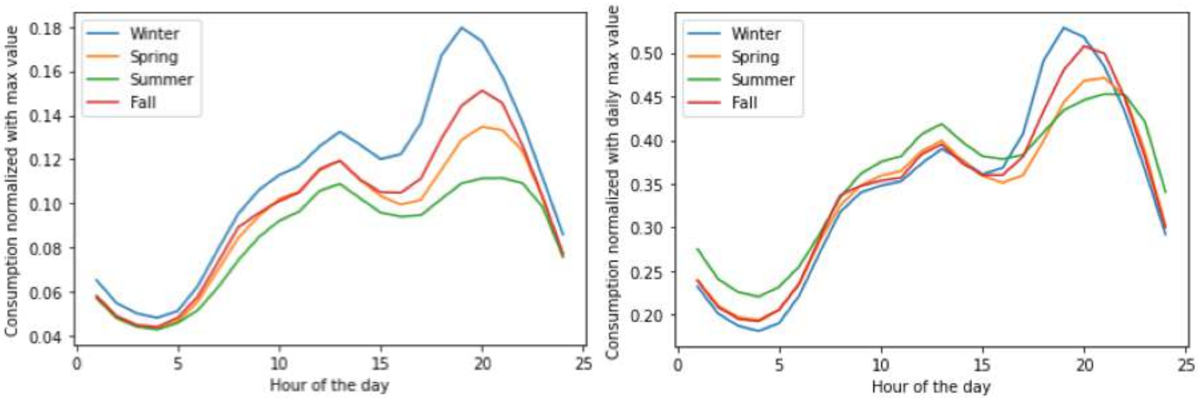


Figure 5.6.: Comparison of normalization methods

among the seasons are useful for the comparison with the PV data of the CoSSMic data set to keep the differences between consumption and PV generation between the seasons.

The impact of the normalization technique on the clustering results is discussed in the following. Therefore, the results of the clustering, conducted with the two different normalization techniques are discussed qualitative and quantitative, supported by the cluster validation criteria defined in section 4.3.

The techniques differ by the reference value used for the normalization. While the first technique uses the absolute consumption maximum and minimum value of the household for normalization, the second technique refers to the daily maximum and minimum value of the household for normalizing the data.

To visualize the differences of the clustering results, the corresponding clustering centers of the two normalization techniques are displayed in figure 5.7. For both techniques the clustering results of two example time periods are shown. Similar to the differences among the seasons, treated before in this section, the normalized consumption value range of the clustering centers differs for the two techniques. For the first technique the normalized consumption is high in the winter time period and lower in the summer time period. Thus, it retains the differences between the seasons and therefore better reflects the reality. On the other hand it is to mention, that the resulting clustering centers of the first technique do not vary as much as the clustering centers of the second technique, in terms of the time when the energy is consumed. The first technique results in more similar profiles which differ by their consumption level and the characteristics of their peaks. The results of the second technique show more differences among the clustering centers with clusters that differ

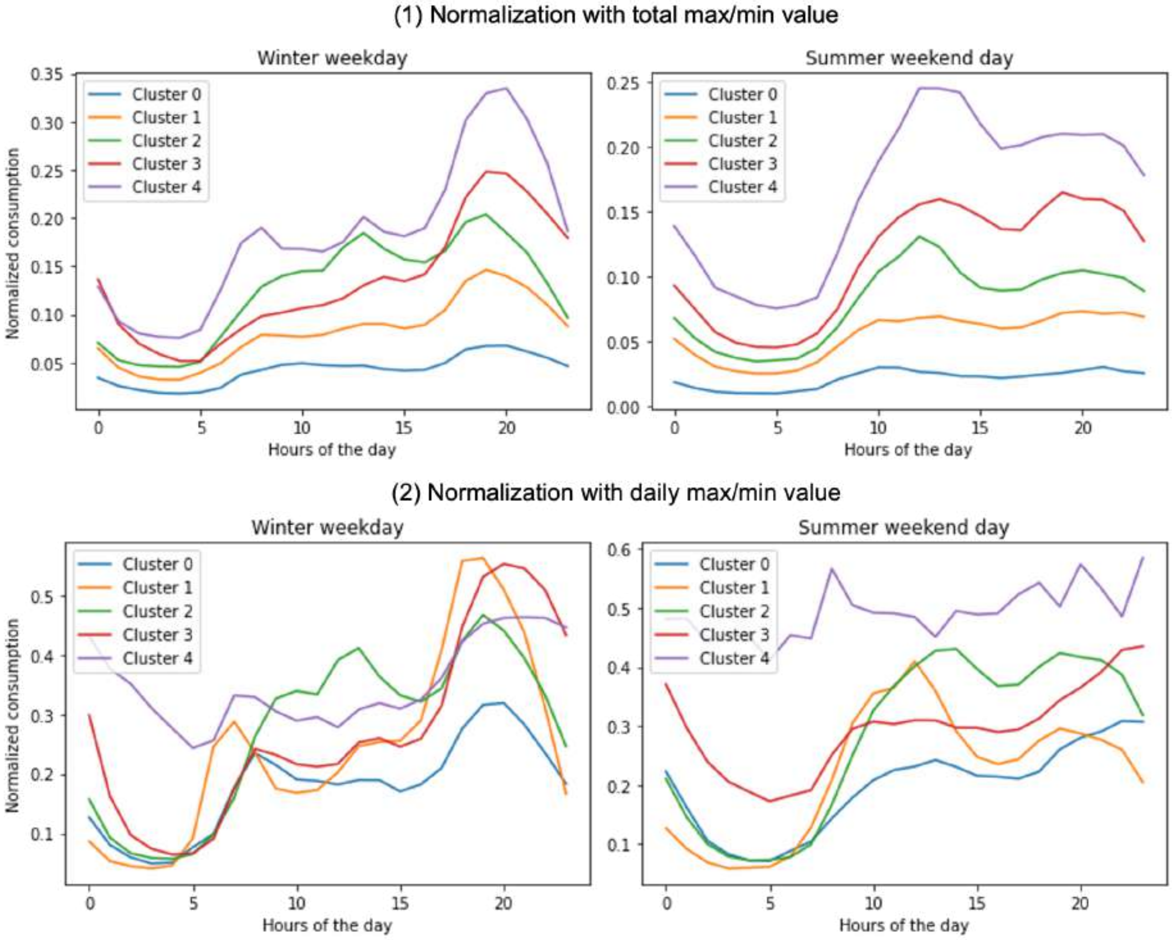


Figure 5.7.: Clustering results for different normalization techniques

from the expected profile structure with peaks during the day or at the evening. For example cluster 4 of the second technique shows a high consumption during the night for the winter weekday time period and an almost constant demand during the summer weekend day time period.

To evaluate the distribution of the households among the clusters the number of cluster members for every cluster is presented in table 5.2. For the seasons marked with (1) or (2) the first, respectively second, normalization technique is used. The number of clustering members indicate a more equal distribution of the households among the clusters for the second normalization technique with the exception of cluster 4. Cluster 4 shows a low number of cluster members for the summer time period of the second technique with only 10 households which explains the more angular shape of the cluster 4.

Cluster number	0	1	2	3	4
Winter weekday (1)	107	185	107	68	45
Winter weekday (2)	97	99	147	133	37
Summer weekend day (1)	65	162	183	114	37
Summer weekend day (2)	143	150	135	120	10

Table 5.2.: Number of households per cluster with the different normalization techniques (1) and (2) for the example time periods

In order to evaluate the clustering results with a quantitative measurement, the SIL and the DBI have been calculated as cluster validation criteria. Table 5.3 shows the validation scores for the two example time periods discussed before and the average validation scores for all time periods for the two normalization methods. It can be recognized that the first technique exhibits higher SIL and a lower DBI, both indicating a higher cluster quality. It can be concluded that the use of the different reference values for the normalization results

Season	SIL	DBI
Winter weekday (1)	0.7612	0.3842
Winter weekday (2)	0.4720	0.8697
Summer weekend day (1)	0.8040	0.3115
Summer weekend day (2)	0.4823	0.8278
All time periods (1)	0.7780	0.3572
All time periods (2)	0.4779	0.8500

Table 5.3.: Validation scores of clustering results with the different normalization techniques (1) and (2)

in different information, represented by the clustering results. While the cluster validation indices indicate a higher quality of the first approach, this must not relate to a increased gain of information on the consumption behavior. For example it is possible that the good validation values of the first normalization approach, are caused by a good separation among the total energy consumption levels. Still, because it leads to more intuitive load shapes and because the total consumption is not unimportant when including PV generation data, this first approach was used for the further analysis and the combination of the data sets.

With the normalization of the data and the other preprocessing steps conducted before, the data sets can be further analyzed. Therefore, the cluster analysis for the Intellikon data set is presented in the next section and the further data processing of the CoSSMic data set is described in the section thereafter.

5.2. Cluster Analysis

In this section the processing steps of the cluster analysis are presented. First it is presented how the validation is performed. Thereafter, it is explained how the appropriate clustering algorithm was selected, supported by the validation criteria, and what parameters have been used. At the end of the section, the results of the cluster analysis with the selected algorithm is presented and discussed.

The clustering was only performed for the Intelliekon data set, because the number of participants of the CoSSMic data set was not sufficient for applying the cluster analysis. The data processing of the CoSSMic data set is described in section 5.4.

5.2.1. Cluster Validation

The validation of the results generated by the compared algorithms was conducted with the help of internal validation criteria that use distances within and among the clusters to evaluate the quality of clusters as described in section 4.3. The internal validation of the clustering results was performed with the SIL and the DBI calculated with the scikit-learn package[73]. For similarity measures between the representative load profiles or clusters the Pearson correlation coefficient was used. Further details on the validation criteria are to be found in section 4.3.

When the number of clusters is to be determined a priori, like for the centroid based clustering method, the so called Elbow Method presented in subsection 4.3.1 can be used.

The resulting graph from this method for the data of all the German households from the Intelliekon data set for 2 to 20 clusters is presented in figure 5.8. Although, it is not possible to identify a clear elbow point it becomes clear that the average SSE is declining less steep from five clusters on, so according to the elbow method five or six clusters are a reasonable choice and are also a manageable number of clusters for load profile classification. Since the Elbow Method does not show a significant value for k, additionally the SIL for the different number of clusters was calculated. The results of this analysis are visible in figure 5.9. First of all it is to mention, that the high SIL of over 0.8 for most values of k is an indicator for well separated clusters as explained in subsection 4.3.2. Moreover, the SIL is increasing slower from five cluster on. With the combined results of the Elbow Method and the SIL the number of clusters was set to five when number of clusters was to be defined as input for an algorithm.

