Time Series Analysis on Energy Consumption Data

Inaugural-Dissertation

zur

Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Christian Bock

aus Düsseldorf

Düsseldorf, Oktober 2019

aus dem Institut für Informatik der
Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. Stefan Conrad

Koreferent: Prof. Dr. Martin Mauve

Tag der mündlichen Prüfung: 20. Dezember 2019

Ich versichere an Eides Statt, dass die Dissertation von mir selbstständig und ohne unzulässige fremde Hilfe unter Beachtung der "Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf" erstellt worden ist.

Die Dissertation wurde in der vorgelegten oder in ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den 22. Oktober 2019

Christian Bock

Danksagung

Als erstes danke ich meinem Doktorvater und Erstgutachter Prof. Dr. Stefan Conrad für die Möglichkeit, Teil seines Teams der wissenschaftlichen Mitarbeiter und Doktoranden zu sein. In den fünf Jahren, die ich an meiner Promotion gearbeitet habe, hat er mir ein angenehmes Umfeld für den wissenschaftlichen Austausch geboten. Er ist stets sehr an meiner Forschung interessiert gewesen, hat immer ein offenes Ohr für meine Fragen gehabt und hat stets mit wertvollem und hilfreichem Feedback zum Gelingen dieser Doktorarbeit beigetragen. Darüber hinaus danke ich Prof. Dr. Martin Mauve für sein Interesse an meiner Arbeit und seine Bereitschaft, als Zweitgutachter zu fungieren.

Des Weiteren danke ich allen Personen, die diese Arbeit direkt oder indirekt thematisch unterstützt haben, darunter Martin Dziwisch, Jan Hoppenkamps, Christian Kretschmann und Jan Sträter. Für ihre kostbare Hilfestellung bei Fragen zum Thema Clustering danke ich darüber hinaus Dr. Ludmila Himmelspach. Nicht unerwähnt bleiben darf außerdem die exzellente technische und administrative Unterstützung, wofür ich insbesondere Sabine Freese und ganz besonders unserem Systemadministrator Guido Königstein danke. Für weiteren technischen Support und die Bereitstellung von Berechnungsinfrastruktur bedanke ich mich beim Zentrum für Informations- und Medientechnologie (ZIM). Außerdem bedanke ich mich bei meinen Forschungskollegen für die stets konstruktiven Gespräche bei unseren Forschungstreffen, darunter Alexander Askinadze, Kirill Bogomasov, Daniel Braun, Janine Golov, Dr. Matthias Liebeck, Julia Romberg, Dr. Michael Singhof und Martha Tatusch.

Selbstverständlich bedanke ich mich ebenfalls bei meiner Familie und Bekannten, insbesondere bei meinen Eltern Martina und Michael, dafür dass sie mir das Studium ermöglicht haben und für die unermüdliche emotionale Unterstützung, sowie bei meiner Nachbarin Anneliese Kaiser dafür, dass sie mir stets wie eine Oma zur Seite stand.

Abstract

Nowadays, electronic devices are used in almost all aspects of everyday lives. Examples for these are computer, radios, television, light bulbs, domestic appliances and electric cars will possibly play a bigger part in the near future. Due to the enormous comfort and gain in efficiency these kinds of devices contribute during both spare time and in a professional environment, it is not possible to imagine one without the other. To enable the usage of these devices a whole industry branch has been created with the task to guarantee the supply of electrical energy almost everywhere. As a consequence of this, a complex electricity grid emerged. In order for the electricity grid to function properly, many roles such as energy producers, grid operators, energy providers, accounting grid coordinator and imbalance energy providers need to cooperate hitchlessly.
Since the supply of electrical energy needs to be ensured at all times, energy providers face the responsible and difficult task of forecasting the actual energy consumption of their customers in advance in order for energy producers to adjust the production accordingly. For this purpose, current business processes rely on so called *load profiles* which are statistical models that yield a good estimate for the total energy demand of all customers based on a set of assumptions about them. Traditionally, these assumptions about the customers are very vague, for example because the electricity meters of customers are read by the energy provider approximately only once per year in conjunction with the yearly accounting of electricity expenses, which does not allow to differentiate between uniform and peak consumption.
As a consequence of the increasing availability of intelligent metering devices, so called *Smart Meter*, it is progressively feasible to read the actual energy consumption of individual customers in real-time and transmit the data as a time series. This enables a variety of possible applications such as target-group-specific tariffs, where the prices can be adjusted in real-time, simplification of customer change processes or pointing out cost-saving opportunities for the customer by visualizing his or her consumption behavior. In addition to this, Smart Metering devices allow for the usage of algorithms that help to semi-automatically extract potentially useful knowledge from these datasets. These algorithms are part of the research area *Knowledge Discovery in Databases*. They are often employed if the data to be processed is too big or too complex for a manual analysis. The extraction of useful knowledge from consumption time series gathered by means of Smart Metering devices is the main topic of this thesis.

Zusammenfassung

In fast allen Bereichen des Alltags werden heutzutage elektrische Geräte eingesetzt. Beispiele hierfür sind Computer, Radios, Fernseher, Glühbirnen, Haushaltsgeräte und in naher Zukunft möglicherweise auch flächendeckend E-Autos. Aufgrund des enormen Komforts und der Produktivitätssteigerung, die derartige Geräte sowohl zur Freizeit als auch im beruflichen Umfeld beitragen, sind sie kaum noch wegzudenken. Um die Nutzung dieser Geräte zu ermöglichen, ist ein ganzer Industriezweig entstanden, um die Versorgung mit elektrischer Energie nahezu überall zu garantieren. Im Zuge dessen ist ein komplexes Stromnetz entstanden, für dessen Funktionieren zahlreiche Rollen wie die des Energieproduzenten, Netzbetreibers, Energielieferanten, Bilanzkreiskoordinators und Regelenergieanbieters reibungslos ineinander greifen müssen.
Da die Versorgung mit elektrischer Energie zu jedem Zeitpunkt sichergestellt sein muss, stehen Energielieferanten vor der verantwortungsvollen und schwierigen Aufgabe, den Strombedarf seiner Kunden möglichst akkurat vorherzusagen, damit Energieproduzenten ihre Produktion entsprechend optimal regulieren können. Hierfür kommen so genannte *Lastprofile* zum Einsatz, das heißt statistische Modelle, welche anhand von getroffenen Annahmen über die Kunden eine Schätzung für den Gesamtverbrauch der Kunden ermöglichen. Traditionell sind diese Annahmen über den Kunden sehr vage, da beispielsweise der Stromzähler der Kunden nur etwa einmal im Jahr im Rahmen der jährlichen Stromkostenabrechnung abgelesen wird, wodurch keine Unterscheidung zwischen gleichmäßigem Stromverbrauch und Lastspitzen möglich ist.
Durch den wachsenden Ausbau von intelligenten Stromzählern, so genannter *Smart Meter*, ist es zunehmend umsetzbar, unter anderem den tatsächlichen Stromverbrauch der einzelnen Kunden in Echtzeit auszulesen und als Zeitreihe zu übertragen. Dies ermöglicht eine Vielzahl von Möglichkeiten, beispielsweise zielgruppenorientierte Tarife, deren Preise in Echtzeit angepasst werden können, eine Verschlankung von Kundenwechselprozessen oder das Aufzeigen von Einsparmöglichkeiten gegenüber dem Kunden durch Visualisieren dessen Verbrauchsverhaltens. Darüber hinaus erlauben Smart Meter die Verwendung von Algorithmen, die dabei helfen, semi-automatisch potentiell nützliches Wissen aus den dabei entstehenden Datenmengen zu extrahieren. Diese Algorithmen sind Komponenten des Themengebiets *Knowledge Discovery in Databases* und werden häufig angewandt, wenn die zu untersuchenden Daten zu groß für eine manuelle Analyse sind; die Extraktion von nützlichem Wissen aus den durch Smart Meter erhobenen Verbrauchszeitreihen ist Hauptbestandteil dieser Arbeit.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Knowledge Discovery . . . 2
  - 1.2.1 Overview of the KDD process . . . 3
  - 1.2.2 Clustering Analysis . . . 5
  - 1.2.3 Dissimilarity measurements for time series . . . 9
- 1.3 Contributions . . . 11
- 1.4 Outline of this work . . . 12

2 Background **13**
- 2.1 Overview of the electricity grid . . . 13
- 2.2 Current challenges of energy providers . . . 16
- 2.3 Structure and purpose of load profiles . . . 18

3 Knowledge Discovery in the energy economy **27**
- 3.1 Impact on consumers . . . 27
  - 3.1.1 Financial benefit and influence on consumer behavior . . . 30
  - 3.1.2 Data privacy . . . 32
    - 3.1.2.1 De-pseudonymisation of customers . . . 33
    - 3.1.2.2 Differential privacy in a Smart Metering environment . . . 35
    - 3.1.2.3 Data reduction and data economy . . . 36
  - 3.1.3 Photovoltaic systems . . . 38
- 3.2 Impact on the electricity grid . . . 39
- 3.3 Impact on energy providers . . . 46
  - 3.3.1 Extraction of customer insight . . . 46
  - 3.3.2 Installation of intelligent metering systems . . . 48
    - 3.3.2.1 Prediction of household properties . . . 48
    - 3.3.2.2 Creation of target-group-specific tariffs . . . 51
    - 3.3.2.3 Forecast of the energy consumption . . . 54

4 Clustering using Smart Meter Data **57**
- 4.1 Description of datasets . . . 57
- 4.2 Assessment of clustering quality . . . 60
  - 4.2.1 Partition Coefficient . . . 61
  - 4.2.2 Compactness & Separation by Xie & Beni . . . 61
  - 4.2.3 Compactness & Separation by Bouguessa, Wang & Sun . . . 62

4.2.4 Fuzzy Hypervolume and Partition Density . . . 64
4.2.5 Silhouette Coefficient . . . 65
4.2.6 Average Clustering Uniqueness . . . 65
4.3 Framework for generating load profiles . . . 66
4.3.1 Day-type segmentation . . . 66
4.3.2 Identification of typical consumption patterns . . . 68
4.3.3 Compilation of load profiles . . . 69
4.3.4 Assessment of load profiles . . . 70
4.4 Evaluation . . . 71
4.4.1 Evaluation using the Euclidean distance . . . 72
4.4.1.1 Experimental Setup . . . 72
4.4.1.2 Results . . . 73
4.4.2 Evaluation using the Manhattan distance . . . 76
4.4.2.1 Experimental Setup . . . 76
4.4.2.2 Results . . . 78
4.4.3 Evaluation using the exponential Manhattan distance . . . 81
4.4.3.1 Experimental Setup . . . 81
4.4.3.2 Results . . . 84
4.4.4 Evaluation using weighted clustering . . . 84
4.4.4.1 Experimental Setup . . . 84
4.4.4.2 Results . . . 88

5 Online Clustering using Smart Meter Data 93
5.1 Motivation . . . 93
5.2 Related Work . . . 94
5.3 Building load profiles using Online Clustering . . . 96
5.4 Evaluation . . . 98
5.4.1 Experimental Setup . . . 98
5.4.2 Results . . . 100

6 Summary 103
6.1 Conclusions . . . 103
6.2 Future Work . . . 104

References 107

List of Figures 121

List of Tables 127

List of Own Publications 129

1

Introduction

1.1 Motivation

One of the most important achievements of humanity is the ability to gather, store and distribute energy. This feat has enabled the *Industrial Revolution* in the 18-th century and the ongoing *Digital Revolution* that has started in the 20-th century, thus laying the foundation for the lifestyle of modern society. Though the comforts of modern electrical devices, such as televisions or smartphones, as well as technologies like the Internet, have drastically changed the habits of citizens and the business models of companies in industrialized countries, these luxuries are typically taken for granted in our day-to-day life, with the average citizen putting very little thought into the inner workings of the energy infrastructure these achievements are based on. When one bears the importance of the security of the energy supply in mind, it is noteworthy that the energy economy has taken a backseat until the end of the 20-th century.
Ever since then, the energy market has begun to significantly change over the course of the following decades. For example, legislators have postulated an expansion in the usage of *renewable energies*, partly to decrease the carbon footprint of the country and partly in an effort to prepare for the increasing shortage of fossil fuels in the future. This has involuntarily let market participants to continously rethink the way of how the energy production and distribution are planned and carried out. Likewise, end-consumers have become more sensitized by the media to reduce their personal carbon footprint. Complementary, end-consumers have progressively been provided financial incentives to install and use personal renewable electricity generation plants such as photovoltaic systems. In addition, because of the liberalization of the energy market in the European Union, the increasing competitive pressure as a consequence thereof, as well as due to policies to adopt digitization and promote energy efficiency, market participants in the energy economy are compelled to engage with new technologies to prevent missing out on potentially valueable knowledge [Hay+14].
One of the new technologies that legislators in particular have placed high expectations in are *Intelligent Metering Systems*, which are commonly referred to as *Smart*

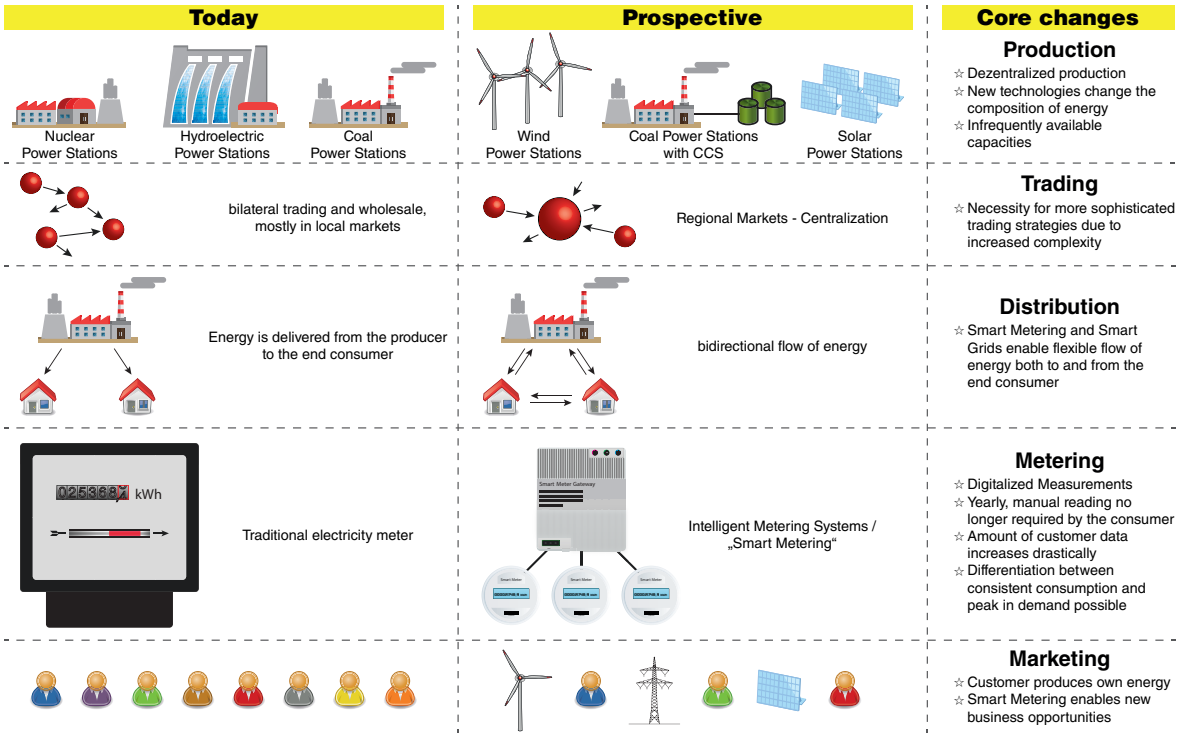


Figure 1.1: Overview of the transformation of the electricity grid due to digitization and growing deployment of *Intelligent Metering Systems*. Adapted from [EK13] with some assets taken from [Wik].