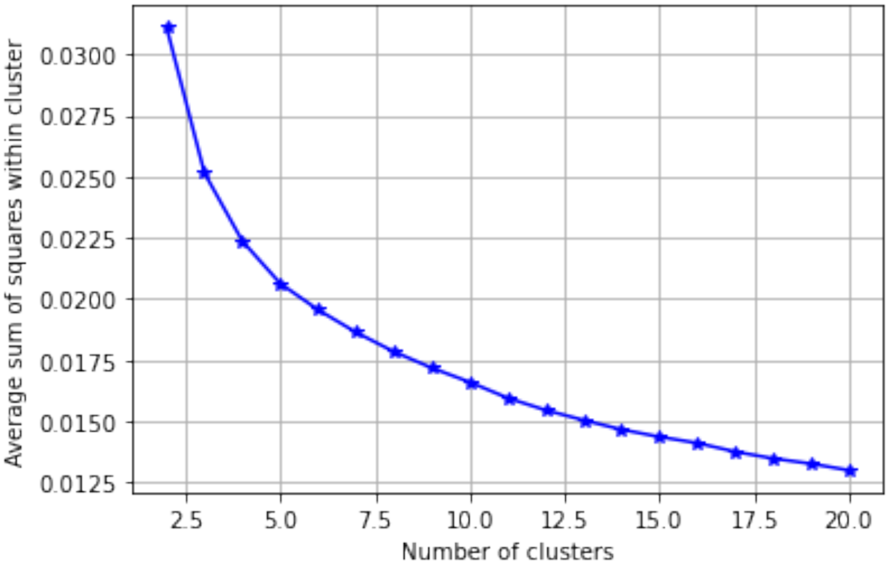


Figure 5.8.: Average SSE within each cluster for different number of clusters

5.2.2. Comparison of Clustering Algorithms

For the cluster analysis the three clustering algorithms described in section 4.2 were compared to find the appropriate clustering method for the Intelliekon data set. The comparison of the clustering algorithms was performed with a hands on approach were the three algorithms, namely hierarchical clustering, K-Means and Fuzzy C-Means, were used on the data set. Then, the best performing algorithm for the further analysis was selected by evaluating the results

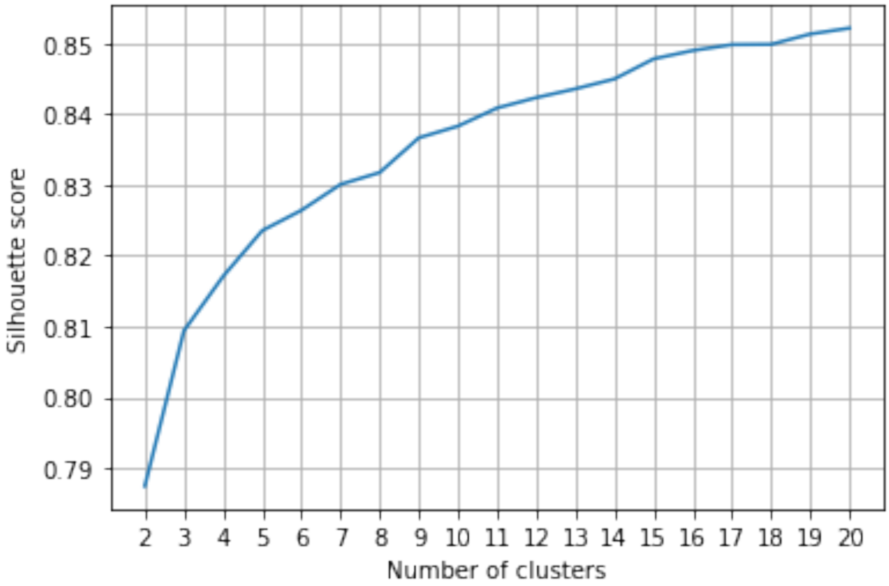


Figure 5.9.: Silhouette score for the different number of clusters

qualitatively and quantitatively with the cluster validation indices. As described in section 4.1.1, the features for the clustering algorithms are the mean of the hourly values for the specific time frame. To compare the algorithms, the hourly average of the whole year data, without the defined time periods, was used. In the following, the set parameters and results of the algorithms are presented and the selection of the used clustering algorithm is described.

Hierarchical Clustering For the hierarchical clustering, an agglomerative clustering approach was performed on the data set. All of the linkage methods implemented in the scikit-learn package have been tested but only the "ward" linkage method that, similar to the k-means algorithm, minimizes the sum of squared differences within all clusters produced a reasonable clustering result. Figure 5.10 shows the resulting dendrogram with the households starting as single clusters at the bottom and hierarchically merging to the top of the chart. A distance matrix with the pairwise calculated Pearson correlation coefficients between the features of the household served as input for the algorithm. Although an interesting approach the hierarchical clustering was not used for the further analysis for the following reasons:

- Despite the additional information that may emerge from this hierarchical approach, the clustering for the different time periods results in different optimal cut-off values and cluster numbers.

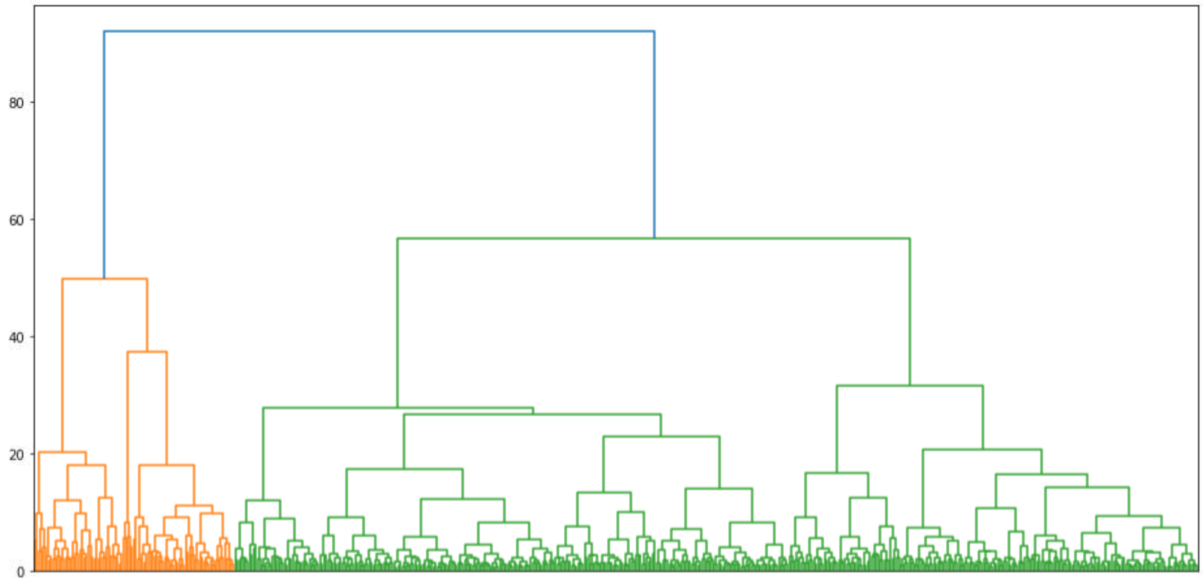


Figure 5.10.: Dendrogram for hierarchical clustering with ward-linkage

- The typical good interpretability of the results from hierarchical clustering is in this case reduced by the high number of households.

- The well known drawback of hierarchical clustering that once assigned to a cluster the households can not be reassigned.

K-Means The K-Means clustering approach is a unsupervised machine learning technique, widely used for many different applications. As shown in chapter 2 it has also proven its performance in the field of energy consumption analysis. The functionality and backgrounds are specified in section 4.2.1.1.

The number of clusters $k = 5$, as defined in section 5.2.1 was used for the K-Means algorithm. The default parameters from the scikit-learn package are 10 initializations to prevent the convergence in local optima and the stopping criteria for the algorithm were 300 maximum iterations or a tolerance of $\varepsilon = 0.0001$ for the change of the clustering centers per iteration. Since the results especially of the SIL were changing when ruining the algorithm several times, this parameters were modified to 100 initializations and 500 maximum iterations per run. With this parameters the algorithm produced more robust results.

The features that serve as input for the clustering algorithm are the representative hourly averaged consumption profiles calculated for each time period. Therefore, 24 features per

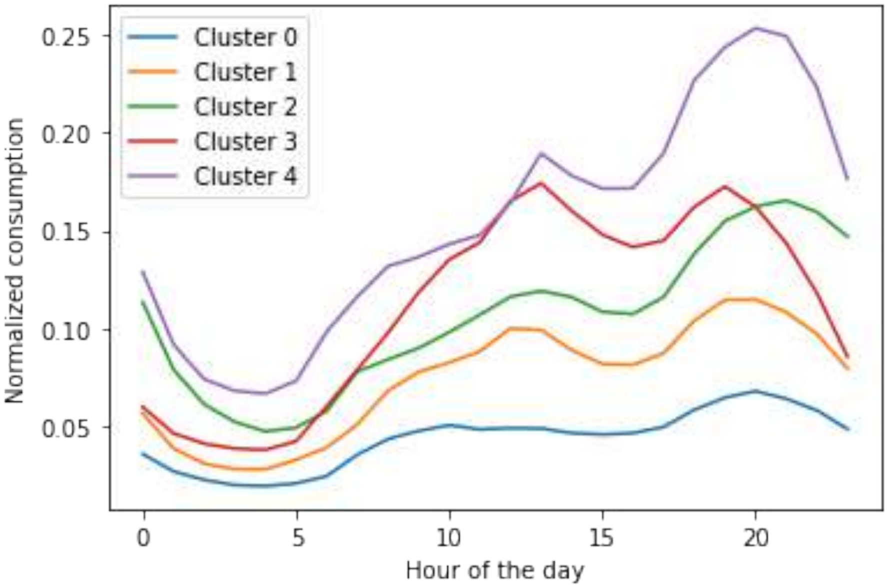


Figure 5.11.: Clustering results of the K-Means algorithm for the one year time period

time period and household are used. For the algorithm comparison the algorithm was used with the mean hourly values for the whole year period, without using the defined time periods. The resulting clustering centers are visible in figure 5.11. The clusters are ordered by the mean value of the clustering centers. To compare the results with other algorithms the global SIL and the DBI where calculated with all the examples and the corresponding cluster membership. The value for the global SIL is 0.8186 with high values up to one indicating a good cluster quality. For the DBI the value calculated is 0.2885 with low values down to zero indicating a better clustering quality. The results are used to compare the results with the Fuzzy C-Means algorithm presented in the following.

Fuzzy C-Means The Fuzzy C-Means algorithm is a soft clustering algorithm were each of the examples can belong to more than one cluster. For the implementation of the algorithm the fuzzy logic toolbox skfuzzy 0.4.2 was used [75]. Further information on this algorithm are presented in section 4.2.1.2.

An additional parameter compared to the K-Means algorithm is the fuzziness parameter q that determines the degree of fuzziness in the results. To visualize the impact of the parameter q the fuzzy partition coefficient is plotted for different possible number of clusters in figure 5.12. The fuzzy partition coefficient is defined for the range from 0 to 1. According to the

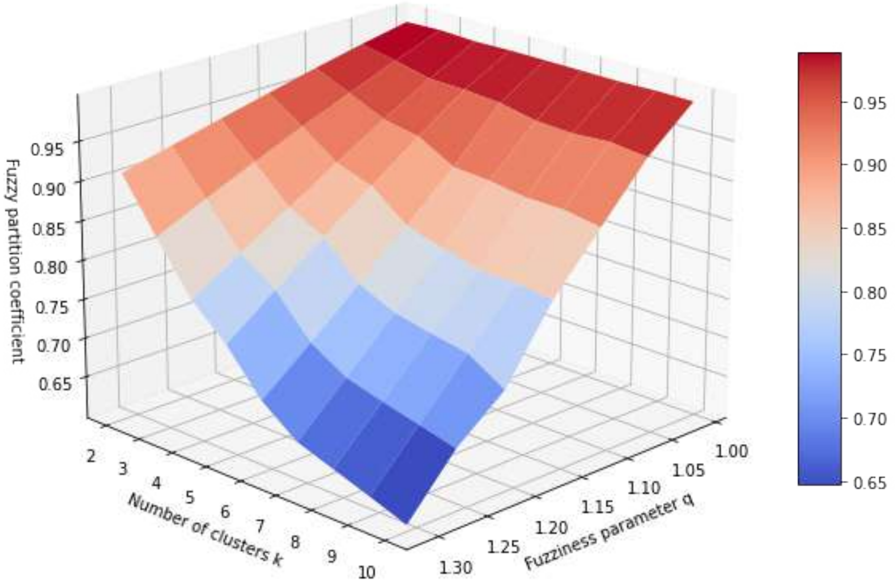


Figure 5.12.: Fuzzy partition coefficient of the Fuzzy C-Means algorithm for different number of clusters k and fuzziness parameters q

documentation of the used package the coefficient is a measure to describe how cleanly the data is described by a certain model, with 1 representing the best value. The plot with the high values for parameters around q = 1 suggests that it is rather a measurement to evaluate how much the results differ from a hard clustering solution. Still, it is worth highlighting the faster decrease of the coefficient from five clusters on, indicating k = 5 as a suitable number of clusters. To further evaluate the results the fuzzy clustering was conducted with the same number of clusters $k = 5$ as the K-Means algorithm and a fuzziness parameter of q = 1.05. The results where than defuzzified by choosing the highest cluster membership as the hard membership. In this way the clustering indices used for the K-Means algorithm could be compared on the results.

Validation Results The comparison of the validation scores for the three clustering approaches with the two validation indices is displayed in table 5.4. The higher SIL index and lower DBI indicate a better clustering performance of the K-Means algorithm.

Algorithm	SIL	DBI
K-Means	0.8186	0.2884
Fuzzy C-Means	0.1952	1.5847
Hierarchical clustering	0.1612	1.688

Table 5.4.: Comparison of cluster validation indices

It is to mention that this comparison is of limited significance, as the internal validation criteria do not offer a conclusion about the information contained in the clusters. For example, the additional information of the soft clustering is a strength of the algorithm providing this additional information that can not be expressed by the validation scores. Still, it is the best possible way to quantify the results of unsupervised learning methods and was performed to evaluate the differences among the results of the clustering algorithms.

Fuzzy C-Means clustering and hierarchical clustering provide additional information that can support the cluster analysis. Also the Fuzzy C-Means algorithm offers the possibility to include an outlier cluster, where outliers with a defined distance to all other examples can be collected. For the present work, the K-Means algorithm provides the necessary information, shows high validation scores, a low computational effort and over all high robustness and simplicity. Therefore, it was selected for the analysis of the location dependency and the linking between the data sets.

5.2.3. Clustering Results

The clustering results with the selected K-Means algorithm for the eight different time periods are displayed in figure 5.13. The results show similar clusters among the seasons. The differences between the seasons and the weekday and weekend days are clearly visible. While showing a morning, midday and evening peak for the weekdays the weekend days show no morning peak, a more distinct midday peaks and an evening peak similar to the weekdays. An exception are the summer weekdays that do not show a morning peak. A possible reason is the six weeks long summer school break. Thus, families with children that might experience the weekdays in this time more like weekend days. Because it is difficult to find out which households have children, with school breaks that influence the consumption behaviour, this exception was not further studied in the present work.

To additionally show the distribution of the households among the clusters in table 5.5 the number of the households with the cluster membership of the corresponding cluster

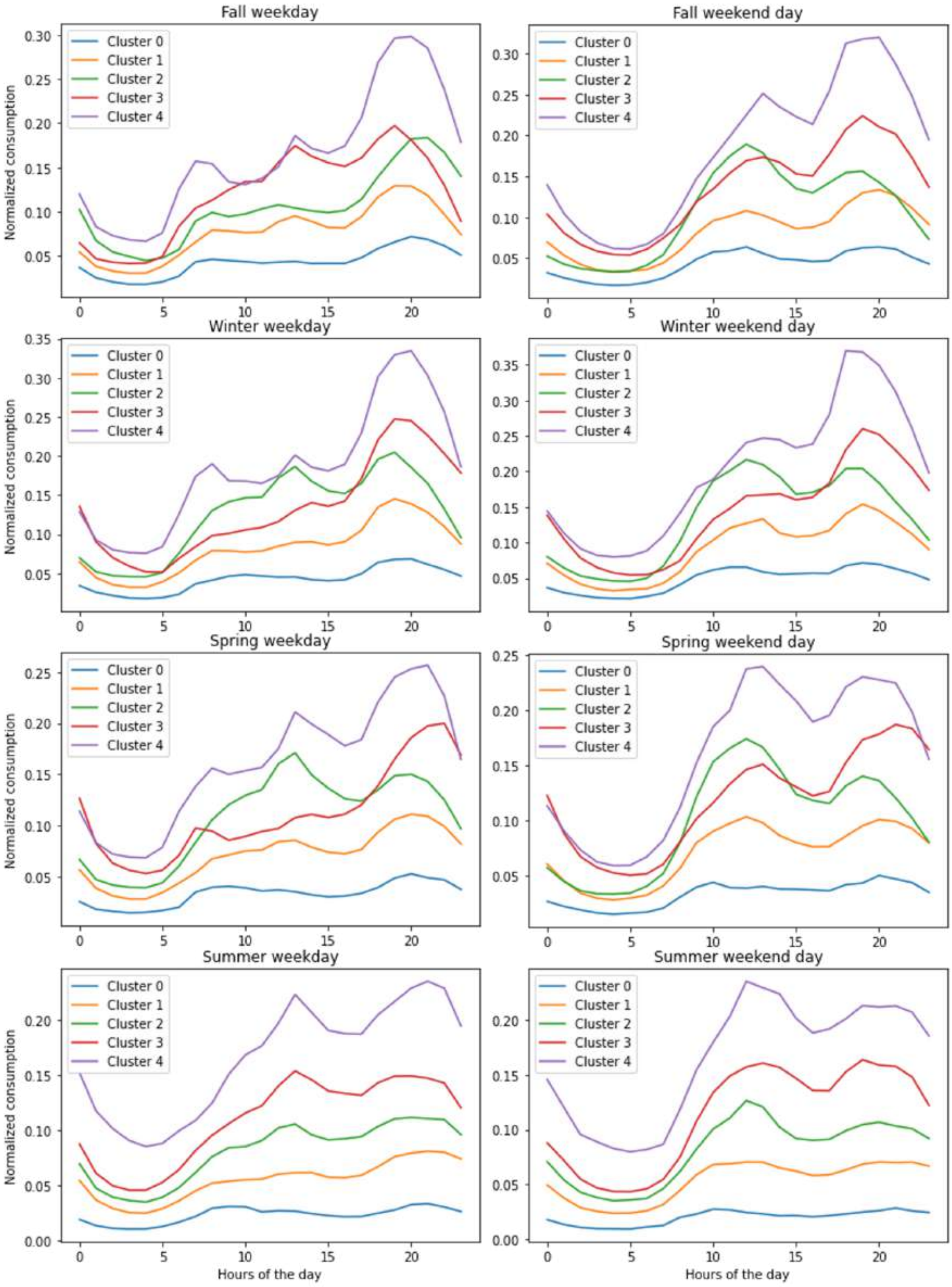


Figure 5.13.: Clustering results of K-Means algorithm for selected time periods with k=5

are displayed. It becomes clear that also the number of cluster members is similar among the seasons. For example cluster 1 includes the most households for most the time periods and cluster 4 exhibits the least cluster members among the time periods. Still the numbers of clustering members are not equal among the time periods, which means some of the households change the cluster membership among the seasons.

Cluster number	0	1	2	3	4
Fall weekday	138	163	84	86	41
Fall weekend day	120	172	92	91	37
Winter weekday	107	185	107	68	45
Winter weekend day	120	176	96	78	42
Spring weekday	105	224	120	77	36
Spring weekend day	95	205	97	98	67
Summer weekday	78	186	162	111	24
Summer weekend day	65	162	183	114	37

Table 5.5.: Number of households with cluster membership for the different time periods

Over all figure 5.13 ilustrates that the summer and spring time period distinguish from the winter and fall time period by a less distinct evening peak and a lower normalized consumption level. Still, most of the time periods have similar characteristics for the different clusters:

- Cluster 0: A cluster with an over all low normalized consumption and only slightly visible peaks for the midday and evening.
- Cluster 1: The cluster with the most cluster members, as shown in 5.5, for most of the time periods. It shows two small peaks during the day and a higher evening peak for the weekdays and two almost equally high peaks at the midday and evening for the weekend days.
- Cluster 2: Cluster 2 is characterized by the high mid day peak on the weekend days and an increasing consumption over the day with an evening peak on the weekdays.
- Cluster 3: A single distinct midday peak that towers or equals the evening peak on weekdays and a higher similarity of weekday and weekend day time periods characterize this cluster.
- Cluster 4: Cluster 4 is characterized by a distinct midday and evening peak. It exhibits a relatively high normalized consumption in the second half of the day.

To validate the clustering results the SIL and the DBI were calculated and are displayed in table 5.6. As stated in chapter 4 positive SILs indicate high clusters quality and a low DBI close to zero indicates a good separation among the clusters. Over all the results indicate a high cluster quality among the different time periods. It is to mention, that the validation scores in unsupervised learning, where the underlying structure is unclear, do not serve as an absolute quality measure. Nevertheless, they are suitable to get an impression of the clustering quality and compare the time periods among each other.