Meter. Though the base functionality of *Intelligent Metering Systems* consists of allowing the frequent remote reading of the meter by the electricity grid operator, governments recommend that Smart Metering devices also support advanced tariff systems and remote control of the supply and energy flow [Com12]. This prospective could enable the electricity grid to be transformed into a complex network, where devices semi-automatically coordinate themselves and make use of a more flexible demand-side according to user-defined parameters, dynamically scheduling tasks in an effort to increase overall energy efficiency, expand the integration of renewable energy sources as well as reduce costs for both energy market participants and end-consumers. This transformation of the electricity grid is also visualized in figure 1.1.

With the energy economy facing ambitious goals, this thesis presents means to address the upcoming challenges for the energy economy and end-consumers. In doing so, we lay the focus on approaches to extract useful knowledge from energy consumption time series yielded by employing Smart Metering devices. In preparation for this, we give a brief introduction into the concept of *Knowledge Discovery in Databases* and relevant techniques in the following section.

1.2 Knowledge Discovery

Over the last decades, advances in computing and storage technology has allowed for a drastic expansion of the collection of data. The term *data* corresponds to arbitrary sets of information recorded by an organization or an individual. In the past, the reasons

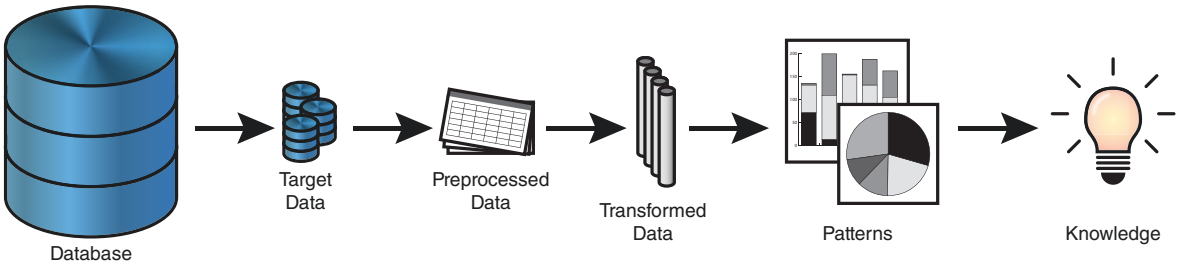


Figure 1.2: Overview of the processes associated with *Knowledge Discovery in Databases*, adapted from [FPS+96].

for collecting data has often been tied to specific use-cases and services, including, but not limited to, contractual details for accounting purposes between multiple parties or keeping track of inventory stock. Data which was not directly related to a given application was often dismissed to clear up capacities for data that was valuable for a given business.

As technological capacities progressed, storing huge amounts of data has become increasingly feasible. Because of this, organizations have started to build databases to be able to record and manage progressively large datasets. The reason in doing so is that these datasets may contain previously unknown knowledge which can give valuable insights vital for business decisions and academic literature. The process of analyzing these large amounts of data in order to gain the sought-after knowledge is called *Knowledge Discovery in Databases*, or *KDD* for short. In [FPS+96], the authors describe KDD as *the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.* In this context, *patterns* describe the results gained when the raw input data has been processed by the KDD framework. Since major parts of this thesis focus on techniques from the research area KDD, we will present this framework in more detail over the course of the following sections.

1.2.1 Overview of the KDD process

The basic steps of the KDD-framework are outlined in figure 1.2 and can be described as follows:

1. **Understanding of the application domain:** Prior to applying any transformation on the data, the goal of the KDD process has to be identified. This also includes acquiring relevant application knowledge.

2. **Selection:** Based on all data that is available for analysis, a *target dataset* is selected on which *Knowledge Discovery* will be performed. This may be a subset of variables or data samples.

3. **Preprocessing:** During preprocessing, the target dataset is pruned of noise and outliers wherever viable. In addition, strategies are chosen to account for missing values as well as for data samples where an unambiguous distinction between signal and noise is not possible. If the target dataset contains heterogeneous data from different data sources, they are integrated into one coherent dataset.

4. **Transformation:** In compliance with the goal of the KDD process and the knowledge about the application requirements, useful features adequate for representing the dataset are defined. If deemed appropriate, dimensionality reduction or transformation is applied on the preprocessed data to exclude variables of lesser importance or to achieve an invariant representation of the dataset.

5. **Data Mining:** After the dataset has been transformed, a fitting Data Mining algorithm, such as clustering, regression, classification, etc., is chosen and applied to extract the desired patterns from the data.

6. **Interpretation / Evaluation:** The discovered patterns are analyzed, visualized and documented. The insight achieved is consolidated with previously gained knowledge.

The goals and tools of KDD make it a strongly interconnected field of research combining statistics, maschine learning and databases [ES00]. Although the KDD process as illustrated above is often presented as a pipeline of processes, *Knowledge Dicovery* is to be understood as an iterative process, where during each step the user may opt to restart the process at an earlier stage. Using this procedure, parameters can be changed if necessary and emerging complications can be avoided. Depending on the complexity of the task, many iterations are required to achieve acceptable results.

Albeit the term *Data Mining* is only one step in the KDD process, it is often used interchangeably with KDD itself. This is because *Data Mining* is often seen as the core part of KDD, even though the results, and thus the success, of *Data Mining* are reliant on the proper execution of all previous steps. Depending on the concrete type of data that is being analyzed, different Data Mining techniques are used to carry out that task. The most common of these techniques include the following [ES00]:

- **Clustering:** The goal of clustering is to partition a given dataset into groups called clusters. The segmentation is processed under the optimization constraint that data objects belonging to the same cluster should be as similar as possible while data objects belonging to different clusters should be as dissimilar as possible. Because clustering algorithms do not rely on labeled or precategorized data objects, clustering belongs to the branch of machine learning called *Unsupervised Learning*. Since major parts of this thesis are utilizing clustering techniques we give a more thorough introduction to clustering in section 1.2.2.

- **Classification:** Given a set of *classes* and data objects where their membership to a *class* is known, *classification* is about training a model to learn to assign a class to data objects of which the class memberships are unknown. Unlike with clustering, the *categories* or *classes* need to be defined ahead of time, which is why classification belongs to the branch of machine learning called *Supervised Learning*.

- **Association Rules:** Association Rule Mining is about finding associations in a set of transactions. Associations describe common or strong correlations between items. These correlation may be of the form *if A and B then C*, which is typically notated as $A, B \rightarrow C$.

- **Generalization:** The goal of generalization is to describe the dataset in a more compact way. This may be achieved by aggregating multiple attributes or by reducing the number of data objects in the dataset.

For each of the *Data Mining* techniques as outlined above, there are numerous algorithms for the analyst to choose from. Each algorithm has its strengths and weaknesses which must be taken into consideration when selecting an algorithm for a concrete Data Mining task; e.g. data objects with categorical attributes require different handling than strictly numerical data.

Although *Knowledge Discovery in Databases* is a very well-known and reliable framework in the field of *Data Analytics*, other guidelines [Mül+16] and variations of KDD have emerged in recent years. While these modifications of KDD maintain the core concepts outlined above, they shift the focus of the discovery process more towards the interests of businesses and decision-making processes. One example for such a variation of KDD is *CRoss-Industry Standard Process for Data Mining (CRISP-DM)* [She00]. The main innovations of CRISP-DM compared to KDD are an additional step named *Deployment* which comes after the *Evaluation* as well as two separate steps called *Business Understanding* and *Data Understanding* for what was unified as *Understanding of the application domain* in KDD.

1.2.2 Clustering Analysis

Clustering is one of the main techniques used as part of the KDD process when it comes to analyzing a large dataset. It is used as a means to segment the elements of a given dataset into groups, such that elements which have been assigned to the same group are as similar to each other as possible while at the same time elements belonging to different groups are as dissimilar as possible. What makes clustering a very versatile technique in the KDD process is the fact that it does not rely on data that is already labeled or categorized and instead mainly relies on a way to quantize the dissimilarity for each pair of elements in the dataset and parameters specific to the chosen clustering algorithm. As each clustering algorithm has different characteristics, the concrete algorithm is typically chosen depending on the application task it is supposed to solve. Among the most important categories of clustering algorithms are the following:

- **Partitioning Clustering:** Clustering methods belonging to this category segment the data objects into a predefined number of groups which is usually given as a parameter for the algorithm. The segmentation can either be *crisp*, meaning a data object belongs to exactly one cluster, or *fuzzy*, meaning a data object belongs partially to multiple clusters depending on the membership degree of the data object to each cluster. Some representatives of this category of clustering algorithms are *K-Means* [Mac+67], *Fuzzy-C-Means* [Bez81], *Gustafson-Kessel* [GK78; BVK02], *Fuzzy-Maximum-Likelihood-Estimation* [GG89] and *ISODATA* [BH65; BD75].

- **Density-based Clustering:** While partitioning clustering algorithms segment the dataset into a number of clusters which is usually predetermined by a parameter given by the user, density-based clustering works by defining clusters as *dense* regions of data objects in the feature space separated by non-dense regions. Density-based methods usually do not require the number of clusters as

an input parameter, but a means to decide if a given data element is part of a *dense* region. A popular representative of this category of clustering algorithms is *DBSCAN* [Est+96].

- **Hierarchical Clustering:** The goal of hierarchical clustering methods is to construct a hierarchy of clusters, where clusters are merged if their distance is sufficiently small. Hierarchical clustering can either be *bottom-up (agglomerative)* by starting at each data object being its own cluster and then subsequently increasing the threshold for the distance at which clusters are merged, or *top-down (divisive)* by starting at one cluster containing the whole dataset and then subsequently lowering the threshold at which clusters are split [ES00]. In addition, several hierarchical clustering algorithms also partially incorporate a strategy from different clustering categories, such as the hierarchical density-based *OPTICS* [Ank+99] and the hierarchical partitioning *CURE* [GRS98].

- **Subspace Clustering:** Subspace clustering is a category of algorithms, where the main motivation is to apply clustering to datasets with a very high number of dimensions. The idea is to identify subspaces of the feature space where data objects form clusters without necessarily computing the distances in the complete feature space. This is to counteract the value of distance functions losing semantic information as the number of dimensions of the dataset increases, a property which is commonly referred to as the *curse of dimensionality* [Bey+99; FWV07]. Similar to hierarchical clustering, the approach to subspace clustering can be either *bottom-up* or *top-down*. Popular subspace clustering methods include *CLIQUE* [Agr+98] and *SUBCLU* [KKK04], both of which pursue the *bottom-up* approach.

As we will see in chapter 4 and 5, a useful approach to analyze datasets for the purpose of finding a solution for current application tasks within the energy economy is to employ partitioning clustering methods, which will be the primary focus for this thesis. Two well-known partitioning clustering methods, *K-Means* [Mac+67] and *Fuzzy-C-Means* [Bez81], solve the generic task of clustering algorithms by minimizing the following objective function:

$$J(\cdot) = \sum_{o=1}^{c} \sum_{i=1}^{N} u_{o,i}^{m} \cdot \|x_i - v_o\|^2 \tag{1.1}$$

Here, $x_i \in X$ corresponds to the data objects, $v_o \in V$ describe the clustering prototypes and $u_{o,i} \in U$ are the membership degrees. In the case of a crisp clustering, such as *K-Means*, one has $u_{o,i} \in \{0,1\}$ with $\forall i : \sum_{o=1}^{c} u_{o,i} = 1$, meaning that a data object x_i belongs to exactly one cluster v_o completely ($u_{o,i} = 1$) and does not belong to other clusters at all ($u_{o,i} = 0$). On the contrary, in the case of fuzzy clustering, such as *Fuzzy-C-Means*, the same objective function is used, but the membership degrees are softened due to $u_{o,i} \in [0,1]$, allowing data objects to partially belong to multiple clusters. The higher the membership degree $u_{o,i}$, the clearer x_i belongs to v_o.
In order to find an optimal solution for equation 1.1, clustering algorithms typically employ an iterative approach, where a set of computational steps are repeated in a cyclic manner until a termination condition is met. For *Fuzzy-C-Means*, these steps consist of first updating the membership degrees and then updating the positions of the clustering

Algorithm 1 Fuzzy-C-Means

Input: X, m, ϵ, c

Output: set of all clustering prototypes V, matrix containing the membership degrees U

1: Generate initial clustering prototypes $V^{(0)} = \left(v_0^{(0)}, \ldots, v_c^{(0)}\right)$

2: $r \leftarrow 0$

3: **repeat**

4: $U^{(r+1)} \leftarrow \begin{pmatrix} u_{1,1}^{(r+1)} & \ldots & u_{1,N}^{(r+1)} \\ \vdots & \ddots & \vdots \\ u_{c,1}^{(r+1)} & \ldots & u_{c,N}^{(r+1)} \end{pmatrix}$ with $u_{o,i}^{(r+1)} := \frac{dist^{\frac{2}{1-m}}\left(x_i, v_o^{(r)}\right)}{\sum_{o'=1}^{c} dist^{\frac{2}{1-m}}\left(x_i, v_{o'}^{(r)}\right)}$

5: $V^{(r+1)} \leftarrow \left(v_0^{(r+1)}, \ldots, v_c^{(r+1)}\right)$ with $v_o^{(r+1)} := \frac{\sum_{i=1}^{N}\left(u_{o,i}^{(r+1)}\right)^m \cdot x_i}{\sum_{i=1}^{N}\left(u_{o,i}^{(r+1)}\right)^m}$

6: $r \leftarrow r + 1$

7: **until** $\|V^{(r+1)} - V^{(r)}\| < \epsilon$

8: **return** V, U

prototypes. This procedure has the advantage of being easily implementable and thus highly practical. In contrast, one disadvantage of this approach lies in the dependency for the starting configuration of the clustering prototypes, causing the algorithm to sometimes terminate in a local optimum which might be far off the global optimum for the objective function. Algorithm 1 presents the general form of *Fuzzy-C-Means*. In literature, the termination condition of *Fuzzy-C-Means* is very commonly expressed as the algorithm continuing indefinitely until the change of the clustering segmentation compared to the prior iteration no longer exceeds a given threshold ϵ as seen in line 7 of algorithm 1, but implementations often also support the algorithm terminating after a user-specified maximum number of iterations, either as an additional or a surrogate termination condition for $\|V^{(r+1)} - V^{(r)}\| < \epsilon$.

While the quality of the results of the clustering process does depend on parameters specific to the clustering process, such as the initial starting configuration for the clustering prototypes, the quality of the input dataset also plays a major role, which is of particular importance if the dataset is based on real-world circumstances, as opposed to synthetic datasets. This is due to the fact that when recording real-world events or objects, external factors may impact the collection of data in a negative way, for example through inaccuracies in the measuring equipment or during preprocessing, technical failures that have occured during the transmission of data, as well as some aspects being outright unknown, for example because of survey participants not answering some questions in a poll. This, among possibly other reasons, may cause the dataset to contain *missing values*. The underlying principle for the absence of a value in the data is referred to as a *missing-data mechanism*. In general, there are three mechanisms for *missing values* [LR02]:

- **Missing At Random (MAR):** If the data is *missing at random*, then this means that the probability of a given value to be missing is dependent on the observed data, but not on the missing attribute value itself. More formally, the values are *missing at random* if the following condition is met:

$$f(M|X, \phi) = f(M|X_{obs}, \phi) \qquad \forall X_{mis}, \phi \tag{1.2}$$

Here, $M = (m_{i,n})$ indicate whether the n-th attribute of the i-th data object is available ($m_{i,n} = 0$) or missing ($m_{i,n} = 1$). Furthermore, X_{obs} denotes the set of observed values, X_{mis} represents the set of missing values, $X = X_{obs} \cup X_{mis}$ is the entire dataset and ϕ are unknown parameters.

- **Missing Completely At Random (MCAR):** If the data is *missing completely at random*, then the availability or absence of values does depend neither on the observed values X_{obs} nor the unobserved values X_{mis}:

$$f(M|X, \phi) = f(M|\phi) \qquad \forall X, \phi \tag{1.3}$$

- **Not Missing At Random (NMAR):** The mechanism of *missing values* is *not missing at random* if the probability of a given value being unobserved is dependent on the *missing value* itself.

While the failure mechanism for *missing values* is usually not known in advance for real-world datasets, the failure mechanism can have a significant influence on the quality of the resulting clustering segmentation [HC10b; HC10a]. In order to probe which *missing-data mechanism* is present for a given dataset, tests based on statistics can be applied; to test for NMAR and MAR, the *one-sample test* and *two-sample test* can be used, respectively [WB05], while a test based on the χ^2 test is presented in [Lit88] for MCAR.
After learning which failure mechanism is present, an analyst may decide on which method is applied to process the data as good as possible while keeping the application task in mind. In general, there are three approaches to handle missing values in the dataset [LR02]:

- **Adaption of analysis methods:** One way to accommodate for *missing values* in the dataset is to modify the methods that are part of the *Knowledge Discovery in Databases* process. This can happen by estimating missing data prior to applying Data Mining methods but still differentiate between estimated and measured data during analysis, or by extending the definition of computational methods that access the data tuples so that they can be applied even if the tuples contain *missing values*.

- **Complete-Case Analysis:** Possibly the simplest approach in processing a dataset that contains *missing values* is to remove all data tuples that are partially unobserved and to analyze only the data tuples that have been observed completely. This approach is particularly tempting if the amount of *missing values* is relatively small. It should be noted however that while this approach is uncritical if the mechanism that let to the occurrence of *missing values* is *MCAR*, it can lead to wrong conclusions if the mechanism is *NMAR*.

- **Imputation of missing values:** This technique works by replacing *missing values* in the dataset with values that are typically derived from observed data. The chosen method to assign a value to fields that are unobserved can be very simple, such as computing the arithmetic mean or median, but also includes arbitrary elaborate methods as long as they are deemed adequate for the given application by an analyst. After *missing values* are replaced with imputed values, these values are treated as measured data, meaning that data analysis can be performed normally without the need to further account for *missing values*.

One of the major drawbacks of *Complete-Case Analysis* is that it artificially reduces the size of the dataset. Similarily, *imputation of missing values* can majorly skew the results of the analysis depending on the amount of unobserved data and the concrete imputational method used. As for *adaption of analysis methods*, while this methodology does not alter the input dataset, the choice of how the data mining algorithm is modified can still have a tremendous impact on the final analysis result. Some approaches on how Fuzzy-C-Means, as presented in algorithm 1, can be modified to work with datasets containing *missing values* are presented in [HB01; SL01], while their impact on the clustering quality is investigated and discussed in [HC10a; HHC11].

1.2.3 Dissimilarity measurements for time series

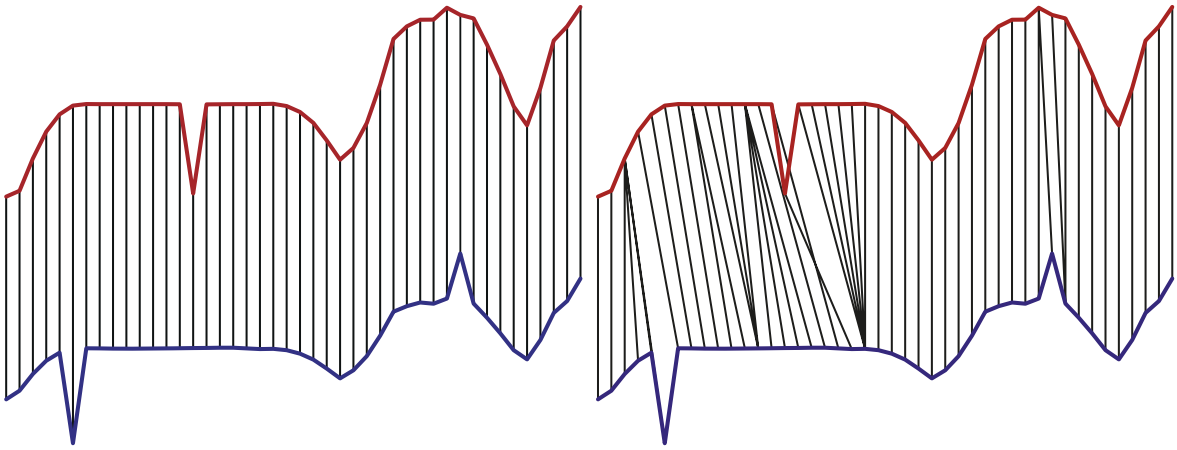
Though clustering methods are a useful tool to find patterns in data, they are primarily designed for *static* data, meaning datasets which are presented as a set of tuples whose features do not evolve over time. This usually makes applying clustering on time series data a non-trivial task. As [Lia05] mentions, further care should be taken in regard as to whether the measured values are discrete or continuous, sampled uniformly or non-uniformly, univariate or multivariate and whether or not the time series are of equal length. Depending on the dataset and the application task, the analyst may opt to modify the function to evaluate the dissimilarity between two data objects to a measure more appropriate for time series. That way, existing clustering algorithms can be used without further modifications. Another possibility would be to convert the time series data or extract features from them in a way such that existing clustering algorithms can handle the resulting data. Over the course of this section, we focus on the first two categories and present examples for dissimilarity measures for time series as well as feature extraction techniques that can help analyze unknown time series.

When trying to evaluate the dissimilarity of two given time series, the most simple case is when both time series are of equal length. In that case, the L_p-norm, sometimes also called the *Minkowski distance*, can be used. It is defined as follows:

$$dist\,(a,b) = \left(\sum_n |a_n - b_n|^p \right)^{\frac{1}{p}} \tag{1.4}$$

For $p = 2$, equation 1.4 results in the well-known euclidean distance. However, depending on the application task, the L_p-norm may be unsuitable to compare the given time series. For example, the time series may be of unequal length, the values of the time series may have different means and variances, or the time series may have local compressions and stretchings.

In order to circumvent the problems outlined above, it is often desirable to mathematically capture the human perception of the *shape* of a given time series and use that to compare the data objects in the dataset instead of focusing on the actual measured values directly; this is the basic idea behind *Shape Definition Language (SDL)* [Agr+95]. Here, an expert first defines an alphabet Σ. The symbols contained in Σ describe the shape of a portion of the time series, for example *ascending*, *strongly ascending* or *descending*. Before analysis, a time series is then encoded using the corresponding symbols of Σ. Due to the concept, the choice of the alphabet strongly influences which curve shapes can be detected at all, but also if a time series representation by a sequence of symbols is unique. Once such a representation by sequence of



(a) (b)

Figure 1.3: Schematic example for two time series whose distance is computed using (a) the euclidean distance and (b) *Dynamic Time Warping (DTW)* [BK59]. The *Warping Path* is visualized by black lines connecting the measurements of the two time series which have been matched up.

symbols is available, one way to evaluate the similarity between two time series might be to check whether the symbols match or not and count the number of non-matching symbols as a measure for the dissimilarity. Another approach might be to treat the sequences of symbols as a string and apply techniques such as the *Damerau-Levenshtein Distance* [Lev66; Dam64; Bar07], sometimes also referred to as the *Edit Distance*, which evaluates the dissimilarity between two strings as the minimal number of operations, consisting of *insert*, *substitution*, *swapping* and *delete*, to transform one sequence of characters into the other. In case the representation of a time series using the alphabet Σ is not unique, additional effort has to be expended when comparing time series, for example by making the dissimilarity measure aware of other representations.

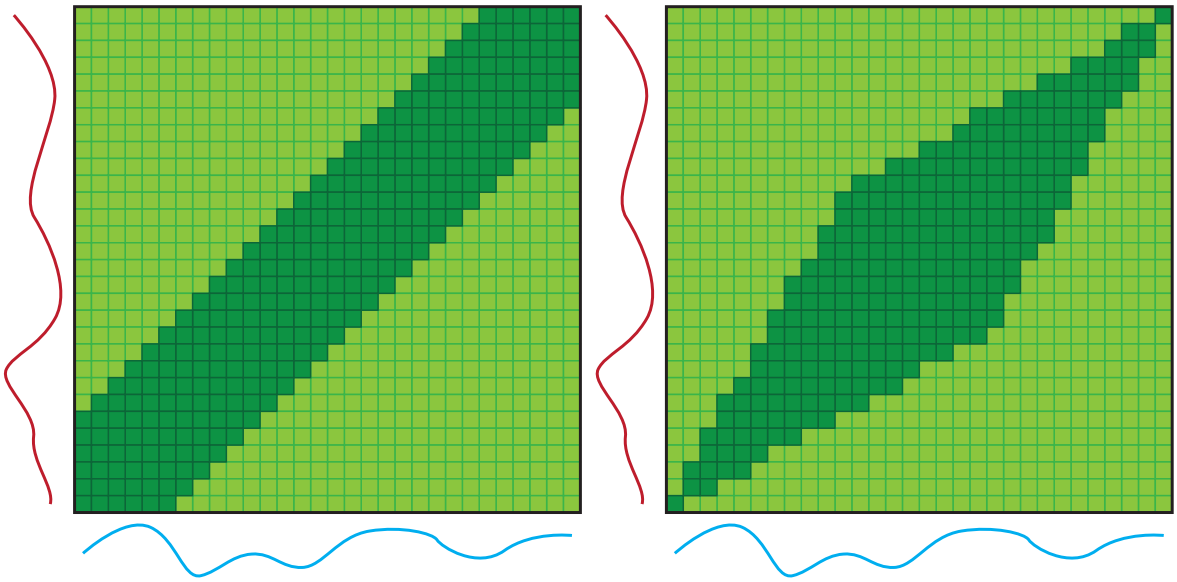
Another approach to compare two time series of possibly different length is to use *Dynamic Time Warping (DTW)* [BK59]. DTW is a *pseudo-distance measure* where the basic idea is to find a non-linear matching for the measurement of two time series. Intuitively, this corresponds to DTW locally compressing and stretching the time series. In a nutshell, DTW implements this as follows:
let $a = (a_1, a_2, \ldots, a_n)$ and $b = (b_1, b_2, \ldots, b_m)$ be two time series. Furthermore, let M be a $n \times m$-matrix where the cell (s, r) contains the distance between the measurements a_s and b_r. Now, the goal of DTW is to find a *Warping Path* $W = (w_1, w_2, \ldots, w_Z)$ with $w_z = (s, r)_z$, $w_1 = (1, 1)$ and $w_Z = (n, m)$ such that the total *cost* of the Warping Path is minimized:

$$DTW\,(a, b) = \min\left\{\sqrt{\sum_z w_z}\right\} \tag{1.5}$$

In addition, the chosen Warping Path must be ordered monotonically:

$$s_{z-1} \leq s_z \qquad \text{and} \qquad r_{z-1} \leq r_z \tag{1.6}$$

Figure 1.3 illustrates this concept by comparing the Warping Path of DTW to the euclidean distance.



(a) (b)

Figure 1.4: Illustration of extensions to *Dynamic Time Warping (DTW)* [BK59]. The figures show two time series (red and cyan graph) aligned to the $n \times m$-matrix used to construct the *Warping Path*. The depicted approaches are (a) the *Sakoe-Chiba Band* [SC78] and (b) the *Itakura Parallelogram* [Ita75]. The area colored in dark green correspond to the region the *Warping Path* of DTW is supposed to not trespass.

Overall, DTW is a *pseudo-distance measure* which allows to incorporate the *shape* of the time series into the comparison much better than by using a distance measure based on the L_p-norm as given in equation 1.4. Because of this, further research has been conducted based on the concept of DTW, for example *Windowing* [BC94], *Slope Weighting* [SC78; KL83] or *Step Patterns* [Ita75; MRR80]. The basic idea of these extensions is to artificially employ a constraint on the Warping Path of DTW to avoid singularities, which denote the case when a certain a_s or b_r gets matched with a large number of measurements in the other time series. Figure 1.4 illustrates some of the extensions to accomplish this goal. Other approaches, such as *Fast-DTW* [SC07], aim to deliver an approximation of DTW with the advantage of an improved runtime. Further research has been conducted in order to improve the overall quality of the DTW measure, for example by using *Derivative Dynamic Time Warping (DDTW)* [KP01] where the basic idea is to construct the warping path based on the derivative of the values of the time series instead of the values directly; this puts an even stronger emphasis on the *shape* when comparing two time series.

1.3 Contributions

After giving a brief introduction into the inner workings of the energy economy as well as upcoming changes of the energy infrastructure, the following chapters present our contributions to the field of energy economy research. These include a new framework for building *load profiles* from Smart Metering data, four approaches based on the presented framework as well as one approach that incorporates *Online Data Mining* techniques. We briefly summarize the main contributions of this thesis as follows:

- We present a framework that describes a methodology to construct *load profiles* using Smart Metering data. This framework sets itself apart from other proposals by constructing load profiles as energy consumptions forecast models that adhere to current business processes within the energy economy, thus being easy to adopt by the industry [Boc16; Boc17; Boc18].
- Another contribution of this work are the experimental evaluation of four approaches based on the framework introduced in [Boc16; Boc17; Boc18] using real-world datasets. Two of these approaches have been presented in [Boc17; Boc18] while the other two are new.
- Lastly, we propose a new approach that combines the advantage of building load profiles that are easy to adopt due to the load profiles being built with existing business processes in mind with *Online Data Mining* techniques. This allows the load profiles to become sensitive to changing customer behavior without the need to restart the entire KDD workflow from scratch. In addition to lowering computational requirements, this approach gives the analyst the opportunity to choose the optimal amount of history to include in the analysis, therefore fine-tuning the performance of the forecast models.

1.4 Outline of this work

The rest of this thesis is structured as follows. In chapter 2, we introduce the basic concepts and business processes in the energy economy. This includes the general setup of the electricity grid, day-to-day challenges of a market participant as well as the basic structure of load profiles, a common technique to forecast the energy demand of consumers.
We expand on this foundation in chapter 3, where we present and discuss approaches from academic literature based on the concept of *Knowledge Discovery in Databases* for gaining useful insight both from the perspective of an end-consumer and an energy provider. We also outline some of the upcoming changes to the electricity infrastructure itself.
In chapter 4, we introduce a framework to construct load profiles using Smart Metering time series in an effort to tackle the challenges energy providers face. In addition, we also present an experimental evaluation of our framework using real-world data. Chapter 5 then further builds upon this framework by introducing the concept of *Online Data Mining* techniques to help keep load profiles up-to-date with changes in consumer behavior as new Smart Metering data becomes available.
Finally, chapter 6 concludes this thesis with a short summary and a brief discussion of potential future work.

2

Background

During major parts of this thesis, the main focus of our attention are approaches to current challenges in the energy economy involving time series data. For this purpose it is important to understand the requirements and procedures of business processes which are in practice today as well as the general setup of the electricity grid. Because of this, we give an introduction to the fundamentals of these areas over the course of this chapter, which is to be understood as the first step in the process of *Knowledge Discovery in Databases*, where relevant application knowledge is collected in order to gain a sufficient understanding of the application domain. The overview we give in the following sections concerns current challenges of utility companies and energy providers as well as approaches implemented at present in the industry to tackle these problems; we introduce the state-of-the-art in academic literature in later chapters of this thesis. For the most part, the business processes presented in this chapter primarily relate to electricity companies in Germany; however it is possible that processes in other regions are also affected.