Season	SIL	DBI
Fall weekday	0.7785	0.3609
Fall weekend day	0.7660	0.3844
Winter weekday	0.7612	0.3842
Winter weekend day	0.7448	0.4079
Spring weekday	0.7823	0.3522
Spring weekend day	0.7720	0.3586
Summer weekday	0.8152	0.2975
Summer weekend day	0.8040	0.3115

Table 5.6.: Validation scores of clustering results for the different time periods

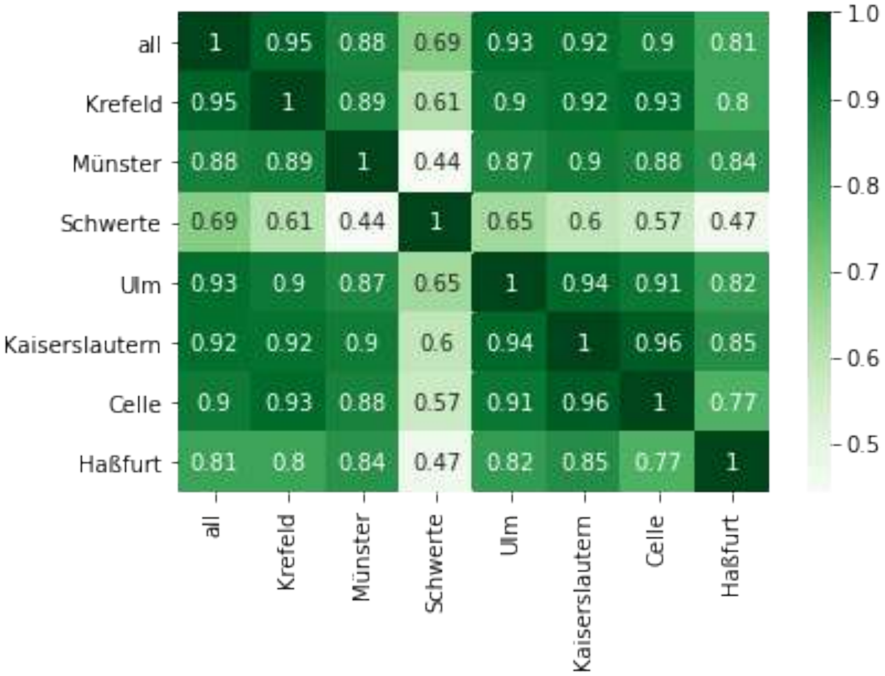
After the comparison of the clustering algorithms and the presentation of the results from the selected K-Means algorithm, in the next section the location dependency of this results is investigated to enable the combined analysis of the two data sets.

5.3. Location Dependency

To examine the influence of the location on the clustering results the data from the German Intelliekon data were used together with the information of the household location. Two different ways of analyzing the location dependency of the data have been used. The two approaches and the corresponding results and are presented in the following sections.

5.3.1. Cluster Center Correlation

The first approach was to perform the clustering analysis on the different subsets of the data set with the same household location. This resulted in eight clustering results of the different time periods for each of the seven locations. To compare the similarity of the clustering results the values of the cluster centers for each result were compared to the results of the

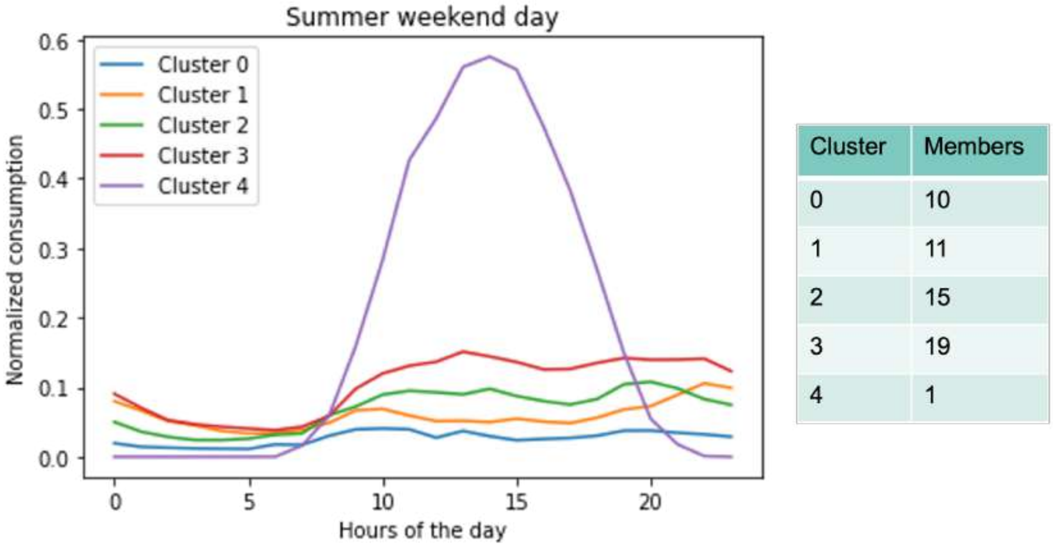


	all	Krefeld	Münster	Schwerte	Ulm	Kaiserslautern	Celle	Haßfurt
all	1	0.95	0.88	0.69	0.93	0.92	0.9	0.81
Krefeld	0.95	1	0.89	0.61	0.9	0.92	0.93	0.8
Münster	0.88	0.89	1	0.44	0.87	0.9	0.88	0.84
Schwerte	0.69	0.61	0.44	1	0.65	0.6	0.57	0.47
Ulm	0.93	0.9	0.87	0.65	1	0.94	0.91	0.82
Kaiserslautern	0.92	0.92	0.9	0.6	0.94	1	0.96	0.85
Celle	0.9	0.93	0.88	0.57	0.91	0.96	1	0.77
Haßfurt	0.81	0.8	0.84	0.47	0.82	0.85	0.77	1

Figure 5.14.: Correlation between clustering centers of the different locations for summer weekend days

other locations for the same season with the Pearson correlation coefficient. The result is a matrix where the correlation of all clusters is displayed. An example of this correlation matrix, for the summer weekend days time period, is displayed in figure 5.14. The matrix is symmetric with 1 as the highest possible value as the diagonal. The example time period illustrates the main problem of this approach: The over all correlation is high for most of the location combinations. For the location "Schwerte" the correlation with the other locations is significantly lower than for all other locations. The clustering centers and a table with the number of cluster members per cluster for the location "Schwerte" on summer weekend days are displayed in figure 5.15. It shows that the reason for the lower correlation values for this location is the unreasonable consumption profile of cluster 4, with only one household as cluster member. Therefore, the main problem of this approach is the low number of clustered households like 56 households in this example combined with unreasonable consumption profiles that may occur due to the limited completeness of the data or failures during the data collection.

Another possible weakness of the approach is that the cluster centers were ordered by the mean value for the comparison with other locations. This approach worked well when the



Cluster	Members
0	10
1	11
2	15
3	19
4	1

Figure 5.15.: Clustering results for location "Schwerte"

time periods were compared among each other, with similar cluster characteristics as shown in section 5.2.3. Other comparison methods for comparing the clusters that have a similar shape are thinkable like calculating the average distance with a suitable distance measure and compare the most similar clusters of the different location. Still, the problem of clusters with only a few members that influence the clustering center correlation remains. In the present work two households with the cluster membership of clusters with only one member were excluded from the further analysis because it occurred in several time periods and the normalized consumption were unreasonable compared to the other clusters like the example showed in 5.15. But even with this two households excluded, still clusters with only a few members appeared in the results for the locations.For this households the distinction, if the household is an outlier that has to be excluded or exception that has to be included in the analysis was more difficult to make. This shows the weakness of the approach and might be cause also by the fixed number of clusters $k = 5$ that does not fit to the number of households clustered for the locations.

Nevertheless, the correlation matrices have been calculated for all of the time periods defined in section 5.1.1. Resulting in eight correlation matrices for the eight time periods displayed in figure 5.16. Although, the over all correlation appears high among the clustering centers, no general conclusion on the location dependency of the clustering results can be drawn. But it can be concluded that there is no clear trend between the locations. For example

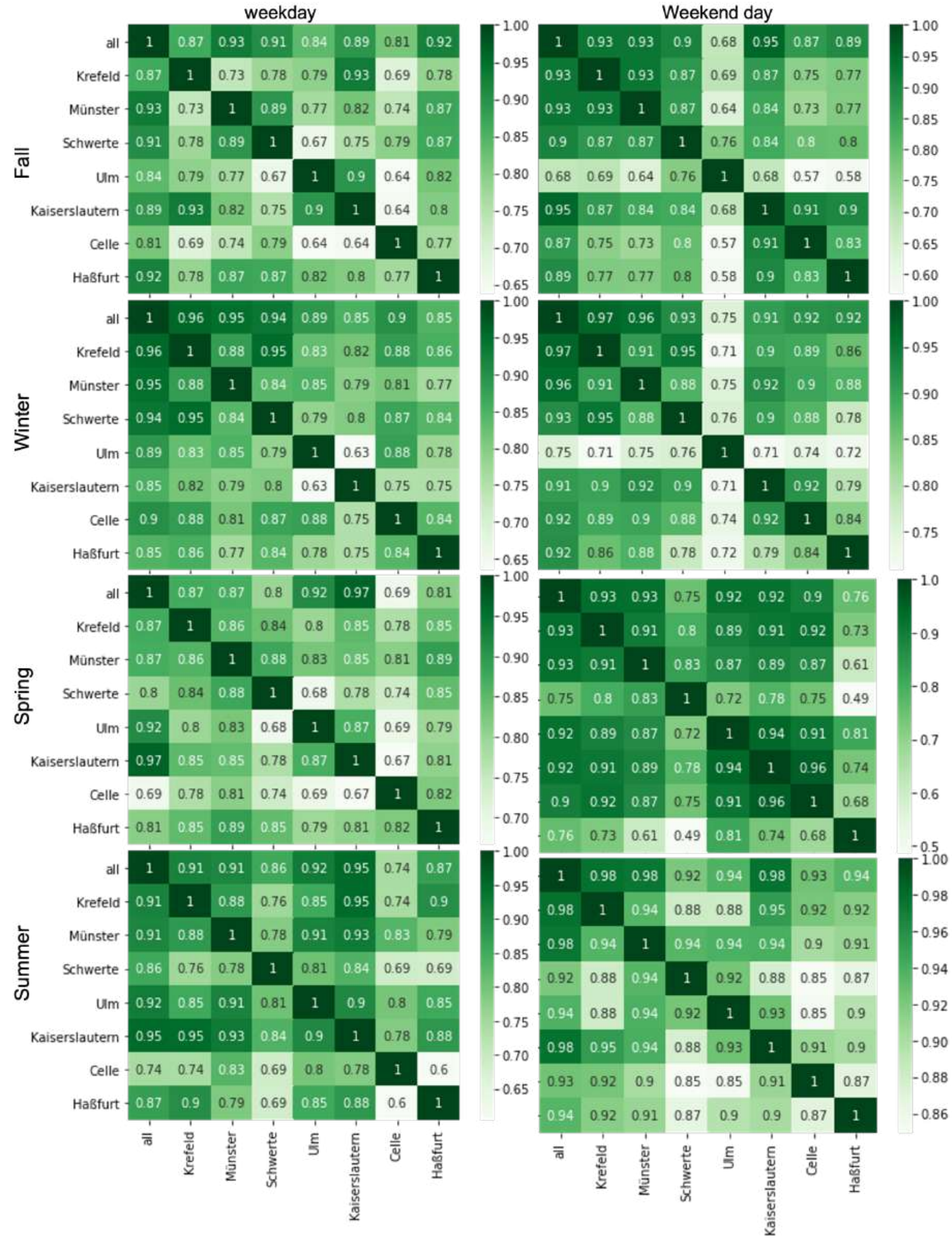


Figure 5.16.: Correlation between clustering centers of the different locations for summer weekend days

the locations Münster, Schwerte and Krefeld are located in the similar area of north western Germany, still the correlation is not higher among the clustering centers of this location compared to the other locations. Still this approach is not sufficient to make assumptions on the location dependency. Therefore, another approach that investigates the location share of the households, within the clusters from a cluster analysis performed for all households together, is conducted and presented in the next section.