2.1 Overview of the electricity grid

One of the most fundamental requirements within the energy economy is the construction and maintenance of infrastructure which allows producers to deliver energy to the consumers. At the same time, the infrastructure must account for physical laws, imposing certain constraints on the way energy can be distributed. To overcome these technical limitations, engineers have elected to build the electricity grid using a tiered structure of different voltage levels, for which an overview is given in figure 2.1. The main idea of using different voltage levels for the various abstract layers of the electricity grid is that in order to deliver energy across large distances, high voltage lanes are preferable. This is due to the fact that increasing the voltage allows for a decrease of the electric current, which helps in reducing the amount of grid losses due to the cables heating up. At the same time, high voltage lanes require more sophisticated isolation to prevent short circuits caused by arcing. This makes them impractical to use all

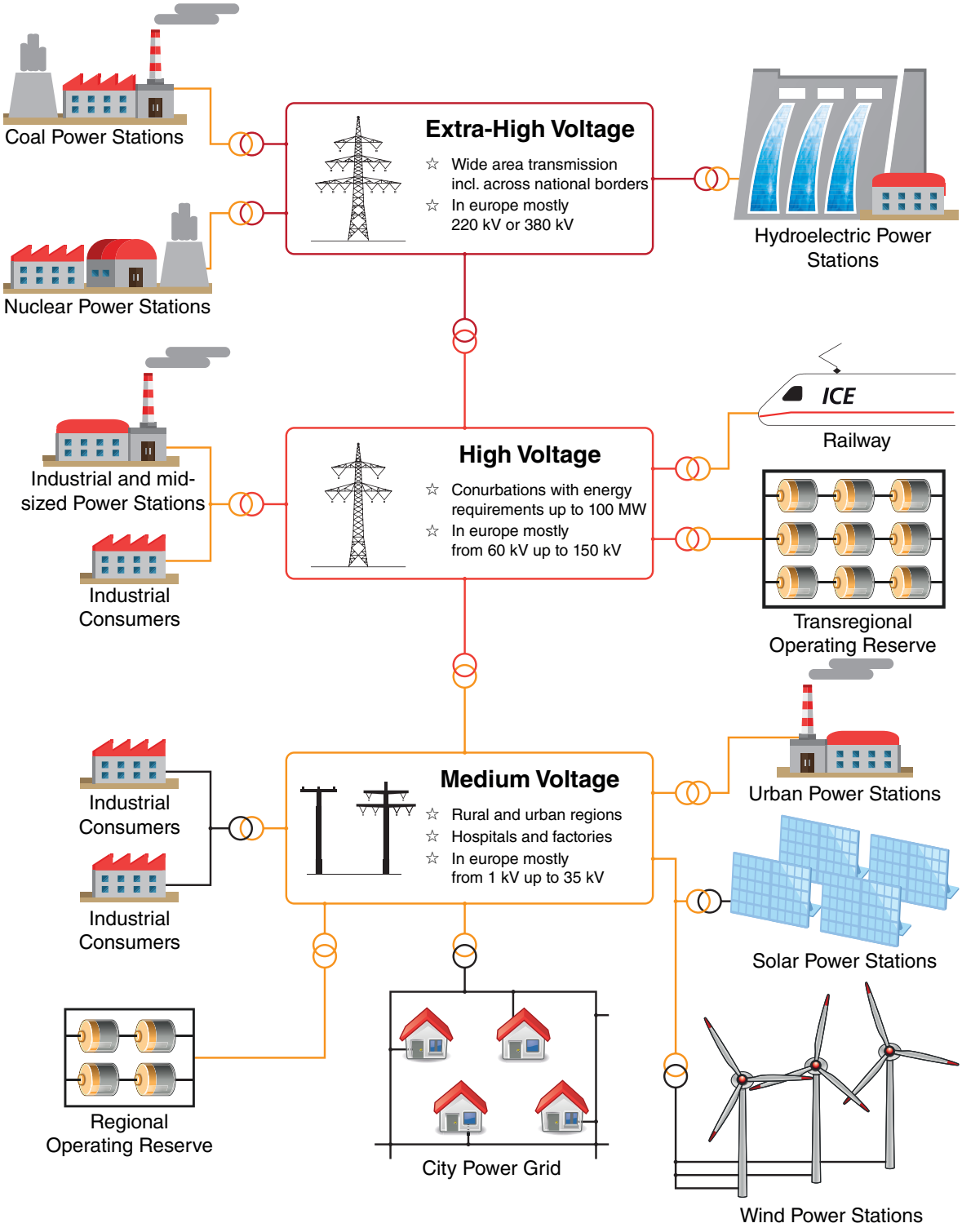


Figure 2.1: Overview of the general layout of electricity grids in europe. *Extra-High Voltage* is notated by dark red lines, *High Voltage* by bright red lines, *Medium Voltage* by yellow lines and *Low Voltage* by black lines. Interleaving circles depict transformer stations. Adapted from [Wik].

the way up to the house or building complex of the customer, particularly in densely populated areas, where a lack of space between cables might lead to short circuits in case the cable isolation is damaged.
In general, the higher the voltage, the larger the area that part of the electricity grid aims to cover. While small, urban power stations usually reside in the medium voltage network, transregional power stations with a large energy output are typically connected to the high voltage or extra-high voltage network.

Until the end of the 1990s, public utility companies had been responsible for both operating and maintaining the electricity grid as well as managing corporate sales. That is, utility companies were purposefully given a monopoly for the region they had been responsible for. This had begun to change with the *Directive 96/92 of the European Parliament* [Par96], which has laid the foundation for the liberalization of the energy market. In the case of Germany, the liberalization was legally finalized as part of the *"Gesetz zur Neuregelung des Energiewirtschaftsrechts" (Law for the revision of the energy economy rights)* [Bun98] in 1998.
Since the liberalization of the energy market, consumers are able to freely choose their energy provider, allowing for grid operators and third party energy providers to be in direct competition with each other. At the same time, even when customers have chosen a third party to be their energy provider, physical lane circuitry still required customers to be served by their local utility company. In order to prevent unfair advantages for the utility company, the liberalization of the energy market also required utility companies to have their grid operation department and their marketing department to function separately and independently. Because of this, customers are able to mandate the energy provider of their choice and pay only an electricity bill according to the prices of their chosen energy provider, while the physical delivery of energy is still conducted by their corresponding local utility company.
As a consequence of customers from different regions being able to require their local grid operator to cooperate with an arbitrary third party energy provider, there has been a need for business processes of all grid operators and energy providers to be compatible to one another to satisfy legal obligations. In Germany, this standardization has been the task of the *BDEW*, which is short for *"Bundesverband der Energie- und Wasserwirtschaft" (Federal association of the energy and water economy)*. For the electricity market, the applicable policy for the business processes themselves is the *"Marktregeln für die Durchführung der Bilanzkreisabrechnung Strom" (MaBiS)* [Bun13], while the technical format for the market communication is documented as a subset of the *UN/EDIFACT* standard called *EDI@Energy* [EDI].
Since business processes and market communication are standardized, notable changes are only possible by new revisions of the *MaBiS* standard, which are then required by all market participants in Germany to be implemented almost simultaneously, with respect to an adequate transition period. Depending on the concrete implementation of the *MaBiS* standard, changes can be very complex and costly. Because of this, we will present some of the most important challenges of energy providers and current solutions within the confinements of the *MaBiS* standard in the following sections.

2.2 Current challenges of energy providers

The most important task of energy providers is to ensure the security of the energy supply, laying down an indispensable necessity of the very foundation of modern society. In order to achieve this goal, it is essential to predict the aggregated energy consumption time series of all customers, that is, the amount of energy the customer base as a whole consume at a given point in time, as accurately as possible. On the basis of this forecasted consumption time series, energy providers allocate production capacities from energy producers at an early stage, giving energy producers enough lead time to ramp production up or down so that energy is injected into the electricity grid exactly according to the consumption time series predicted and announced by the energy provider. This procedure is necessary due to the fact that adjusting the production rate of energy takes time, which is dependent on the type of the power station. While hydroelectric power stations can adjust their production within seconds, most other power plants, such as coal or nuclear power stations, require a startup time of multiple hours up to several days. Because energy is conserved in a closed system, special precautions need to be taken to protect sensitive industrial and consumer electronics from damages caused by undervoltage or overvoltage. This is one of the reasons why the electricity grid need to be balanced at all times, thus requiring the forecast of the total energy consumption compiled by the energy providers need to match the actual consumption as closely as possible, with deviations by overestimating and underestimating the actual load being equally undesirable.
However, since the consumption forecast by energy providers is, by definition, merely an estimate of the future actual consumption, forecast errors are inevitable. These forecast errors require adjustments in real-time, which is referred to by the industry as *imbalance energy* or *operating reserve*, in order to keep the electricity grid balanced. Any technology that is able to both absorb and provide energy within seconds can function as imbalance energy, such as parallel-connected batteries at large-scale or pump-driven water containers which store and retrieve energy by converting between electricity and potential energy. Due to their limited availability, their increased wear and tear because of constant readjusting, their reduced energy efficiency in order to prioritize response time, as well as their importance to keep the electricity grid balanced, imbalance energy is usually much more financially volatile than regular energy. On an abstract level, imbalance energy can be understood as a battery which aims to stay at 50 percent charge status and is installed between energy producers and consumers; depending on whether the energy provider has under- or overestimated the actual total energy load of customers, the battery charges or discharges accordingly to make up for the difference. As depicted in figure 2.1, based on whether the provider of imbalance energy is regional or transregional, the provider typically resides in either the medium voltage or high voltage electricity grid.

In order to minimize the amount of imbalance energy that has to be injected into or extracted from the electricity grid, the forecast models used by energy providers need to both be accurate and provide long-term predictions about the consumption behavior of customers. This enables the energy providers to take a long view on their future buy-in of energy, allowing for more attractive terms in cooperation with energy producers. As a rule of thumb, the more uniformly and smoothly the total energy consumption changes, the easier it is both to adjust to those changes by requiring less

imbalance energy and to forecast the energy consumption beforehand. While for most of the time, prices for imbalance energy are one order of magnitude more cost-intensive than regular energy, in cases where the operating reserve is almost depleted, but also for other reasons, it is possible for the prices of imbalance energy to be multiple orders of magnitude higher than for regular energy. For example, on the 17th October 2017, the prices of imbalance energy have reached an all-time high with a price of 24.455,05 € per MWh of energy, a huge step up compared to the prices of regular energy which are typically within the price range of 30 to 60 € per MWh of energy for *Day-Ahead Auctions*, causing controlling authorities to intervene [Bun18]. As a consequence thereof, sudden, abrupt changes in the total energy consumption time series which cause imbalance energy to be used are generally undesirable to energy providers due to the financial risk associated with them.

In practice however, the total energy consumption is only known in hindsight and electricity meters of customers are usually read only once per year as part of the annual accounting. Although the total energy consumption time series is in fact the only requirement of energy providers in order to plan the future buy-in of energy, the absence of detailed knowledge about the consumption behavior of customers leads to the absence of knowledge about consumption causes of which the total energy consumption time series is composed of. In particular, since in most cases only one meter reading from each customer is available per year, there is no way of differentiating between uniform and peak consumption. Aside from one annual meter reading, in most cases the energy provider only knows contractual details about the customer such as his or her name and the full postal address. High resolution measurements of the consumption behavior of individual customers as well as an intensive, continuous communication so to cater to their specific needs are often economically justifiable only for very large customers, that is, customers with an annual consumption of at least 100.000 kWh of energy.

The resulting lack of insight about the consumption behavior of customers is in stark contrast to the amount of information that online merchants and service providers, as well as retailers using campaigns such as *Payback*, have gathered on their customers, which can then be analyzed using techniques such as *Association Rule Mining* that one may derive customer habits and preferences [Sch12]. Although, for billing purposes, having only one meter reading per year is sufficient, assuming that the retail energy price is constant, differentiating between uniform and peak consumption is necessary to plan the buy-in of energy and avoid imbalance energy as much as possible. It should be noted however that even if the past consumption behavior of a given customer is known, his or her behavior might change the following day. For example, it is intuitively plausible for employed end-consumers to have a different daily routine, and thus a different consumption behavior, on a working day than during the weekend. This factor can possibly cause the base load to change significantly or for some load peaks to not occur at all, occur pronounced differently or at different times of the day. Furthermore, customers can genuinely differ in their consumer behavior due to the customers practicing different hobbies, having made different lifestyle decisions, being either a single or a family household, being a low income or a high income household, or being either an end-consumer of a business, among other factors. The accurate handling of the superimposition of all distinct unknown requirements of all customers is one of the main challenges of an energy provider when planning the future buy-in of energy. The

goal of finding a feasible solution to this problem is a core component of this thesis. For this purpose, we introduce the concept of *load profiles* in the following section. Currently, load profiles are the main tool of energy providers in order to tackle the problem of forecasting the energy consumption. Based on this, we will then elaborate on our optimization approach in chapter 4 and 5 of this thesis.

2.3 Structure and purpose of load profiles

The forecast of the energy consumption and thus the planning of the future buy-in of energy in order to overcome the difficulties outlined in the previous section is typically achieved using so called *load profiles*. Load profiles are designed to offer a way to differentiate between uniform and peak consumption along with the possibility to cope with variable daily routines of customers. To accomplish this goal, additional knowledge about customers or, if such knowledge is not available, additional assumptions about customers need to be incorporated. Over the course of this section, we will give an overview of the mechanics of *SLP* load profiles, the standard forecast model of the energy economy as outlined by the *"Marktregeln für die Durchführung der Bilanzkreisabrechnung Strom" (MaBiS)* industry policy [Bun13].

In essence, load profiles consist of a set of *customer groups* and a segmentation of all calendar days for a year into so called *day-types* as well as a consumption pattern for each combination of a customer group and a day-type. The basic idea here is that while customer groups offer a way to address the needs of distinct customers, such as businesses having different use-cases and energy requirements than home consumers, day-types allow to define temporal intervals during which the consumption behavior of customer is significantly different than for other day-types. What is important to note here is that load profiles, as implemented by the energy economy, have a strict periodicity of 1 year. Because of this, special events that happen more rarely than 1 year or events whose appointed date is not known 1 year in advance, but which possibly have a significant influence on the consumption behavior of customers, such as the *FIFA World Cup* or other major sports events, can not be modeled using load profiles. Figure 2.2 depicts some of the *Standard Load Profiles* as provided by the BDEW. As shown in figure 2.2 and in compliance with intuitive understanding, private households and businesses have significantly different consumption behaviors. While the industrial load profile *G0* expects the energy consumption of businesses to strongly correlate with business hours, the consumption of private households is much more spread out according to the load profile *H0*, with two noticeable consumption peaks at noon and in the evening and a drop in energy consumption during the afternoon. By assigning a load profile to each customer, energy providers do not presume that every customer behaves exactly as expected by the consumption pattern of the load profile; instead, energy provider merely assume the consumption pattern of a given load profile to be representative for all customer assigned to that customer group, with deviations between individual customers and the consumption pattern of their assigned load profile to balance out.