5.3.2. Location Distribution within the Clusters

The second approach was to evaluate the share of locations within the clusters. The number of clusters was set to five as this was the number of clusters found to work well with the data set in section 5.2.1. Afterwards, the clustering was executed for all of the households for the one year period. For all of the clusters the distribution of locations within the cluster was determined from the survey data.

The results of this approach are visible in table 5.7. The column "Data set" shows the distribution of the household locations for all the analysed households followed by the remaining columns that show the corresponding distribution within the clusters found by the algorithm. The first column serves as the reference column and the more similar the locations are compared to the first column with values, the smaller the probability that the location has a high impact on the cluster membership. With the exception of a few examples, like the high share of households with the location Münster and low share of households from Kaiserslauatern in cluster 4 as well as the other way around in cluster 0, the location share within the clusters is similar to the data set reference column. To further evaluate the results the absolute differences to the reference column is displayed in table 5.8. The mean difference of the location shares among the clusters is calculated from this table to 4.99 %.

Location	Data set	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Münster	0.207815	0.123077	0.174603	0.117021	0.322581	0.473684
Kaiserslautern	0.177620	0.346154	0.174603	0.117021	0.086022	0.052632
Hassfurt	0.163410	0.123077	0.185185	0.244681	0.096774	0.157895
Krefeld	0.154529	0.153846	0.190476	0.170213	0.107527	0.087719
Ulm	0.110124	0.076923	0.105820	0.138298	0.139785	0.105263
Schwerte	0.099467	0.092308	0.084656	0.138298	0.129032	0.052632
Celle	0.087034	0.084615	0.084656	0.074468	0.118280	0.070175

Table 5.7.: Location shares within the data set and the clusters

Location	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Münster	0.084738	0.033212	0.094413	0.121855	0.274328
Kaiserslautern	0.168534	0.003017	0.064218	0.089708	0.124048
Hassfurt	0.040333	0.021775	0.073703	0.064509	0.002696
Krefeld	0.000683	0.035947	0.020728	0.055628	0.065244
Ulm	0.033201	0.004304	0.034206	0.021744	0.002981
Schwerte	0.007159	0.014811	0.034553	0.032401	0.045896
Celle	0.002418	0.002378	0.004560	0.033845	0.033462

Table 5.8.: Absolute difference to location share in the data set

In addition to the location, the influence of some other explanatory variables on the clustering results is examined. Since the clustering centers are ordered by increasing mean values of the normalized consumption it is expected that explanatory variables, that influence the energy consumption, found in the literature should impact the cluster membership. In [21] the floor area of the households was found to be an important determinant on the residential electricity use. Therefore, the average floor size of the households of the different clusters is calculated from the survey data that comes with the Intelliekon data set. Furthermore, the average hourly energy use in Wh is calculated for the clusters. The results are presented in table 5.9. It is remarkable that the floor size is not increasing among the cluster while the total energy consumption is increasing. This indicates on the one hand that the different clusters still represent the total energy use even with normalized values as input features and on the other hand, that for this data set the floor size may not be as important for the energy consumption as assumed in the literature.

Cluster	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Consumption Wh	205.39	346.49	436.71	486.20	586.48
Floor size in m^2	115.10	134.40	161.01	157.01	132.56

Table 5.9.: Average floor size of the households within the corresponding clusters

In conclusion, both approaches indicated a low location dependency within the country. It is important to mention that the distances between the locations are rather small with the maximum direct distance of 470 km between Ulm and Celle and the climate and geographical conditions are similar among the locations. All in all, the results support the assumption that a combined analysis of the two data sets from Germany is not obstructed by the different locations. Therefore, in the next sections the data processing of the CoSSMic data set, followed by the combined analysis, are presented.

5.4. Data Proccessing CoSSMic

For the analysis of the CoSSMic data set, the one-year time period from March 2016 to February 2017 was used. In order to display the results similar to the clustering results of the Intellikon data set, the mean values of the households for the hours of the days have been used. In figure 5.17 the average of the values for the import, export and PV generation data available for the households are shown. It is noticeable that for resident 1 only the PV generation data but no export data is available. For this resident, it is assumed that all the energy generated by the PV system is exported directly to the grid because the import data shows no lower values during the hours where PV energy is generated. Unfortunately, the documentation of the data set contains no information to verify this assumption. Further it stands out that for resident 6, there is a high difference between the generated PV energy and electricity exported to the grid. The traced appliances do not explain this difference. In the documentation on the data in [56] there is mentioned, that the PV system of this resident is composed out of a 4 kWp system and an additional 5 kWp PV system on the neighbour roof. Thus, the export data probably includes only one of the systems while the generation data is from both systems. Again this assumption could not be verified because of limited documentation for the data set.

The two data sets, that are to be combined in this thesis, differ in the way the consumption data is collected. In the Intelliekon data set the values are the hourly consumption data, while in the CoSSMic data set the grid import and export data as well as PV generation data and appliance level data represent the raw data. In order to enable the combined analysis of the data sets, in the present work the consumption data of the households from the CoSSMic data set was reconstructed for households with PV generation data from the available data. Therefore, the import data and the PV generation data where summed as electricity input while the export data was subtracted as electricity output. Figure 5.18 visualizes the reconstruction of the consumption data. The reconstruction was used on the hourly values of the whole one year time period after the normalization. Then the resulting consumption data was grouped by the hour of the day values. The consumption was calculated with the explained method for resident 3 and 4. For the other residents the import data was used as the consumption

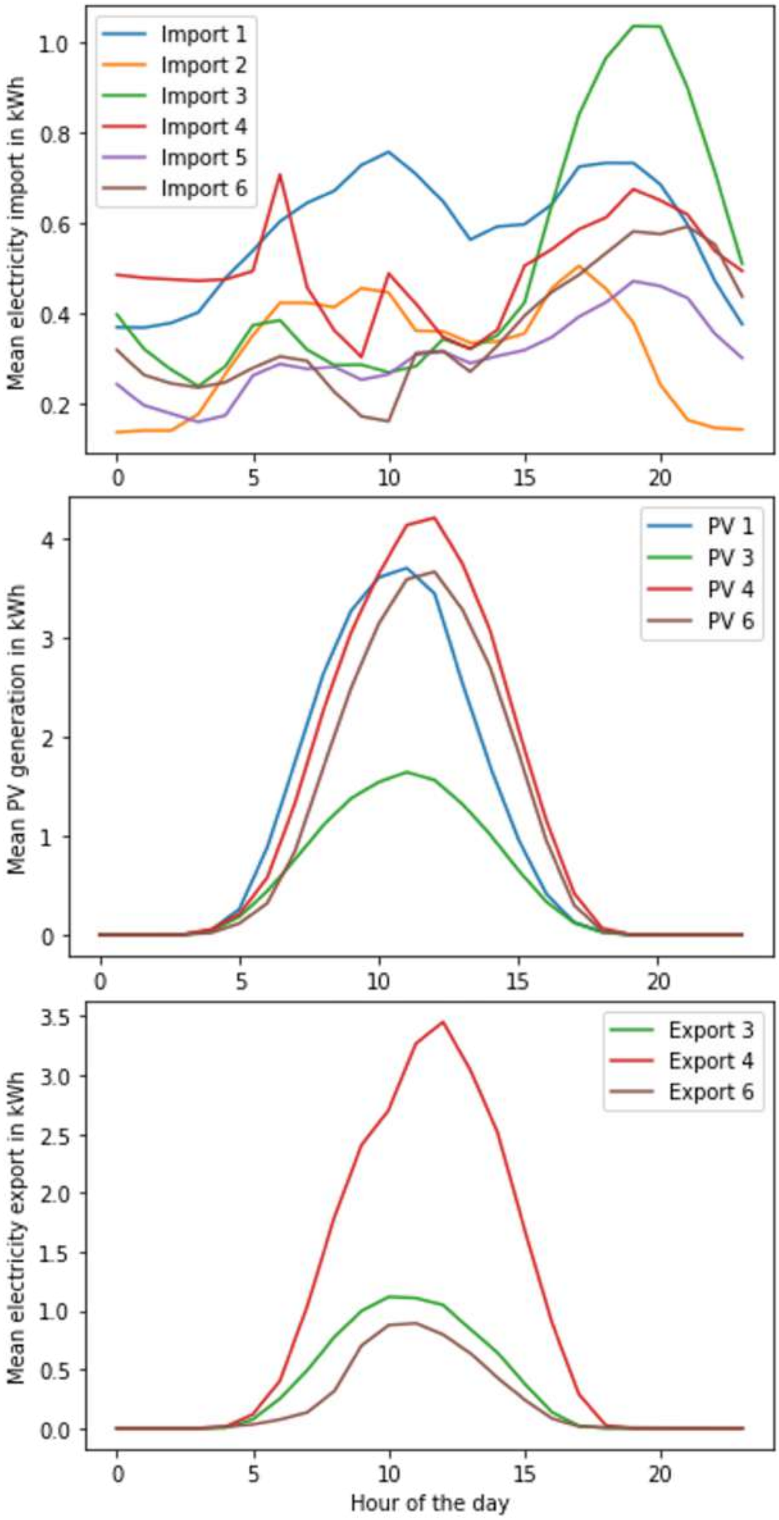


Figure 5.17.: Mean import, PV generation and export data for the households of the CoSSMic data set

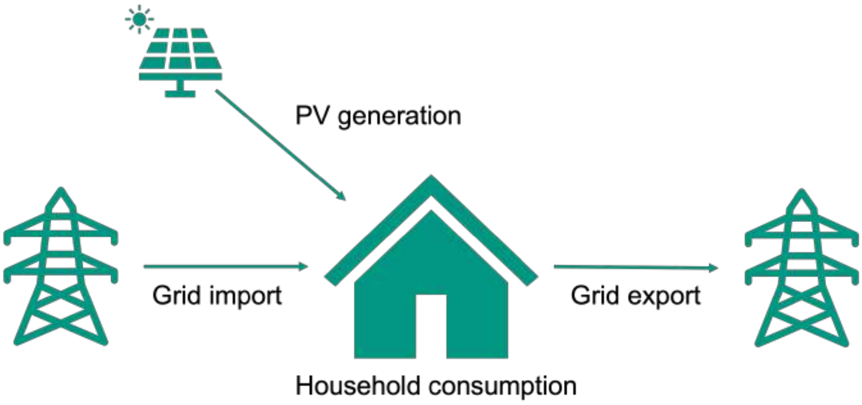


Figure 5.18.: Reconstruction of the consumption data for the CoSSMic households

data for the following reasons: Resident 2 and 5 do not have PV and export data. For resident 1 there is no export data available, therefore, as explained earlier in this section, it is assumed that the energy generated is exported directly to the grid. Resident 6 the export and PV data are disregarded because of the high difference of export and PV generation, probably due to split configuration of the PV system as explained before. The results are displayed in figure 5.19. The consumption reconstructed for resident 3 appears reasonable with the typical high evening peak and a distinct midday peak. For resident 4 the constant normalized consumption

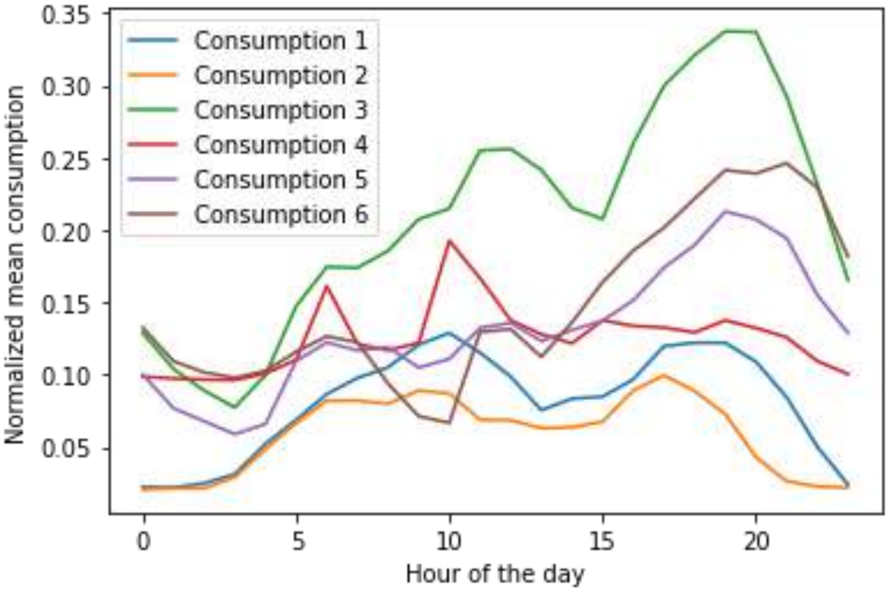


Figure 5.19.: Mean consumption profiles of residents from the CoSSMic data set

level over the day may be caused by the fact that it is the only resident with a reported EV and additional a heat pump is installed. The representative load shapes of households 1 and 6 appear reasonable which supports the assumptions made for this households. Still, especially for household 6 the decision to disregard the PV and export data of this household might lead to false consumption data and is to be treated carefully. It shows that a detailed documentation of the data is crucial for combined analysis of data sets.

5.5. Combined Analysis

For the linking the two data sets described in section 3.3 the clustering results of the Intelliekon data set have been compared to the CoSSMic data set. The data sets differ in several aspects that are explained in the following:

- Temporal aspect: For both of the data sets the one year period with the highest rate of available data has been used for the comparison. For the Intelliekon data set this is the periode from December 2009 to November 2010. The used period of the CoSSMic data set is from March 2016 to Febuary 2017
- Number of participants: While after the pre-processing 563 German households remained for the analysis, only 6 residential households have been used of the CoSSMic data set.
- Data structure: The different structure of the data is presented in section 3.3 and made different steps for the preprocessing of the data necessary.
- Additional data: While survey data on socio-economic factors and household characteristics is available for the Intelliekon data set, the CoSSMic data set exhibits traced appliances and PV generation data with a limited documentation.

Due to the differences, the data processing steps could not be performed equally for the two data sets. For the Intelliekon data set sufficient households were available to perform a cluster analysis, to find clusters that represent the consumption behaviour of the households. On the other hand, since only six households were analyzed of the CoSSMic data set, performing the cluster analysis applied to the Intelliekon data set would not provide any additional information for this data set. Therefore, the clustering centers of the Intelliekon data set were compared to the mean consumption data of the households from the CoSSMic data set. The consumption for the CoSSMic households is composed out of the electricity import, export and the generated PV data as explained before in this chapter in section 5.4. To keep a good interpretability of the results a whole year period was used for the combination. The combined visualization of the clustering centers of the Intelliekon data set and the mean consumption values of the CoSSMic data set is displayed in figure 5.20. The clustering centers

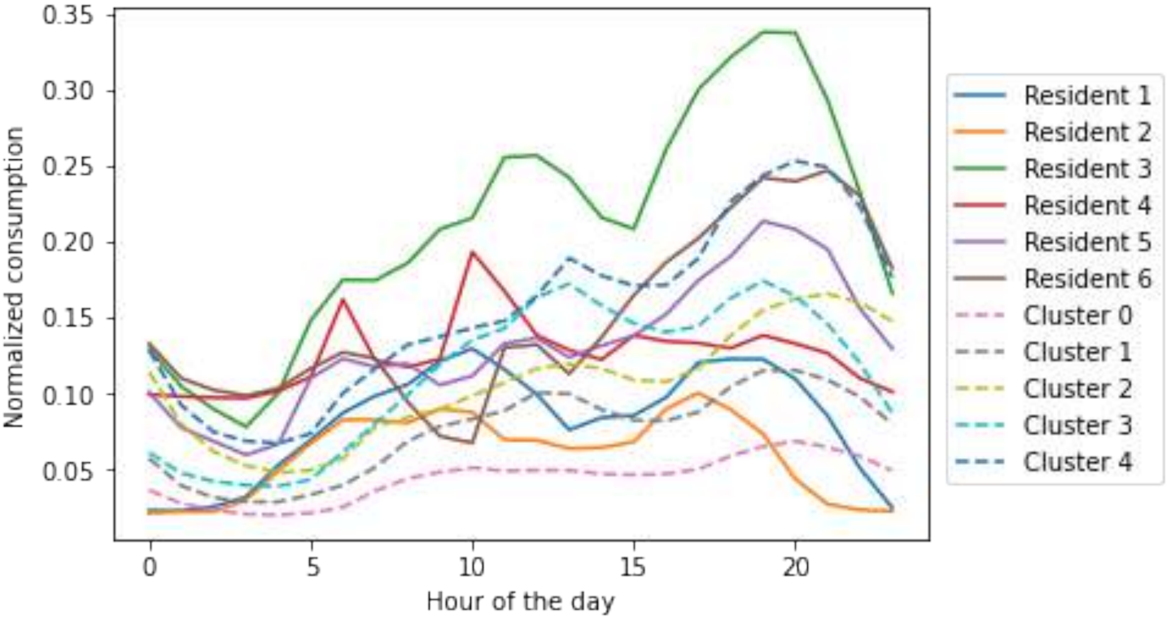


Figure 5.20.: Normalized mean electricity consumption of the CoSSMic residents and clustering centers of the Intelliekon data set

of the Intelliekon data set are marked with dashed lines, while the consumption data from the CoSSMic data set are characterized by solid lines. It is visible that there are clear similarities between the clustering centers and the residential mean consumption. For example the reconstructed consumption profile of resident 3 is similar to the shape of cluster 4 with the difference of a higher level of the normalized consumption.

In section 5.3.2 it was shown, that the order of the clusters ordered by the mean normalized consumption also relates to the total consumption level. The residents with PV generation, especially resident 3,4 and 6, exhibit a high normalized consumption level compared to the households without PV generation in 5.20. To support this assumption the total mean electricity value for the different residents is visible in table 5.10. For resident 3 and 4 the reconstructed consumption from import, export and PV data as well as the import data itself is displayed. The mean consumption of the households without PV are the lowest among the

Resident	Mean hourly consumption in kWh
Import 1 (PV)	0.5826
Import 2 (no PV)	0.3164
Import 3 (PV)	0.5175
Reconstructed consumption 3 (PV)	0.6195
Import 4 (PV)	0.4936
Reconstructed consumption 4 (PV)	0.7564
Import 5 (no PV)	0.3034
Import 6 (PV)	0.3564
Intelliekon mean	0.3708

Table 5.10.: Mean electricity consumption of the CoSSMic residents

households which supports the assumptions drawn from the comparison with the clustering results.

Interestingly, the mean hourly electricity import of the residents with PV generation (resident 1,3,4 and 6) is 0.4815 kWh compared to 0.3099 kWh for the residents without PV generation (resident 2 and 5), representing a 57.32 % higher electricity import of households with PV generation. With the reconstructed consumption of households 3 and 4, the difference is even higher with 86.74 %. Compared to the mean hourly electricity consumption of the Inteliekon data set with 0.3708 kWh, the import of PV households increased by 31.48 % and considering the reconstructured consumption by 56.07 %. This finding is interesting when considering the over all decreasing electricity demand of households in Germany by -10.5 % from 2010 to 2019[76]. Explaining this behaviour is difficult, since the information about the CoSSMic households is limited. It is possible that the PV residents have different household characteristics that are related to the electricity consumption, which could not be investigated due to no available survey data for the CoSSMic data set. Also, because of the low number of residents of the data set the results are not representative. Still, PV related mechanisms like the rebound effect, where the more sustainable or economical production leads to an increased consumption, are possible explanations to consider. Another reason could be the different appliance configuration among the households, for example two of the households with PV have a heat pump installed and one owns a EV. This new technologies can increase the electricity demand. Resident 4 is the only household of the data set with PV, Ev and a heat pump installed and the representative consumption profiles visible in 5.20 is different to the other residents and the cluster of the Intelliekon data. It exhibits a more even electricity consumption over the hours of the day without distinct peaks, indicating a changed consumption behavior.