It should be noted however that the load profile themselves are normalized, meaning they do not directly predict the amount of energy required by a given customer, but merely the expected consumption behavior. In order to derive an actual forecast for a given customer, load profiles require additional external input in the form of the

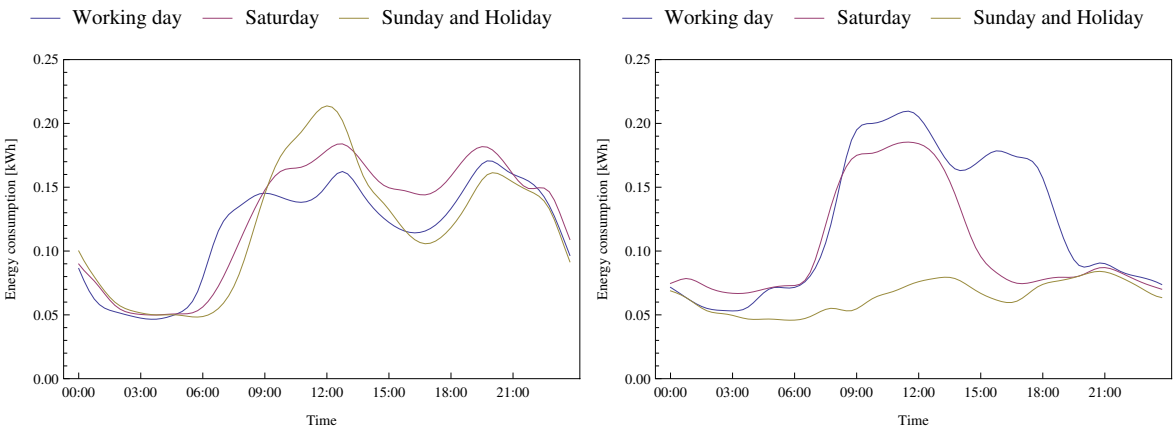


Figure 2.2: Overview of the Standard Load Profiles (a) *H0* (household profile) and (b) *G0* (general purpose industrial profile) from the *BDEW*, valid only during summer. The time series shown are normalized and describe the expected consumption behavior for the day-types *Working day*, *Saturday* and *Sunday and Holiday*.

so called *Year Consumption Forecast (YCF)*. The Year Consumption Forecast is a customer-specific property assigned by the energy provider and equals to the amount of energy the corresponding customer is expected to consume over the course of 1 year. Since the energy provider does know 1 meter reading per year of every customer as part of the annual accounting, most energy providers derive the Year Consumption Forecast of their customers for the following year by computing the actual energy consumption of the previous year as follows:

$$\begin{aligned} YCF_{2016} &= \text{actual total consumption in 2015} \\ &= \text{reading}\,(01.01.2016) - \text{reading}\,(01.01.2015) \end{aligned} \tag{2.1}$$

Even though the formula in equation 2.1 is possibly the most used method to define a Year Consumption Forecast, depending on the choice of the individual energy provider, other options for the definition of a Year Consumption Forecast can be used. For example, instead of using just the actual energy consumption of a given customer of the previous year as a prediction for the energy consumption of the same customer of the following year, a weighted average over the last couple of years can be used. In addition, without loss of generality, equation 2.1 can also be applied if energy providers opt to perform the meter reading not at the beginning of the year, but rather perform the meter reading of all customers over the course of the year in a staggered manner. If both the load profile and the Year Consumption Forecast of a customer are known, the forecast of that customer's energy consumption $E_i\,(t)$ is computed as follows:

$$E_i\,(t) = YCF_i \cdot \frac{E_{\text{normalized profile}}\,(t)}{1.000.000\ \text{kWh}} \tag{2.2}$$

The denominator of 1.000.000 kWh in equation 2.2 is a consequence of the load profiles being normalized; according to [Bun13], load profiles have to be tuned so that $E_{\text{normalized profile}}\,(t)$ adds up to 1.000.000 kWh over the course of a year chosen by the authors of the load profile. Subsequently, however, the normalization of a load profile

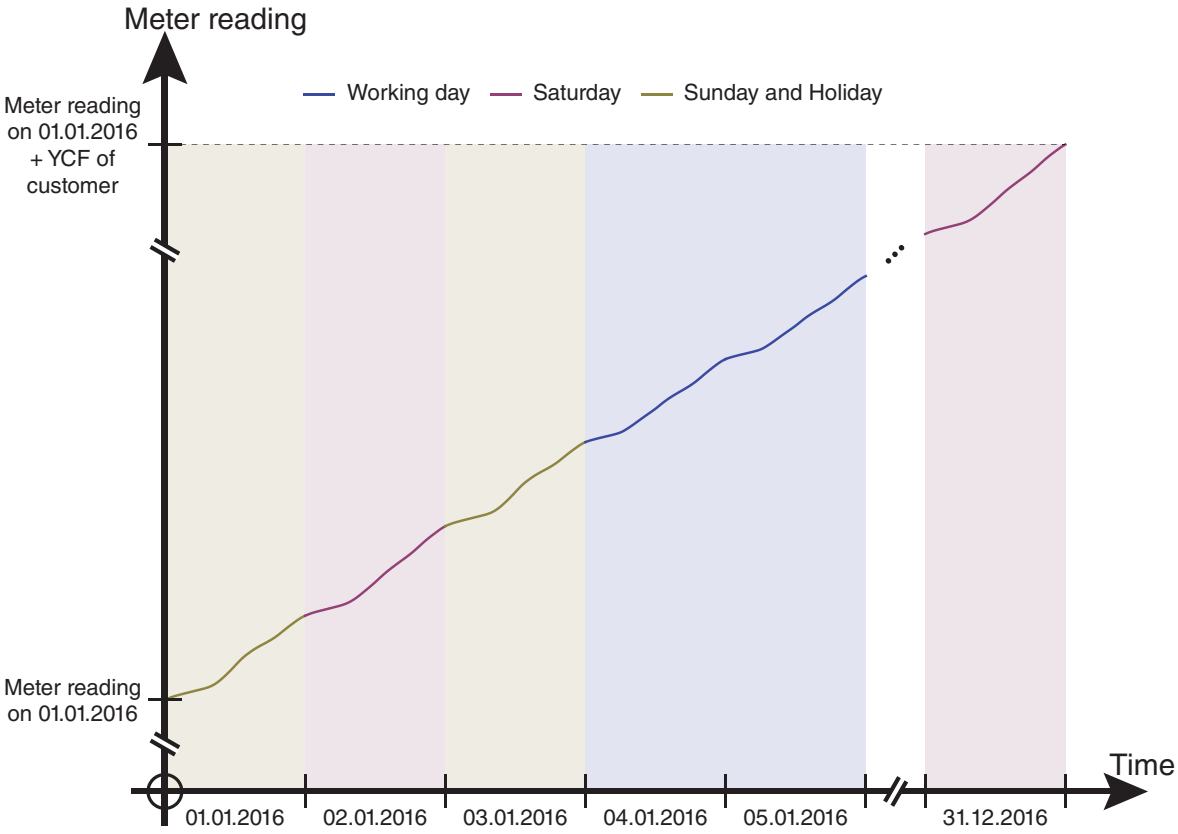


Figure 2.3: Depiction of how load profiles are used to forecast the electric meter reading of a given customer. The normalized time series of a load profile are concatenated based on the day-type segmentation and integrated over the course of one year to form a time series describing the normalized meter reading of the customer, which is then scaled based on the *YCF* to yield a forecast for the actual meter reading of the customer.

is not readjusted for the following years, meaning that depending on the amount of holidays or the current year being a leapyear, $E_{\text{normalized profile}}(t)$ may not add up to exactly 1.000.000 kWh each year. The concept behind equation 2.2 is also visualized in figure 2.3. Since load profiles are normalized and use the Year Consumption Forecast as a scaling factor, they can be understood as a weighting function, where the load profile itself does not directly determine the amount of energy the customer will consume, but rather outputs when the customer will consume what percentage of his or her annual energy consumption as given by the Year Consumption Forecast. In particular, if the actual annual energy consumption of a customer has changed compared to the previous year, the difference between the actual annual energy consumption and the Year Consumption Forecast will result in a forecast error in the form of a multiplicative offset which will remain until the customer's Year Consumption Forecast is readjusted by the energy provider, for example as part of the next annual meter reading.

While the dependency on an accurate Year Consumption Forecast is one of the biggest weaknesses of load profiles, their main advantage is that load profiles allow the total energy consumption to be forecasted 1 year in advance. For this purpose, energy providers repeat the steps outlined above for each customer and compute the sum time series of the individual customer forecasts. This time series is called the *SLS*

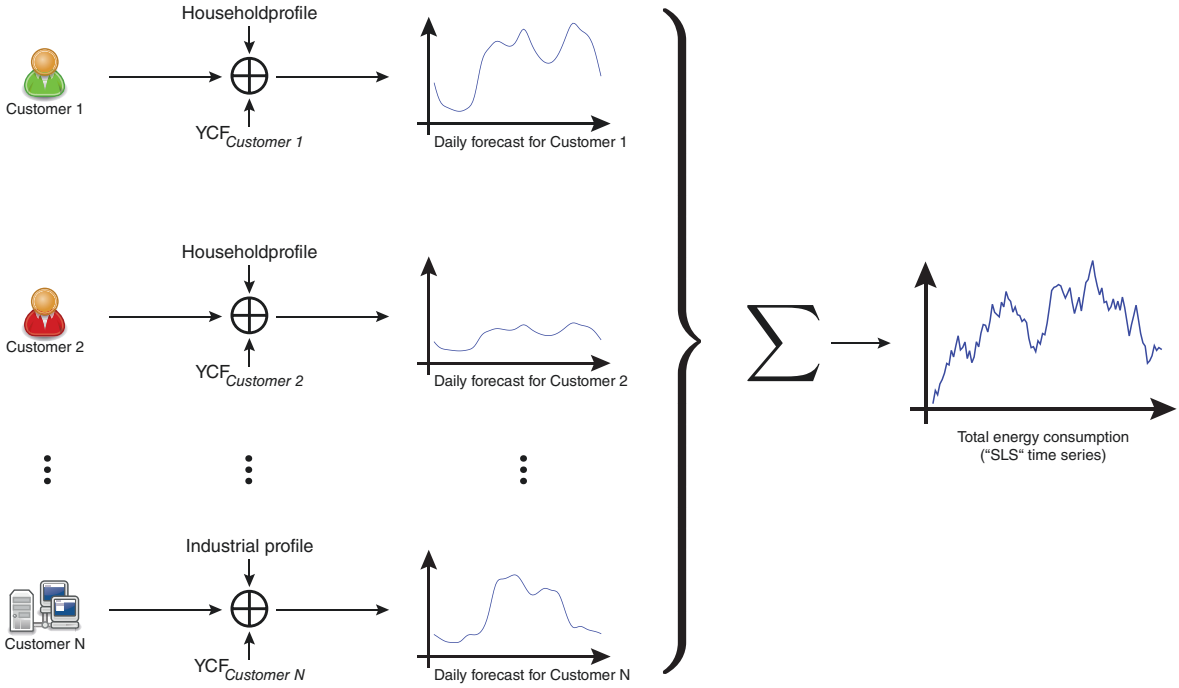


Figure 2.4: Schematic representation of how the forecast of the total energy consumption is derived based on the customer base of an energy provider, their individual *YCF* as well as the load profile assignments.

time series and aims to be as equal to the actual total energy consumption as closely as possible. Energy providers which plan their future buy-in of energy will therefore allocate capacities from energy producers according to the *SLS* time series for the most part. Figure 2.4 visualizes the computation of the *SLS* time series. Due to the fact that load profile models the total energy consumption as the sum of individual consumption time series, load profile offer a straightforward way to handle the business process of customers switching to or from a different energy provider by simply computing the *SLS* time series as the sum of more or fewer customers.

While general purpose electricity usage as predicted by the *SLS* time series is by far the most critical component of forecasting the energy usage of customers, it is noteworthy that some aspects of energy usage are reliant on external factors. For example, the decision of customers to turn on the heating is typically made depending on the temperature and possibly other weather conditions. In contrast, external factors such as the temperature can not be accounted for with *SLP* load profiles as outlined above. Because of that, if energy providers know that a given customer owns a night storage heating or other devices whose usage is likely dependent on weather conditions, the energy providers may opt to employ special-purpose load profiles, so called *TLP* load profiles, in addition to the *SLP* load profiles [Bun13]. These *TLP* load profiles are not able to predict the energy consumption of customers far into the future as they require weather conditions, such as the temperature, as input. Since the contribution of *TLP* load profiles to the total energy consumption forecast is typically very small compared to *SLP* load profiles, they are sometimes deemed negligible and therefore not employed in the first place. For a more detailed introduction to *TLP* load profiles, we direct the interested reader to [Bun13].

Up to this point we described how load profiles work in helping to forecast the energy demand of customers, assuming they are chosen carefully and contain day-type and customer group segmentations that accurately represent the customer behaviors. Traditionally however, energy providers do not have detailed, up-to-date knowledge about their customers to ponder what load profile best suits a given end-consumer of business. Instead, the segmentation of customer groups has been based on conjectures; for example, it seems intuitively plausible that private households and businesses have different energy requirements, causing a segmentation of private households and businesses in different customer groups to likely be reasonable. In the case of Germany, most energy providers use the *Standard Load Profiles* as published by the BDEW, which have been ordered from the academic chair for energy economy of the *Brandenburg University of Technology* during the 1990s [Mei+99]. To compile the *Standard Load Profiles*, the researchers of [Mei+99] resorted to field measurements of approximately 1.500 low voltage customers in cooperation with a selection of energy providers. Some of the resulting load profiles are shown in figure 2.2. As for the *day-type* segmentation, the *Standard Load Profiles* differentiate between *working day*, *saturday* and *sunday and soliday* for each of the seasons *summer*, *winter* and *interim period*, with the corresponding customer groups being introduced in table 2.1. According to the researchers, the *Standard Load Profiles* are applicable to customers with an annual consumption up to 30.000 kWh or a peak demand of 30 kW of energy [FT00], but are often used for customers with an annual consumption up to 100.000 kWh.
In addition to the concept of *day-types* and *customer groups*, the *Standard Load Profiles* also define a so called *dynamization function*. The idea behind such a function is to account for reoccurring patterns in the total energy consumption time series with a periodicity of 1 year, such as the sine-shaped patterns visible in the datasets introduced in section 4.1. In the case of the *Standard Load Profiles*, the dynamization function proposed by the BDEW is as follows:

$$\begin{aligned} DynFactor\,(doy) = &- 3{,}92 \cdot 10^{-10} \cdot (doy)^4 + 3{,}2 \cdot 10^{-7} \cdot (doy)^3 \\ &- 7{,}02 \cdot 10^{-5} \cdot (doy)^2 + 2{,}1 \cdot 10^{-3} \cdot (doy) + 1{,}24 \end{aligned} \tag{2.3}$$

Here, *doy* corresponds to the *day of year*. By incorporating a dynamization function into the forecast of the energy consumption, energy providers are able to continuously distribute the energy allocation over the course of a year, increasing the energy allocation during winter and at the same time decreasing the energy allocation during summer. As a consequence, succeeding calendar days can have slightly different energy forecasts even if they belong to the same *day-type*. Thus, if a dynamization function is used, equation 2.2 is modified as follows:

$$E_i\,(t) = YCF_i \cdot DynFactor\,(doy\,(t)) \cdot \frac{E_{\text{normalized profile}}\,(t)}{1.000.000\ \text{kWh}} \tag{2.4}$$

In principle, an energy provider may choose their preferred dynamization function freely, as the determined dynamization function is broadcasted to other market participants as part of the normal market communication [Bun06]. When deciding for a dynamization function however, it is important to tune it such that the average value of the dynamization function over the course of a whole year is equal to 1 to avoid the overall amount of energy allocated, as predicted by the sum of the Year Consumption Forecasts of the customers, from being distorted.

Profile	Description	Target audience
H0	Private households, but also minor commercial needs	End-consumers, commercial agents, not applicable to thermal heat pumps and thermal storage heating devices
G0	General purpose industrial profile, defined as a weighted mean of all commercial customers	Assigned if none of the profiles *G1* to *G6* apply
G1	Industrial profile for businesses which operate on weekdays from 8 to 18 o'clock	Offices, workshops, kindergartens, public administration facilities, doctor's office
G2	Industrial profile for businesses which operate mostly during the evening	Street lights, gas stations, evening restaurants and recreational facilities (if their peak consumption is not during the weekend)
G3	Industrial profile for continuous, relatively uniform demand, including a noticeable, continuous peak demand	Purification plants, drinking water pumps, communal facilities in residential complexes, cold storage warehouses
G4	Industrial profile for businesses which are mostly dependent on business hours	Shops, hairdresser
G5	Industrial profile for bakeries with bakehouses which typically start operating at 3 o'clock during weekdays and at midnight on saturdays	Bakeries with bakehouse
G6	Industrial profile for businesses with a strong consumption focus during weekends	Youth clubs, cinemas, restaurants, petrol stations
L0	General purpose agricultural profile, defined as a weighted mean of all agricultural customers	Assigned if the energy provider does not differentiate between agricultural customers according to the profiles *L1* and *L2*
L1	Agricultural profile for dairy farms and sideline stockbreeding businesses	Dairy farms, sideline stockbreedings farms
L2	Agricultural profile for businesses with a mixture of household and farming	Assigned if neither *L1* nor time-of-the-day-independent industrial profiles apply

Table 2.1: Overview of the *BDEW Standard Load Profiles* used by most German energy providers and their common use-cases as defined in [Mei+99; FT00].