It can be concluded that the general consumption behavior of residential households has not changed fundamentally in the seven year time difference between the data sets but the PV rooftop systems and other technologies like EV or heat pumps are influencing the consumption level.

Moreover, despite the necessary afford to align the data sets with different preprocessing and processing steps, it still can support a better understanding of the temporal development of the consumption behaviour of residential households. An important aspect is a good documentation of the used data sets, as otherwise assumptions are necessary to conduct the combined analysis. Further research on the impact of the new technologies is necessary for example on the impact of different appliances like PV, heatpumps or EV among the different seasons and with a more comprehensive inclusion of household characteristics, if available. Also it is to be investigated, how transferable the combination of data sets is for other countries and data set specification.

6. Summary and Outlook

In this chapter, a summary of the motivation and the background of the thesis are given, followed by the main conclusions of this thesis. In the end, an outlook on further research directions is pointed out.

6.1. Motivation and Background

The need for sustainable transformation in the society is one of the main challenges in this century. Thereby, the energy sector, especially the residential electricity supply, inhibits an important role, as it is the base for multiple new technologies, like EVs and heat pumps. In addition, residents participate more actively in the energy system by generating energy with PV systems and even storing and managing energy with batteries and intelligent systems. A better understanding of the temporal development of the residential electricity consumption behaviour is crucial to ensure a fact based and effective decision and policy making for the energy system stakeholders. The limited data availability of electricity consumption data, in particular with trial duration of more than one years, is a special challenge on this pathway. Therefore, in the present thesis, a workflow for a combined analysis on datasets with different temporal origins and specifications, is proposed, which includes data preprocessing, clustering, location dependency validation and a comparison of the consumption behaviors.

6.2. Conclusions

In the following, the main conclusions of the thesis are summarized.

- **Dataset combination:**

  Due to the generality issue of smart data analysis methods, combining different datasets of residential electricity consumption is seldom performed in the literatures. In the present work, a combination of two smart meter datasets with seven years time difference and different specifications were compared. Therefore, a workflow for combining different smart meter data set is proposed and visualized with an intuitive approach of representative consumption profiles. The combination showed comparable representative profiles of the electricity consumption behavior among the data sets, despite the seven year time difference. Still, the PV rooftop systems and other technologies like EV or heat pumps are influencing the consumption level and behavior. In particular, resident 4 from the CoSSMic dataset, the only resident with PV, EV and a heat pump installed together, exhibits a different representative load shape compared to the other resident and compared to the clusters found in the Intelliekon dataset, for example by not showing an distinct evening consumption peak and a relatively even consumption among the day.

- **Clustering methods:**

  For one of the datasets the unsupervised machine learning method clustering was used on the data to distinguish groups of households with a similar energy consumption behaviour. Therefore K-Means, Fuzzy C-Means and agglomerative hierarchical clustering were compared with the internal validation criteria SIL and DBI. The performance of the K-means algorithm, with internal validation indices of SIL = 0.8186 and DBI = 0.2884, was by far the best among the algorithms. Together with the over all simplicity and robustness this algorithm is qualified for a combined analysis between the two electricity consumption datasets of the present thesis.

- **Location dependency:**

  The location dependency of clustering results within a country has barely been investigated in the studies available to the author. It was evaluated with two different methods. The first approach showed a high Pearson correlation coefficient among the clustering results from different locations of 0.8048 indicating a high similarity of shape for the representative patterns. With the drawback of small clusters and lack of interpretability of the results a second approach was performed. The second approach compared the location share within the Intelliekon datasets with the location shares within the clusters, showing an average difference of 4.99 %. Therefore, the impact of different locations within Germany on the clustering results, and thus electricity consumption, was considered to not hinder the combined analysis of the datasets.

- **Normalization techniques:**

  Two normalization techniques have been investigated and the impact on the clustering results was evaluated. The first technique, where the highest/lowest values of the dataset were used as reference values, showed a 62.80 % increased SIL and a DBI decreased by 57.98 % compared to the technique where the daily maximum/minimum value was used. Thus, both indices indicate a better cluster quality of the first approach. Still, the techniques show different characteristics of the consumption behavior and the selection of the technique is case dependent.

- **Clustering results:**

  The cluster results show the different consumption behavior among the clusters. Various clusters have been distinguished, with different characteristics regarding the energy consumption levels, the distinction of the different peaks and the time when the energy is consumed. This characteristics as well as the internal validation criteria show clear similarities among the considered time periods, while reflecting the seasonality and differences within the week. In addition, a similar distribution of the number of households per cluster indicates a small exchange of households between the clusters, among the selected time periods.

6.3. Outlook

The combined analysis of smart meter sets enables different further research possibilities. For example, it revealed the interesting finding that the households with PV installed, show a 57.32 % higher electricity import than the households without PV in the same dataset and a 31.48 % higher import than the mean consumption from the Intellikon dataset. Because of the limited number of households wit PV, that have been analysed, and limited information from the datasets it only can be speculated about possible explanations, like the impact of modern technologies or rebound effects. Therefore, this high electricity import, despite the decentralised PV generation, calls for further research with more households included and a comprehensive inclusion of appliances, household characteristics and different time periods.

Another aspect, that can be an object of further research, is the combination of other smart meter datasets. For example, in chapter 3.2 two datasets of the UK that feature temporal differences and interesting specifications are presented as possible candidates. With the combined investigation of the more datasets, also the research on the impact of the household location on the electricity consumption behavior within other countries or even among different countries is to be can be investigated.

Bibliography

[1] Bundesverband der Energie- und Wasserwirtschaft (BDEW). “Stromverbrauch in Deutschland nach Verbrauchergruppen 2020”. Available online: `https://www.bdew.de/service/daten-und-grafiken/stromverbrauch-deutschland-verbrauchergruppen/`. Accessed: 2021-9-29.

[2] C. Breyer, O. Koskinen, and P. Blechinger. “Profitable Climate Change Mitigation: The Case of Greenhouse Gas Emission Reduction Benefits Enabled by Solar Photovoltaic Systems”. In: *Renewable and Sustainable Energy Reviews* 49 (2015), pp. 610–628.

[3] K. Bódis et al. “A High-Resolution Geospatial Assessment of the Rooftop Solar Photovoltaic Potential in the European Union”. In: *Renewable and Sustainable Energy Reviews* 114 (2019). 109309.

[4] B. Yildiz et al. “Recent Advances in the Analysis of Residential Electricity Consumption and Applications of Smart Meter Data”. In: *Applied Energy* 208 (2017), pp. 402–427.

[5] G. Chicco. “Overview and Performance Assessment of the Clustering Methods for Electrical Load Pattern Grouping”. In: *Energy* 42.1 (2012), pp. 68–80.

[6] M. Kochanski, K. Korczak, and T. Skoczkowski. “Technology Innovation System Analysis of Electricity Smart Metering in the European Union”. In: *Energies* 13.4 (2020). 916.

[7] Yi.Wang et al. “Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges”. In: *IEEE Transactions on Smart Grid* 10.3 (2019), pp. 3125–3148.

[8] Q. Sun et al. “A Comprehensive Review of Smart Energy Meters in Intelligent Energy Networks”. In: *IEEE Internet of Things Journal* 3.4 (2015), pp. 464–479.

[9] J. N. Adams et al. “How Smart Meter Data Analysis Can Support Understanding the Impact of Occupant Behavior on Building Energy Performance: A Comprehensive Review”. In: *Energies* 14.9 (2021). 2502.

[10] B. Völker et al. "Watt's Up at Home? Smart Meter Data Analytics From a Consumer-Centric Perspective". In: *Energies* 14.3 (2021). 719.

[11] Smart Energy International. "Global Trends in Smart Metering". Available online: `https://www.smart-energy.com/magazine-article/global-trends-in-smart-metering/`. Accessed: 2021-09-29.

[12] F. Gangale et al. "Smart grid projects outlook 2017: facts, figures and trends in Europe". Available online: `https://ses.jrc.ec.europa.eu/sites/ses.jrc.ec.europa.eu/files/publications/sgp_outlook_2017-online.pdf`. Accessed: 2021-10-13. 2017.

[13] Energy Information Administration (EIA). "Frequently Asked Questions (FAQs): How many smart meters are installed in the United States, and who has them?". Available online:`https://www.eia.gov/tools/faqs/faq.php?id=108&t=3`. Accessed: 2021-09-29. 2019.

[14] T. Serrenho, P. Zangheri, and B. Bertoldi. "Energy Feedback Systems: Evaluation of Meta-studies on energy savings through feedback". EUR 27992 EN. Luxembourg (Luxembourg): Publications Office of the European Union. JRC99716. 2015.

[15] J. Kelly and W. Knottenbelt. "Does Disaggregated Electricity Feedback Reduce Domestic Electricity Consumption? A Systematic Review of the Literature". In: *arXiv:1605.00962* (2016).

[16] B. P. Esther and K. S. Kumar. "A Survey on Residential Demand Side Management Architecture, Approaches, Optimization Models and Methods". In: *Renewable and Sustainable Energy Reviews* 59 (2016), pp. 342–351.

[17] F. McLoughlin, A. Duffy, and M. Conlon. "A Clustering Approach to Domestic Electricity Load Profile Characterisation Using Smart Metering Data". In: *Applied Energy* 141 (2015), pp. 190–199.

[18] J. L. Viegas et al. "Classification of New Electricity Customers Based on Surveys and Smart Metering Data". In: *Energy* 107 (2016), pp. 804–817.

[19] Commission for Energy Regulation (CER). "Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010". [dataset]. 1st Edition. Irish Social Science Data Archive. SN: 0012-00. 2012.

[20] P. Esmaeilimoakher et al. “Identifying the Determinants of Residential Electricity Consumption for Social Housing in Perth, Western Australia”. In: *Energy and Buildings* 133 (2016), pp. 403–413.

[21] A. Kavousian, R. Rajagopal, and M. Fischer. “Determinants of Residential Electricity Consumption: Using Smart Meter Data to Examine the Effect of Climate, Building Characteristics, Appliance Stock, and Occupants’ Behavior”. In: *Energy* 55 (2013), pp. 184–194.

[22] Y. Iwafune and Y. Yagita. “High-Resolution Determinant Analysis of Japanese Residential Electricity Consumption Using Home Energy Management System Data”. In: *Energy and Buildings* 116 (2016), pp. 274–284.

[23] M. A. Andor, D. H. Bernstein, and S. Sommer. “Determining the Efficiency of Residential Electricity Consumption”. In: *Empirical Economics* 60.6 (2021), pp. 2897–2923.

[24] B. Mashhoodi, D. Stead, and A. van Timmeren. “Local and National Determinants of Household Energy Consumption in the Netherlands”. In: *GeoJournal* (2019), pp. 1–14.