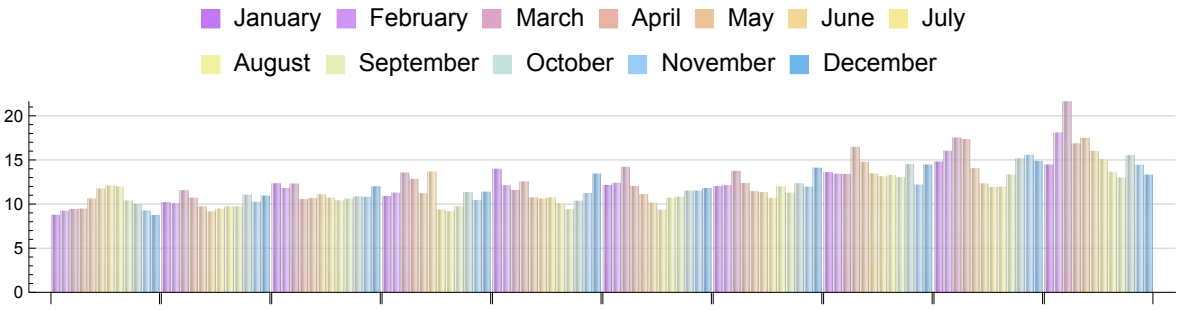


Figure 2.5: Comparison of the ratios of the absolute difference between the actual consumption and the predicted load yielded when using the *BDEW Standard Load Profiles* to the actual consumption in percent for 10 randomly selected energy providers in Germany. The monthly ratios shown correspond to the year 2017. The energy providers are sorted ascendingly according to their yearly average deviation ratio.

Although the *Standard Load Profiles* as published by the BDEW have played a vital role in securing the stability of the energy supply since they have come into effect, studies have shown that, at least for the load profiles for the gas economy, the current load profile have become an increasingly bad model to forecast the energy demand of customers [Roo+14]. Possible reasons for this trend include social changes as well as technological advances in recent years. In the case of the *BDEW Standard Load Profiles*, an evaluation of the performance of the load profiles can be done using freely available data in accordance with disclosure requirements for energy providers exceeding a certain size due to the *"Verordnung über den Zugang zu Elektrizitätsversorgungsnetzen" (Electricity Supply Grid Access Act)* [Jus05]. Using a random sample of energy providers, depicted in figure 2.5, the data suggests that in practice the load profiles yield an average forecast error of approximately 12,21 percent in relation to the actual total energy consumption. Other researchers mention a forecast error 13,67 percent using the *BDEW Standard Load Profiles* [SM17]. Though regional and possibly other factors likely play an important part in whether load profiles fit one customer base better than the other, in general there is the risk of the accuracy of the *Standard Load Profiles* to worsen in the future due to upcoming changes in the consumption behavior of customers, for example by an increase in electromobility, unless adequate countermeasures are taken [Wüs17b; Wüs17a]. Aside from gathering new data to adjust existing consumption patterns of load profiles it begs the question whether the segmentation of *day-types* and *customer groups* themselves should be reconsidered. For example, [Sil+05; Chi+01] indicate that customers may have very different consumption behaviors even within the same industry branch. Their research suggest that category based customer groups, such as the industrial load profiles *G0* till *G6* in table 2.1, are deprecated and should be replaced by *macro categories* such as *residential*, *industrial*, *commercial*, *electric lighting* or *traction* [Chi12]. Due to the growing availability of Smart Metering technology and increasing computational power, this issue can be tackled using techniques from *Knowledge Discovery in Databases*. When evaluating segmentations of *day-types*, it is important to recall that load profiles are designed to primarily forecast the total energy demand as accurate as possible. Thus, a good *day-type* segmentation partitions the calendar days such that the daily total energy demand of calendar days belonging to the same *day-type* are as similar as possible

while aiming to make the total energy demand on calendar days belonging to different *day-types* as dissimilar as possible. Analogous, and in accordance with intuitive understanding, the same argument holds true for *customer groups*. The task description for such a good segmentation closely coincides with the goal of generic clustering algorithms as we have introduced in section 1.2.2. In chapter 4 and 5, we will get back to this problem and present solutions approaches.

3

Knowledge Discovery in the energy economy

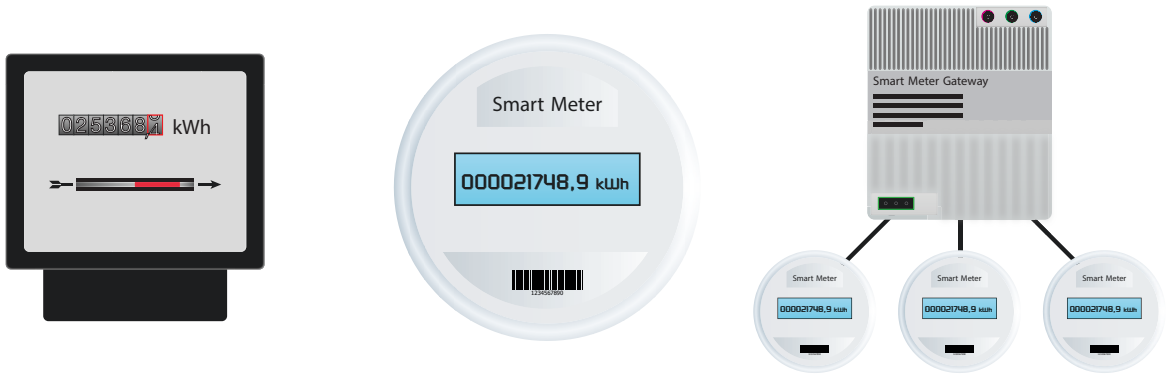
Knowledge Discovery in Databases (KDD) is applied in many areas of academic research and decision-making processes in order to analyze huge amounts of data. In section 1.2, we have given an introduction to the principles of the KDD-process. Over the course of this chapter, we present concrete approaches from academic literature on how data collected by the growing digitization can be used as a chance for both utility companies and consumers via KDD-techniques. In doing so, we focus on 3 main aspects in the energy economy in conjunction with digitization while at the same time keeping the requirements as outlined in chapter 2 in mind:

- Savings benefiting consumers, especially through a more *energy-aware usage of electricity*, resulting in a *reduction of energy cost*.
- Modifications of the electricity grid itself (*smart grid*), which is of particular interest with regard to electric vehicles as this technology puts enormous stress on the energy infrastructure which currently is not designed to be able to handle such a task [Wüs17b; Wüs17a].
- Optimizations benefiting energy providers by being able to better plan their buy-in of energy as well as improving tariff offerings according to customer behavior.

Although we present and discuss the aspects of digitization mentioned above in separate sections of this chapter depending on the entity affected most by these changes, they are *not* to be understood as isolated modules as each aspect involuntarily influences the other aspects in their goals and methods.

3.1 Impact on consumers

End-consumers primarily participate in the digitization of the energy economy through the integration of *Photovoltaic Systems*, but also via the installation of *Intelligent*



Traditional electricity meter

- ☆ mechanical electricity meter, able to measure the consumption of energy
- ☆ unable to be read remotely

Modern Metering Device

- ☆ digital electricity meter, able to measure and store the consumption of energy
- ☆ unable to be read remotely

Intelligent Metering System

- ☆ consisting of one or more Smart Meter devices that are directly connected to a Smart-Meter-Gateway
- ☆ integrated rights management
- ☆ able to process and transmit measurements

Figure 3.1: Comparison of capabilities of traditional electricity meters against *Modern Metering Devices* and *Intelligent Metering Systems*.

Metering Systems, the latter of which consist of one or more *Modern Metering Devices* and one *Smart Meter Gateway* per building complex. Although the term *Smart Meter* has become an accepted usage for the umbrella term covering all aspects of digitized metering, technically it merely describes *Modern Metering Devices*. Modern Metering Devices are solely able to measure and store the amount of energy consumed and do not possess any capabilities involving a sophisticated network stack, to manage access control lists or to process the gathered data. To provide these missing functionalities, one or more *Modern Metering Devices* are required to be directly connected to a Smart Meter Gateway, which are multitenant devices and allow for remote servicing. This enables even big apartment buildings to be managed by a single Smart Meter Gateway. The combination of one Smart Meter Gateway and possibly multiple Modern Metering Devices is called an *Intelligent Metering System*. Figure 3.1 provides an overview of this set of facts.

To guarantee that *Smart Metering* devices are rolled out in a timely manner, the European Parliament has enacted that at least 80 percent of all consumers must be equipped with Intelligent Metering Systems by the year 2020 provided that studies deem the rollout to be economically reasonable [Par09]. One of the main motivations behind the rollout of Smart Metering devices from a consumer standpoint is to improve the amount of information readily available to consumers about their energy consumption behavior. Traditionally, end-consumers have next to no information about their energy consumption behavior aside from their yearly consumption in the form of an annual account on their electricity bill. With consumers only knowing their yearly consumption, they have hardly any way to critically question their own consumption behavior. Therefore, some of the main benefits of Smart Metering for consumers are to raise awareness for their own consumption behavior, to identify major consumption appliances and to increase customer comfort. Ways to accomplish this include the visu-

alization of the customers energy usage in real-time, the possibility of individual tariff offerings tailored to the specific needs of the consumer, faster rate conversions, shorter billing cycles as well as less reading errors and thus less customer complains through automated remote readings. In the case of Germany, the rollout of Smart Metering devices is regulated by law through the *"Gesetz zur Digitalisierung der Energiewende" (Law for the digitization of the energy transformation)* [Wir16; Ene16] as well as the *"Messstellenbetriebsgesetz" (Law for the operation of metering points)* [Jus16] which are implementations of the EU guidelines 2015/1535 [Par15] and 2006/32 [Par06].

Even though the digitization of the energy economy offers great benefits for the end-consumer, they come with a lot of risks attached, some of which are as follows:

- Since Smart Meter Gateways are essentially general purpose computers, they impose the risk of being hacked by a malicious entity, possibly causing a huge *financial loss* for the targeted consumer by corrupting the measurements transmitted to the energy provider or by cutting off the energy supply entirely.

- Because Intelligent Metering Systems by design allow for high resolution measurements of the energy consumption, they also allow for detailed insights into the lifestyle habits of the consumer, including, but not limited to, deducing the time when he or she usually leaves the house or determining the period when the customer has gone on holiday, which is of particular interest to burglars.

- If the typical behavior of each individual customer is known, the energy provider may opt to offer financial incentives (or penalties) to urge the customer to change his or her consumption behavior in a certain way, by which the consumer might feel being patronized.

In contrast to these fears of customer advocates are statutory provisions such as article 8 of the *Convention for the Protection of Human Rights and Fundamental Freedoms* by the *Council of Europe* which warrant the right of privacy to each individual [Eur50]. These requirements constitute a great responsibility for legislators, market participants and device manufacturers to be mindful of data and device security as well as privacy demands. The security aspect so far has been countered by requiring all parties participating in controlling Intelligent Metering Systems to be certified according to *ISO/EIC 27001* [Sta13] as mandated by §25 of the aforementioned *"Messstellenbetriebsgesetz"* [Jus16]. In addition, market participants are required to secure all communication regarding Smart Metering devices by using cryptographic functions compliant with the technical guidelines *BSI TR-03109-2* and *BSI TR-03109-1* [Sic].

Even though the requirements as described above cover the most important criticisms of Intelligent Metering Systems, we will discuss the privacy aspects in more detail in section 3.1.2. In that section we will also present examples on why the security and data privacy controversy is by no means conclusively clarified and should be openly discussed further in order to prevent the failure of Smart Metering due to consumer mistrust. Additionally, we will present some of the possibilities on how the end-consumer can benefit from the digitization of the electricity grid in the following sections.

3.1.1 Financial benefit and influence on consumer behavior

Aside from the criticisms of Smart Metering mentioned at the beginning of this chapter, one of the main concerns of consumers is an increase in cost for energy providers, which is then passed on to their customers. Possible reasons for energy providers to claim an increase in their operating costs include the acquisition costs of the Smart Metering devices themselves, the labor costs associated with the physical installation in the homes of the consumers as well as the perpetual costs for operating and maintaining the devices. In order to protect end-consumers from being faced with unreasonably high additional costs because of Intelligent Metering Systems, §31 of the German law *"Messstellenbetriebsgesetz" (Law for the operation of metering points)* [Jus16] dictates a limit on how much energy providers may charge annually to cover all costs related to Smart Metering. The costs are tiered according to the customer's total energy consumption per year. For typical private households (up to 10.000 kWh per year), customers have to expect an initial increase in cost ranging from 23 to 100 € per year. While doing so, legislators anticipate that the main increase in cost originates from upfront one-time payments, namely device acquisition and deployment. Once these one-time payments have been amortized, advocates expect the benefits of Smart Metering to outweigh the ongoing costs for operation and maintenance, causing the consumer to see a decrease in energy cost compared to prior of the installation of the Intelligent Metering System.

In order to verify that Smart Metering is profitable for the end-consumer in the long run and to quantize that financial benefit, several countries have launched pilot projects and auditors compiled cost-benefit analyses [EK13]. In the case of Germany, auditors estimated in 2013 that 68 percent of all customers can realistically be equipped with either Intelligent Meter Systems or Modern Metering devices by the year 2022 via an average annual cost of 58 € for consumers outfitted with the new technology. This scenario estimates that approximately one third of installations of Smart Metering devices will be Intelligent Metering Systems and two thirds will be Modern Metering devices. Since Modern Metering devices are relatively cheap compared to Intelligent Metering Systems, consumers can be offered to just have an Modern Metering device installed which may later straightforwardly be upgraded to an Intelligent Metering System. This constitutes an economic alternative for customers who do not put emphasis on being connected to the external communication infrastructure. Altogether, this strategy for the rollout of Smart Metering devices would require energy providers to perform a mixed calculation, where at first the providers accept financial deficits by underselling the devices to keep the economic burden for end-consumers low. In contrast, energy providers save money as Smart Metering devices are able to cover functionality required by some *renewable energy power plants* which would otherwise depend upon a further expansion of capacities within the energy grid. Thus, the overall net capital value of the Smart Metering rollout according to this scenario is estimated to be positive 1,5 billion € with the period under review being from the year 2012 until 2032, a strong signal in the eyes of Smart Metering advocates.

In addition to the cost-benefit analysis for Germany, the authors of [EK13] also present Smart Metering pilot projects in 6 european countries, the most important key figures of which have been summarized in table 3.1. In all countries that are part of this study, the net benefit of the end-consumer is positive, with the average net benefit per consumer being approximately 70 €. It is noteworthy however, that the

Country	Rollout period	# metering points	Consumer net benefit
Great Britain	$2014 - 2019$	27 million	3.042 million €
Ireland	$2015 - 2019$	2,2 million	179 million €
Italy	$2001 - 2011$	32 million	*N/A*
France	$2013 - 2018$	35 million	$100 - 700$ million €
Netherlands	$2014 - 2020$	7,7 million	770 million €
Sweden	$2006 - 2009$	5,1 million	230 million €

Table 3.1: Overview of the Smart Metering pilot projects in different european countries. Adapted from [EK13].

statistical dispersion of the net benefit per consumer is very high, possibly depending on demographic and region-specific factors, which makes it difficult to derive the net benefit of consumers in other countries or even to conclude that the net benefit of consumers in other regions will be positive at all.

Aside from the pilot projects mentioned above which mainly focus on the overall cost benefit of the consumer, there have been studies which concentrate on if and how the consumer changes his or her behavior when being provided with certain information. For instance, in [Deg+13] the authors have randomly segmented study participants in Switzerland into groups, each group consisting of roughly 1200 participants. Each group has been provided with one of the following:

- No additional information at all (control group).
- Fine-grained consumption information about the customer's own energy consumption via the installation of Smart Metering devices.
- Free energy counseling to highlight possibilities for the customers to save energy.
- Fine-grained consumption information about the energy consumption of a mutually assigned, group-internal partner.
- Fine-grained consumption information about the customer's own energy consumption as well as information about the energy consumption of a mutually assigned, group-internal partner via the installation of Smart Metering devices.

Comparisons with the control group have shown that the mere installation of Smart Metering devices has caused a lasting reduction of the energy consumption of roughly 3,2 percent which approximately corresponds to 0,2 kWh per day. The field measurements have also demonstrated that customers with Smart Metering devices partly shift their energy consumption into off-peak tariff periods, for example during the night, causing their consumption during the evening hours, where the peak load period takes place, to drop by as much as 8 percent. The consumption behaviors of the customers themselves however has not changed, for example by watching less television, meaning that information provided by Smart Metering helps to improve the efficient usage of energy without diminishing the customer's quality of life. At the same time, assuming a constant electricity price of 0,36 € per kWh, the end-consumer is able to save roughly 0,07 € per day, or 26,28 € annually. In addition to these savings gained by reducing the amount of energy consumed, the aforementioned shifting of consumption into off-peak

tariff periods likely causes the financial gain to be notably greater, possibly resulting in an overall net benefit more in line with the findings of the pilot project described in table 3.1. The precise amount of money that can be saved by shifting consumption depends on the off-peak tariff terms negotiated between the energy provider and the end-consumer. Other sources of information however, such as free energy counseling or information about the energy consumption of a mutually assigned, group-internal partner has not shown to have a significant and lasting impact on the consumption behavior.

3.1.2 Data privacy

Data privacy is a very important and sensitive subject for many end-consumers, with data collected by the energy economy being no exception. On the one hand this encompasses the granularity as well as the sheer amount of data being gathered, but also how the data is being processed and transmitted, for example whether or not the data is encrypted while transferred. In addition, since most data collected and processed within the energy economy, as well as by Smart Metering devices because of the planned digitization, are inherently sensitive personal information, the presence of data protection concepts will become more and more important for market participants. For example, the German energy provider *Yellow Strom* was awarded the negative award *BigBrother 2008* in the category *technology* because of its intent to install Smart Metering devices in the homes of its customers without developing an adequate data protection concept first [Ren08].

To protect end-consumers from unauthorized usage of their data, German legislators have decided and pronounced via §25 of the *"Messstellenbetriebsgesetz"* [Jus16] that all market participants that come into contact with Intelligent Metering Systems and data gathered from them must be certified according to *ISO/EIC 27001* [Sta13]. Furthermore, Smart Metering devices must use cryptographic functions compliant with the technical guidelines *BSI TR-03109-2* and *BSI TR-03109-1* [Sic] to secure all communication. This marks an important step to protect consumer rights and to prevent misuse of data. Malicious exploitation of sensible personal information including data about the consumption behavior makes end-consumers vulnerable to a great extend. If Smart Metering data were to fall into the wrong hands or if legal obstacles are set inappropriately low, the possible ramifications include, but are not limited to, the following [MRT12]:

- Burglars could deduce from the absence of energy consumption over a sufficiently long period that the home of their potential target is currently uninhabited.
- Stalkers would obtain a new tool to monitor their victims.
- Law enforcement authorities could easily check whether or not a subject had indeed been at home at the time of the crime.
- Insurance companies would be able to abruptly *adjust* the tariffs of customers who behave in a way the insurance company deems unpleasant, for example by keeping some devices switched on when leaving the house.

The more fine-grained the consumption data of a customer is, the more information about the consumer can be derived from it and the more invasive these deductions

become, up to a highly personalized profile describing the customers behavior patterns and family status.

3.1.2.1 De-pseudonymisation of customers

In section 3.3, we will discuss approaches on how household properties and lifestyle habits of end-consumers can be deduced by analyzing sensitive personal information such as consumption time series measured by Smart Metering devices. To prevent analyses of customer data for unauthorized purposes, some energy providers implement technical safety mechanisms in addition to the data protection guidelines mandated by law. One of the mechanisms is to store consumption data and personal affiliation separate. In this case, the matching table between the pseudonym and the real customer identity is kept confidential by the energy provider. Thus, in the event of the consumption data becoming compromised, the idea is for an attacker to only retrieve consumption data with random labeling, making it harder for the malicious party to derive the true identity of a potential victim. This safety measure is often reinforced by re-scrambling the labels of the consumption data regularly, for example once each month. Even though this is a well-meant approach to increase customer safety, academic literature has demonstrated a variety of attack vectors, which the authors have called *linking by behavior anomaly* and *linking by behavior pattern* [JJR11]:

- **Linking by behavior anomaly:** This attack vector is applicable if two distinct data sources (with different pseudonymisation) are available. The time intervals of these data sources must be overlapping and within these overlapping time intervals the same *anomaly* must occur in both data sets. An *anomaly* refers to any action by the consumer which leaves a rare, possibly unique signature pattern in the energy consumption time series, for example a craftsman executing a particular assignment or character-specific changes in consumption behavior during holidays.

- **Linking by behavior pattern:** This attack vector can be employed if the pseudonymisation of the data has changes, for example if the customer changes his or her energy provider or the time series labels get re-scrambling as part of data privacy efforts. The idea behind this approach is that the behaviors of customers fundamentally do not change, making lifestyle habits such as early rising recognizable even if the data labels are altered.

In order to de-pseudonymize a given customer, different techniques are employed to perform the attacks briefly described above. For this purpose, let $F_i = f_{i,1}, \ldots, f_{i,n}$ be the Smart Metering consumption time series of the i-th customer with n calendar days of data available. Furthermore, let $f_{i,l} = (m_{i,l,1}, \ldots, m_{i,l,H})$ describe the individual measurements for the l-th calendar day with H total measurements per day.
To perform the *Linking by behavior anomaly* attack, every $f_{i,l}$ is mapped into a binary feature space, visualized by a $g \times H$ grid, where for each time of the day a g-dimensional vector is used to tier the continuous consumption values into discrete bins of ranges. For example, if the resolution of the Smart Metering time series is 1 measurement per hour (which equates to $H = 24$), with the highest consumption reading of any customer being less than 2.000 Wh and the user specifies $g = 100$, then the feature space corresponds to a 2.400-dimensional binary vector $\phi(f_{i,l})$, where each measurement

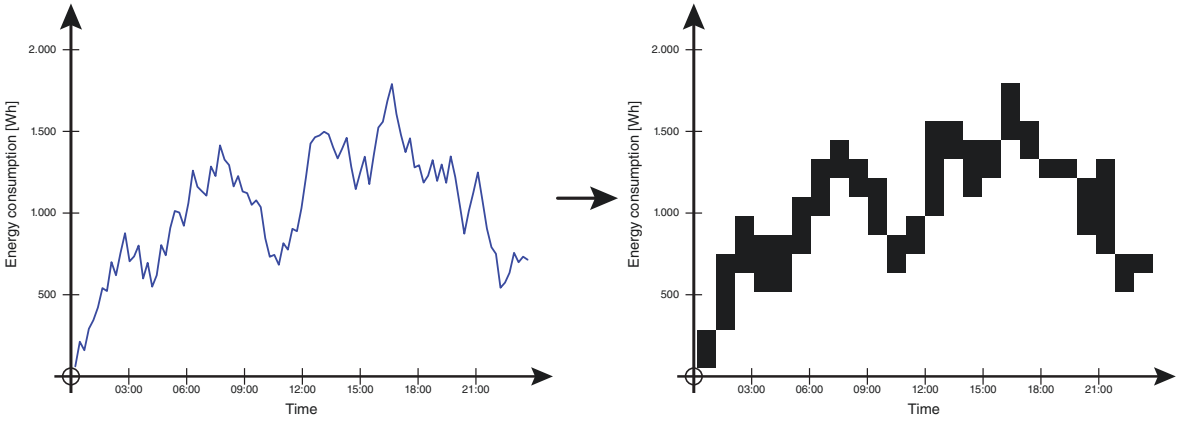


Figure 3.2: Visualization of how a consumption time series (left side) is transformed into a binary feature-vector $\phi(f_{i,l})$ depicted as a grid (right side) as part of the *Linking by behavior anomaly* attack vector [JJR11]. Black colored bins correspond to a value of 1, white colored bins to a value of 0.

is tiered into the ranges $[0; 20), [20; 40), \ldots, [1.980; 2.000)$. Since the parameter g is chosen by the user, it provides the possibility for *parameter-tuning*, meaning that the quality of the experiments is likely dependent on the value of g and should be repeated with different values for g. The process of mapping a given consumption time series into the binary feature space is visualized in figure 3.2. To determine if a feature-vector describes a day where an anomaly occured, the arithmetic average μ of the feature-vectors of consecutive days from a given customer is computed. If the euclidean distance between a given feature-vector and μ is greater than a user-defined threshold, that day is classified as containing an anomalous event. If the same anomalous event occurs during overlapping time intervals from different data sources, the two consumption time series most likely describe the same customer.
To employ the *Linking by behavior pattern* attack, the same feature-vector $\phi(f_{i,l})$ as before is used. As part of this attack however, *Support Vector Maschines* [HPK11] are used to separate the $\phi(f_{i,l})$. To ensure best results, a *one-against-all* approach is used, meaning that for each customer, an optimal hyperplane is determined to separate that customer from any other. Assuming the lifestyle habits of customers do not shift when changing the energy provider or when the energy provider re-scrambles the pseudonymisation, the feature-vectors from different data sources are likely to be assigned the same class using the hyperplanes of the Support Vector Maschines, from which a matching of the pseudonymisation labels can be derived. Using this approach, the authors of [JJR11] have achieved an accuracy of 83 percent with as little as 14 days of training- and test data on a data set consisting of 53 household customers covering 221 days with an hourly resolution of the consumption time series. Furthermore, the authors have shown that these results improve to over 90 percent accuracy for 30 days of training- and test data.
What is important to note here is that these serious privacy implications are already possible with a single consumption time series per customer, which would be classified as *None-Instrusive Load Monitoring (NILM)* according to [Zha+14]. In the case of there being individual consumption time series for each appliance the customer owns, the possible privacy implication could be even more severe, at which point one would

refer to this practice as *Instrusive Load Monitoring (ILM)*.

3.1.2.2 Differential privacy in a Smart Metering environment

Overall, even though the dataset used by the experiments in section 3.1.2.1 is relatively small, it shows how privacy-invasive KDD-techniques can be and thus how important compliance with data protection guidelines is to maintain customer trust. Possible ways to make unauthorized de-pseudonymisation more difficult for an attacker is to drastically lower the resolution of the consumption time series, for example by only storing one measurement per day, or to re-scramble the labels much more frequently. Other possibilities to preserve consumer privacy include methods to smooth out the consumption time series or to mask the consumption pattern with noise in order to minimize the amount of meaningful information that can be extracted from the data, for example by employing *Battery-based Load Hiding (BLH)*, which is sometimes also called *Load Signature Moderator (LSM)* [Zha+14]. In this scenario, the consumer decides to install a battery at home to act as a buffer between the actual energy consumption caused by household peripherals and the consumption time series recorded by the electricity meter. The basic idea here is that since the battery does not alter the total amount of energy consumed, it does not hamper the ability of the energy provider regarding billing purposes. It is however able to charge and discharge in order to mask sensitive consumption behavior, restricted only by physical limitations given by the chemistry of the battery. This chemistry of the battery determines the maximum amount of privacy it can provide, given by its maximum capacity as well as its maximum charge and discharge rate. To accomplish its task, the battery may provide the user the option of several strategies to hide the consumption behavior. These strategies determine what the battery aims the consumption time series to look like *"to the outside world"* and how it may behave when the battery charge approaches its maximum or minimum.
One such strategy is what the authors of [Zha+14] have called the *Best Effort* strategy. Here, the battery produces artificial load peaks (by charging the battery) when the actual demand is low in order to discharge and thus mask or mitigate actual peak loads caused by peripherals of the consumer. In essence, the *Best Effort* strategy corresponds to a low-pass filter, where the battery tries to maintain the externally visible load as being constant as much as possible. However, since the battery comes with physical limitations such as a maximum capacity and maximum charge or discharge rate, usage patterns that push the battery to its limits will cause the externally visible load to once again contain private information. Overall, the authors of [Zha+14] have shown that the *Best Effort* strategy does not provide *differential privacy* [Dwo06].
Differential privacy or more specifically ϵ-*differential privacy* [Dwo06] is a statistical concept that measures the amount of private information that is exposed if a database is asked a given query. To ensure the disclosed information about a private individual does not exceed a user-defined threshold, *random noise* is added to the result of queries. The amount of noise added depends on the selectivity of the query. As a rule of thumb, the more specific a query is, the more noise needs to be added to retain the desired minimum privacy for all individuals in the database. More formally, given two dataset D_1 and D_2 which differ at most in one element, a randomized function $\mathcal{K}$ provides

ϵ-*differential privacy* if for all $S \in Range(\mathcal{K})$ the following condition is met [Dwo06]:

$$\Pr[\mathcal{K}(D_1) \in S] \leq e^{\epsilon} \times \Pr[\mathcal{K}(D_2) \in S] \tag{3.1}$$

In simplified terms, this means the difference of the query-results that would emerge if a single piece of information about an individual was part of the dataset or not (denoted by the hypothetical datasets D_1 and D_2) must remain smaller than a user-defined threshold. This is achieved by not evaluating the query on the actual datasets D_1 and D_2, but on $\mathcal{K}(D_1)$ and $\mathcal{K}(D_2)$, which corresponds to adding noise to the datasets by applying a *randomized function* $\mathcal{K}$. If $\mathcal{K}$ is chosen carefully, this concept provides provable privacy in statistic queries.
In order to translate this concept on the problem of preserving ϵ-differential privacy in a scenario where a Smart Metering device queries the energy consumed in regular intervals, the authors of [Zha+14] propose for the battery to periodically chose a random value of the binomial distribution $B\left(n, \frac{1}{2}\right) - \frac{n}{2}$, where n corresponds to a sufficiently large number, at least as large as the number of queries on the dataset, but within the physical limitations given by the chemistry of the battery. Following that, the battery charges with a rate according to the chosen value if it is positive, or analogously discharges if the value is negative. Since the expected value of the aforementioned binomial distribution is 0, the overall charge of the battery is expected to stay the same, meaning that the battery is able to provide this privacy measure indefinitely.
Overall, the authors of [Zha+14] have presented an effective means to counteract espionage or profiling of customers by deliberately distorting the energy extracted from the energy grid by the end-consumer. Since these methods are about random charging or discharging of a battery with an expected value of 0, this distortions even themselves out if used by a sufficiently large user base, meaning this privacy measure does not hamper the energy providers duty to maintain the stability of the energy grid. However, a possible weakness of this procedure emerges if there is no consumption at all caused by the user over a sufficiently long period, for example if the consumer has gone on holiday. In that case, depending on the strength of the battery, the noise it can provide might be unable to mask that the average consumption has decreased by such an amount. Because of that, the absence of the user might still be visible under some circumstances, which constitutes some valuable information for burglars.