[25] G. Deng and P. Newton. “Assessing the Impact of Solar PV on Domestic Electricity Consumption: Exploring the Prospect of Rebound Effects”. In: *Energy Policy* 110 (2017), pp. 313–324.

[26] X. Li et al. “Sustainability or Continuous Damage: A Behavior Study of Prosumers’ Electricity Consumption After Installing Household Distributed Energy Resources”. In: *Journal of Cleaner Production* 264 (2020). 121471.

[27] M. Hansen and B. Hauge. “Prosumers and Smart Grid Technologies in Denmark: Developing User Competences in Smart Grid Households”. In: *Energy Efficiency* 10.5 (2017), pp. 1215–1234.

[28] M. Barreto, L. Pereira, and F. Quintal. “The Acceptance of Energy Monitoring Technologies-the Case of Local Prosumers.” In: *Proc., 6th International Conference on ICT for Sustainability.* 2019.

[29] B. V. M. Vasudevarao, M. Stifter, and P. Zehetbauer. “Methodology for Creating Composite Standard Load Profiles Based on Real Load Profile Analysis”. In: *Proc., Innovative Smart Grid Technologies Conference Europe (ISGT-Europe).* IEEE. 2016, pp. 1–6.

[30] A. Tureczek, P. S. Nielsen, and H. Madsen. “Electricity Consumption Clustering Using Smart Meter Data”. In: *Energies* 11.4 (2018). 859, p. 859.

[31] S. Waczowicz et al. “Demand Response Clustering-How Do Dynamic Prices Affect Household Electricity Consumption?” In: *Proc., Eindhoven PowerTech*. IEEE. 2015, pp. 1–6.

[32] Y. Wang, Q. Chen, and C. Kang. *Smart Meter Data Analytics: Electricity Consumer Behavior Modeling, Aggregation, and Forecasting*. Springer Nature, 2020.

[33] S. Ramos et al. “A Data-Mining-Based Methodology to Support MV Electricity Customers’ Characterization”. In: *Energy and Buildings* 91 (2015), pp. 16–25.

[34] A.-L. Klingler and F. Schuhmacher. “Residential Photovoltaic Self-Consumption: Identifying Representative Household Groups Based on a Cluster Analysis of Hourly Smart-Meter Data”. In: *Energy Efficiency* 11.7 (2018), pp. 1689–1701.

[35] R. Scitovski et al. *Cluster Analysis and Applications*. Springer, 2021.

[36] G. Gan, C. Ma, and J. Wu. *Data Clustering: Theory, Algorithms, and Applications*. SIAM, 2020.

[37] L. Jin et al. “Comparison of Clustering Techniques for Residential Energy Behavior Using Smart Meter Data”. In: *Proc., AAAI Workshop-Technical Report*. 2017, pp. 260–266.

[38] S. Park et al. “Data-Driven Baseline Estimation of Residential Buildings for Demand Response”. In: *Energies* 8.9 (2015), pp. 10239–10259.

[39] A.-L. Klingler, F. Schuhmacher, and K. Wohlfarth. “Identifying Representative Types of Residential Electricity Consumers—A Cluster Analysis of Hourly Smart Meter Data”. In: *Proc., 4th European Conference on Behaviour and Energy Efficiency (Behave 2016)*. 2016, pp. 8–9.

[40] A. Ozawa, R. Furusato, and Y. Yoshida. “Determining the Relationship Between a Household’s Lifestyle and Its Electricity Consumption in Japan by Analyzing Measured Electric Load Profiles”. In: *Energy and Buildings* 119 (2016), pp. 200–210.

[41] J. P. Gouveia and J. Seixas. “Unraveling Electricity Consumption Profiles in Households Through Clusters: Combining Smart Meters and Door-to-Door Surveys”. In: *Energy and Buildings* 116 (2016), pp. 666–676.

[42] L. Ruiz et al. “A Time-Series Clustering Methodology for Knowledge Extraction in Energy Consumption Data”. In: *Expert Systems With Applications* 160 (2020). 113731.

[43] P. Nystrup et al. “Clustering Commercial and Industrial Load Patterns for Long-Term Energy Planning”. In: *Smart Energy* 2 (2021). 100010.

[44] M. M. Gösgens, A. Tikhonov, and L. Prokhorenkova. “Systematic Analysis of Cluster Similarity Indices: How to Validate Validation Measures”. In: *Proc., International Conference on Machine Learning*. PMLR. 2021, pp. 3799–3808.

[45] C. Li et al. “Electricity Consumption Behaviour Analysis Based on Adaptive Weighted-Feature K-Means-Ap Clustering”. In: *IET Generation, Transmission & Distribution* 13.12 (2019), pp. 2352–2361.

[46] M. Syakur et al. “Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster”. In: *Proc., IOP Conference Series: Materials Science and Engineering*. Vol. 336. 1. 012017. IOP Publishing. 2018.

[47] M. R. Berthold and F. Höppner. “On Clustering Time Series Using Euclidean Distance and Pearson Correlation”. In: *arXiv:1601.02213* (2016).

[48] E. L. Ratnam et al. “Residential Load and Rooftop PV Generation: An Australian Distribution Network Dataset”. In: *International Journal of Sustainable Energy* 36.8 (2017), pp. 787–806.

[49] CoSSMic Deliverable D6.2. “Report on Collected, Analyzed and Evaluated Measurement Data”. Available online: `https://cordis.europa.eu/docs/projects/cnect/6/608806/080/deliverables/001-D62Reportoncollectedanalyzedandevaluatedmeasurementdata.pdf`. Accessed: 2021-8-21. 2016.

[50] Open Power System Data. “Data Package Household Data”. Available online: `https://data.open-power-system-data.org/household_data/`. Accessed: 2021-09-20. 2020.

[51] M. Pullinger et al. “The IDEAL Household Energy Dataset, Electricity, Gas, Contextual Sensor Data and Survey Data for 255 UK Homes”. In: *Scientific Data* 8.1 (2021), pp. 1–18.

[52] J. Schleich et al. *Smart Metering in Germany and Austria: Results of Providing Feedback Information in a Field Trial*. Tech. rep. Working paper sustainability and innovation, 2011.

[53] J. R. Schofield et al. “Low Carbon London Project: Data From the Dynamic Time-of-Use Electricity Pricing Trial, 2013”. In: *uK Data Service, SN* 7857.2015 (2015), pp. 7857–1.

[54] Pecan Street. “Real Energy. Real Customers. in Real Time”. Available online: `https://www.pecanstreet.org/work/energy/`. Aaccessed: 2021-09-25. 2012.

[55] Pecan Street. “Pecan Street Dataport”. Available online: `https://www.pecanstreet.org/dataport/`. Accessed: 2021-10-7.

[56] CoSSMic Deliverable D5.1. “Report of Buildings, Systems, Equipment and Users Involved in the Trials”. Available online: `https://cordis.europa.eu/docs/projects/cnect/6/608806/080/deliverables/001-D51Reportofbuildingssystemsequipment.pdf`. Accessed: 2021-8-18. 2014.

[57] A. M. Tureczek and P. S. Nielsen. “Structured Literature Review of Electricity Consumption Classification Using Smart Meter Data”. In: *Energies* 10.5 (2017). 584.

[58] J. MacQueen et al. “Some Methods for Classification and Analysis of Multivariate Observations”. In: *Proc. 5th Berkeley symposium on mathematical statistics and probability.* Vol. 1. 14. Oakland, CA, USA. 1967, pp. 281–297.

[59] I. P. Panapakidis and G. C. Christoforidis. “Optimal Selection of Clustering Algorithm via Multi-Criteria Decision Analysis (MCDA) for Load Profiling Applications”. In: *Applied Sciences* 8.2 (2018), p. 237.

[60] D. Arthur and S. Vassilvitskii. “K-Means++: The Advantages of Careful Seeding, P 1027–1035”. In: *Society for Industrial and Applied Mathematics* (2007).

[61] O. P. Mahela and A. G. Shaik. “Power Quality Recognition in Distribution System With Solar Energy Penetration Using S-Transform and Fuzzy C-Means Clustering”. In: *Renewable Energy* 106 (2017), pp. 37–51.

[62] T. M. Kodinariya and P. R. Makwana. “Review on Determining Number of Cluster in K-Means Clustering”. In: *International Journal* 1.6 (2013), pp. 90–95.

[63] R. Nainggolan et al. “Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) Optimized by Using the Elbow Method”. In: *Proc., Journal of Physics: Conference Series.* Vol. 1361. 1. 012015. IOP Publishing. 2019.

[64] J.-O. Palacio-Niño and F. Berzal. “Evaluation Metrics for Unsupervised Learning Algorithms”. In: *arXiv:1905.05667* (2019).

[65] P. J. Rousseeuw. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis”. In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65.

[66] P.-N. Tan, M. Steinbach, and V. Kumar. *Data Mining Introduction.* Bei Jing: The people post and Telecommunications Press, 2006.

[67] C. C. Aggarwal. *Data Mining: The Textbook.* Springer, 2015.

[68] M. E. Celebi and K. Aydin. *Unsupervised Learning Algorithms.* Springer, 2016.

[69] D. L. Davies and D. W. Bouldin. “A Cluster Separation Measure”. In: *IEEE Tansactions on Pattern Analysis and Machine Intelligence* 2 (1979), pp. 224–227.

[70] M. Goswami, A. Babu, and B. Purkayastha. “A Comparative Analysis of Similarity Measures to Find Coherent Documents”. In: *International Journal of Management, Technology and Engineering* 8.11 (2018), pp. 786–797.

[71] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual.* Scotts Valley, CA: CreateSpace, 2009.

[72] T. Kluyver et al. “Jupyter Notebooks – A Publishing Format for Reproducible Computational Workflows”. In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas.* Ed. by F. Loizides and B. Schmidt. IOS Press. 2016, pp. 87–90.

[73] F. Pedregosa et al. “Scikit-Learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[74] H. Meier et al. “Repräsentative VDEW-Lastprofile”. In: *VDEW Materialien M-32/99* (1999).

[75] J. Warner et al. “Scikit-Fuzzy version 0.4. 2” Available online: `https://pythonhosted.org/scikit-fuzzy/overview.html`. Accessed: 2021-8-15. 2019.

[76] Statistisches Bundesaamt. “Stromverbrauch der privaten Haushalte nach Haushaltsgrößenklassen”. Available online: `https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Umwelt/UGR/private-haushalte/Tabellen/stromverbrauch-haushalte.html`. Accessed: 2021-10-10. 2021.

A. Appendix

The source code of the analysis for the present work is available at:
`https://github.com/togg-dot/Master-thesis`
A revision of the source code is planned to make the code more comprehensible and to adapt it to the structure of the thesis.