3.1.2.3 Data reduction and data economy

Another approach for privacy is presented in [MRT12]. Here, the authors debate whether the data can remain mostly local while still providing the benefits of having Smart Metering devices installed, eliminating the need for anonymization in the first place. For this purpose, several applications of Smart Metering data are discussed, which we will present hereafter.
The primary application of the energy provider, where having access to Smart Metering data of consumers is a technical necessity, is improving load balancing and the buy-in of energy from the energy producers. One way of improving the buy-in of energy using Smart Metering data is by analyzing the data to build *load profiles* (see section 2.3), for example by performing the experiments we present in chapter 4 and 5. In essence, the authors of [MRT12] argue that while this utilization of the data is ultimately legitimate, with the benefits outweighing the data privacy concerns, it *might* suffice to

be content with just using anonymized consumption time series gathered on an entirely voluntary basis. Among satisfied customers, this would simplify legal questions if energy providers plan to engage the services of cloud providers to put up with the amount of computing power required to apply KDD-techniques to these huge amounts of data.
Another aspect where having high resolution consumption data available is advantageous is *Demand Reduction*. For this purpose, Smart Metering devices usually provide the end user with up-to-date consumption statistics via a build-in display, a smartphone or web application. Using this data, users can get an idea of their consumption, identify high consumption devices and thus specifically focus on reduce their demand. Since smartphone or web applications can work within the local home network of the user, the data does not need to be transferred into the cloud. If the Intelligent Metering System does offer *Remote Feedback*, meaning that data processing does happen in the cloud, the protection of the user's data privacy can consist of *Remote Feedback* being entirely optional, meaning that no data is transmitted outside of the user's home unless the user explicitly consents. This facet of *Remote Feedback* is primarily relevant if the user is either living alone or consists of a family, where mutual trust can be assumed. However, if a building complex hosts a shared apartment or something similar, even local-only feedback can give undesirable deep insights into private matters. In this scenario, a solution might be to allow local feedback only based on aggregated data, such as half-hourly or hourly resolution, which limits the amount of critical information in terms of data-protection law that can be extracted from such data.
Lastly, the authors of [MRT12] discuss the extensive aspect of *Demand Response* in the context of data privacy. Although we present this subject as the last facet of this chapter, it is by far the most significant. This is a consequence of the fact that Smart Metering is considered the great hope of politics for end-consumers to voluntarily adjust their consumption behavior. For example, the *Directive 2009/72/EC of the European Parliament* [Par09] describes Intelligent Metering Systems as the best possibility to *provide energy management services* and *innovative pricing formulas* in order to *promote energy efficiency*. As a consequence, [Par09] recommends the wide-spread roll-out of Smart Metering devices so that they *shall assist the active participation of consumers in the electricity supply market*. Notably, the *active participation of consumers* raises the question of which data must necessarily be made available publicly to the market for this purpose and what the benefits for the consumers are. One of the most relevant aspects for consumers are price advantages. Here, the likely most simple, but also most privacy-unfriendly approach would be to transmit the consumption time series in its highest available resolution to the energy provider. Subsequently, the energy provider could apply *Time-Of-Use Pricing (TOU)*, meaning that the energy price is not constant, but rather dynamically calculated based on customer-specific agreements, the total load of the energy grid, the time of the day and possibly other factors. This scenario is consistent with the *European Commission Recommendation 2012/148/EU* [Com12], which describes *remote reading of meters* as a key functionality and advocates the support for *advanced tariff systems*. A more privacy-friendly approach to dynamically handle different tariffs consists of the Intelligent Metering System managing separate meters for each tariff. In this case, capable Intelligent Metering Systems require only the time intervals describing which tariff is valid at a given time to be told by the energy provider. At the end of the accounting period, only

the cumulative total consumptions readings per tariffs need to be submitted to the energy provider. Additionally, if the Intelligent Metering Systems is able to process dynamic pricing tables published by energy provider, invoicing can be done completely local from a technical standpoint. In this last case, privacy concerns can be avoided completely as the energy provider only learns about the total invoice amount. This concept of keeping data local also corresponds to the suggestion of [JJR11] to prevent de-pseudonymisation as much as possible.

3.1.3 Photovoltaic systems

In 2014, the European Commission has stated that at least 27 percent of the energy consumed should be produced by renewable energy sources by the year 2030 [Com14]. This strong commitment towards renewable energies can also be considered as a signal for investors as it has been feared that investments would level off through 2020 if not stimulated and backed up by corresponding long-term policies [Age14]. Due to technological advances in an effort to make solar power plants more economically sustainable, the price of energy produced by photovoltaic systems continues to decrease: in 2018, a tender for a 300 megawatts solar power plant was won in Saudi Arabia at a price of 0,0234 US-Dollars per kWh [WS18]. Even though this project corresponds to a rollout at a rather large scale, a photovoltaic system can also be economically reasonable for a home consumer without the user being aware of that fact. Because general energy prices are expected to increase over time due to the growing scarceness of fossil fuels, it becomes increasingly important for consumers to have the benefits of photovoltaic systems presented to them. This amplifies the incentive of end-consumers to consider to invest in photovoltaic systems even today. However, one huge challenge for the widespread adoption of photovoltaic systems is that the economic potential differs greatly per customer, which greatly hampers the consumers ability of making an informed choice. This is, among others factors, because of the different east-west orientation of the roofs, foreground objects casting shadows on the roof, roof inclination and roof type. Since the factors are often unknown to private individuals, the assessment of economic feasibility often requires a costly on-site meeting with a professional.

To avoid having to perform large-scale, yet detailed assessments for accurate simulations, some approaches settle with using commonly available statistics like typical sizes of buildings, population density or gross domestic product to evaluate the economic feasibility of photovoltaic systems [Löd+10], while other methods include using small random samples which are then extrapolated to the size of whole regions [Ord+10]. While these approaches are useful for politics and investors to get a rough idea which countries or regions of countries to subsidize or invest in, it is of little help for individual home consumers to decide whether or not the installation of photovoltaic systems is reasonable for them.

In order to overcome this difficulty, research has started to include freely available geospatial information systems, also referred to as *Volunteered Geographic Information (VGI)*, as a source of data. In [Hop+17], the authors make use of these services, especially *OpenStreetMap*, to build a Data Mining based *decision support system* to help predict the solar energy potential of customers. The idea here is to resolve the address of a customer into coordinates using OpenStreetMap and then to derive the floor area A_F by choosing the nearest building polygon for the address coordinates. By

assuming a *gabled roof* with an inclination angle of $\alpha = 35°$ for all homes, which is the most common roof type according to the authors, the usable area A_c for photovoltaic systems can be calculated as:

$$A_c = \frac{1}{2} \cdot \frac{A_F}{\cos(\alpha)} \tag{3.2}$$

In conjunction with the roof orientation angle β, which can be derived using the orientation of the polygons from OpenStreetMap, as well as historical solar radiation and temperature data, an estimate for the solar energy potential can be given. The implementation of these semantics can be simplified using purpose-built software libraries such as *solaR* [Lam12], a framework written in the statistical programming language *R* to perform calculations involving photovoltaic energy, sun geometry and solar radiation.

While the procedure outlined above is a good approach for end-consumers to get a first impression about their solar energy potential, it is noteworthy that the estimate is highly dependent on the accuracy of the geospatial data used. Furthermore, the approach does not take into consideration that only parts of the roof may be usable due to roof windows, chimneys, antennas or foreground obstacles casting shadows. In this regard, computer ray-tracing simulations in 3 urban sites in Switzerland have shown that the average usable area of the roofs for photovoltaic systems is ranging from 49 to 95 percent [WNP10]. In addition, evaluations of this OpenStreetMap-based approach with validation data in 3 German cities have shown that the average deviation in estimating the usable roof area is $-9{,}6m^2$ for private households, meaning that the OpenStreetMap-based approach tends to underestimate the usable area. With the absolute deviation for gabled roofs being $20{,}14m^2$, this corresponds to an average prediction error of 27 percent. However, since this approach assumes gabled roofs with an inclination angle of $\alpha = 35°$ for all buildings, the estimation error is much higher for non-gabled roofs. Overall, the OpenStreetMap-based approach achieves an average estimation error of 55,15 percent when considering all roof types of private homes.

In conclusion, the main strength of this approach lies in the fact that VGI-data is widely available in contrast to more accurate cartographic material, however caution should be exercised when estimating the photovoltaic potential of an end-consumer whose roof type is unknown. If more accurate *Geographical Information System (GIS)* data is available, errors in the rooftop estimation of approximately 15 percent can be achieved [WNP10]. With that said, due to the wide availability of VGI-data, the OpenStreetMap-based approach presented in this section poses a means of a quick and approximative preselection, for example to identify regions where more precise, but also more expensive, measurements might be profitable.

3.2 Impact on the electricity grid

Due to the growing digitization within the energy economy, changes to the way energy is delivered to consumers have become mandatory in order to accommodate the plans for the installation of Intelligent Metering Systems as a consequence thereof. These changes to the electricity grid are described by the umbrella term *Smart Grid*. The term Smart Grid was first characterized by the *US Energy Independence and Security Act of 2007* [Con07], which lists the following requirements for Smart Grids:

1. Increased use of digital information and controls technology to improve reliability, security and efficiency of the electric grid
2. Dynamic optimization of grid operations and resources with full cyber-security
3. Deployment and integration of distributed resources and generation, including renewable resources
4. Development and incorporation of demand response, demand-side resources and energy-efficiency resources
5. Deployment of *smart* technologies (real-time, automated, interactive technologies that optimize the physical operation of appliances and consumer devices) for metering, communications concerning grid operations and status, and distribution automation
6. Integration of *smart* appliances and consumer devices
7. Deployment and integration of advanced electricity storage and peak-shaving technologies, including plug-in electric and hybrid electric vehicles, and thermal-storage air conditioning
8. Provision to consumers of timely information and control options
9. Development of standards for communication and interoperability of appliances and equipment connected to the electric grid, including the infrastructure serving the grid
10. Identification and lowering of unreasonable or unnecessary barriers to adoption of smart grid technologies, practices and services

The implementation of these properties requires a transformation of the electricity grid with profound effects on the way energy is produced, traded, distributed and measured. An overview of these changes has been given in figure 1.1 in section 1.1 of this thesis. The way of how the concept of a Smart Grid will transform the energy supply and force energy providers, which to date have had next to nothing to do in terms of *Big Data*, to rethink their business practices, is part of active research. Among other topics, current research is concerned with processing huge amount of data in the cloud, *load classification* and *short-term load forecasting* with respect to security policies [DKK15] as well as means to predict technical failures [RW+12].

One of the main advantages of the digitization of the electricity grid is that it allows for technologies that can not be facilitated in a conventional electricity grid. An example for this are *electric cars*: current electric cars have batteries ranging in capacity from 16 to 53 kWh and consume as much as 7 to 10 kWh of energy to drive a distance of 50 to 65 kilometers. In order to charge these cars in a reasonable amount of time, for example between 3 and 4 hours, electric outlets from 6,6 to 16 kW of power are required [IA09]. However, this means that when an end-consumer charges his or her electric car, the overall energy consumption of that customer more than doubles for the duration the car is charging [IA09]. Depending on the distance a given customer has driven during the day, this can cause high peak amounts of energy to be needed to be injected into the electricity grid in order to ensure the stability of the energy supply.

On a larger scale, this can lead to problems when the electricity grid is not designed to handle the use-case of a significant proportion of customers switching to electric cars [IA09; Wüs17b; Wüs17a], providing the necessity of practical solutions to enable electric cars and other high-power household appliances to use energy in a staggered manner.

One inconvenient property of customer behavior for energy providers is when the consumption occurs in bursts in contrast to a more steady load. This is due to the fact that high peaks of energy load, such as the load generated by charging electric cars, are generally difficult to predict accurately and thus can not be properly accounted for when energy providers are planning their future buy-in of energy, instead requiring cost-intensive real-time adjustments as soon as a peak load becomes apparent. These adjustments typically are executed by using *imbalance energy*, which we have introduced in more detail in section 2.2. Due to the financial risk imbalance energy poses, energy providers generally aim to smooth out the overall load on the electricity grid as much as possible. One way to quantize how abrupt the short-term adjustments need to be is to compute the *Peak-To-Average Ratio (PAR)* which [Moh+10] defined as follows:

$$PAR = \frac{L_{peak}}{L_{avg}} = \frac{H \cdot \max_{h \in \mathcal{H}} L_h}{\sum_{h \in \mathcal{H}} L_h} = \frac{H \cdot \max_{h \in \mathcal{H}} \sum_{i \in \mathcal{N}} l_i^h}{\sum_{h \in \mathcal{H}} \sum_{i \in \mathcal{N}} l_i^h} \tag{3.3}$$

Here, $\mathcal{N}$ corresponds to the total number of customers and l_i^h describes the amount of energy the i-th customer has consumed at the time $h \in \mathcal{H} \triangleq \{1, \ldots, H\}$. Since the PAR is computed per day, the value for H is equal to 24 in case the energy time series have an hourly resolution. The smaller the value of PAR, the more uniform the consumption can be characterized.

With the development of the Smart Grid, several technologies have emerged in order to alter the overall shape of the consumption time series:

- **Demand-Side Management (DSM):** This term describes any model which incentivizes the consumer to become more energy efficient in the long term, for example by *lighting retrofits*, *building automation upgrades*, *re-commissioning*, *improvements to Heating, Ventilation and Air Conditioning (HVAC)*, *variable frequency drives*, etc [Alg+15], as well as other measures to reduce the Peak-To-Average Ratio without introducing external causes such as dynamic pricing models.

- **Demand Response (DR):** This term covers any models to encourage the end user to make short-term adjustments in his or her energy demand. This is typically achieved by creating financial incentives in the form of dynamic, but predetermined pricing models, for example by offering more attractive prices during times where the overall energy consumption is low and more expensive prices during peak load times. This latter model is sometimes referred to as *Time-Of-Use Pricing (TOU)* [WL11; Str08; AE08].

- **Direct Load Control (DLC):** This term describes a subcategory of Demand Response, in which the reaction to dynamic tariffs including the control of some electronic appliances, for example the act of turning them off or on, is carried out directly by the energy provider. The degree to which energy providers assume control over the devices is negotiated with the customer beforehand [Alg+15].

DLC is most frequently applied in conjunction with air conditioning systems and thermal heat pumps [AE08]. While DLC is a simple and cost-effective means for energy providers to shape the overall energy load of large institutions, where the consumption behavior is primarily dictated by the business practices, end-consumers might reject this patronization as being privacy-invasive and intrusive [RV08].

The application of these models have been tried in pilot projects over the last years. For example, in the USA, Demand Response using dynamic pricing models have been deployed [SL07], more specifically *Real-Time Pricing (RTP)* and *Critical-Peak Pricing (CPP)*, which are variations of the more general *Time-Of-Use Pricing (TOU)* [WL11]. However, in order not to lose large industrial customers, these models where offered on a voluntary basis. As [SL07] mentions, the primary motivation to roll out dynamic pricing models have been *customer retention* for enterprises seeking a way to optimize their electricity bill; *peak management* was of far lesser importance to energy providers. The success and amount of savings for customers through dynamic pricing models, such as RTP or CPP, strongly depend on the *elasticity of demand* of the customer. In [SL07], the authors describe an elasticity of $-0{,}2$ as meaning that a price increase of 10 percent will cause the customer to reduce his energy demand by 2 percent. Participants opting for RTP with an elasticity of $-0{,}1$ have achieved a reduction in their invoicing amount of between 3,51 and 6,52 percent, with the overall peak power being lowered by between 14,0 and 24,5 percent [SL07].
A similar success could be observed in the case-study of a library [Alg+15], where the managers of the library and the energy provider agreed upon deploying *Direct Load Control*. In exchange for direct or short-term control of *Heating, Ventilation and Air Conditioning* equipment, the energy provider has agreed on a price discount in the amount of 3 € per MWh. This method caused the amount of energy consumed to be reduced by 4 percent as well as financial savings of 7 percent annually.

Another way for customers to increase their elasticity of demand is by installing a battery to buffer energy locally. In section 3.1.2.2, we have introduced this concept as *Battery-based Load Hiding (BLH)* to help protect user privacy. In this case however, the basic idea is for the battery to charge during low-price time slots and to discharge when energy prices are high. This process could possibly be automated by equipping the batteries with a network interface, analogous to *Intelligent Metering Systems* (see section 3.1), which could enable the battery to query current prices. Subsequently, it would be possible for the battery to autonomously *learn* to classify prices as *low*, *average* and *high* by computing the average price over a sufficiently long time period, thus decide on its own when to charge and when to discharge. Alternatively, the charging and discharging of the battery of the customer could be executed by the energy provider as part of *Direct Load Control*. In this case, the user participates in the energy market as a supplier for *imbalance energy* and would be entitled to claim a financial compensation from the energy provider depending on the amount of battery capacity allocated. Using this technology, the end user could automatically benefit from fluctuating prices and thus amortize the acquisition cost of the battery without any compromises in lifestyle habits. Depending on the functionality the battery provides and the demand of the customer, an optimal balance for the end user between a financial gain by utilization of dynamic pricing models and means to protect user privacy, such as the ones presented in section 3.1.2, is arrangeable.

Further research has been conducted on *Intelligent Load Management* using cooling systems and refrigeration plants [GP11]. The idea here is to suspend or lower cooling activities from time to time and to play on *thermal inertia* to prevent defrosting of chilled goods. By also taking Demand-Side Management into account, cooling systems can be used to supply imbalance energy, albeit to a lesser extend.
However, using the example of the *Model City Mannheim* from 2013 [And13], it has been revealed that large, industrial consumers are missing the necessary communications infrastructure between refrigeration plants and control instances to participate in an intelligent, decentralized Smart Grid. Instead, a unidirectional communication of price data from the energy provider to the customer has taken place, with the customer manually deciding on which measures to adopt. Overall, this likely results in industrial customers, which store chilled goods and thus are liable to strict legal temperature policies, to deem such measures as being especially risky and intrusive for their core business. One of the main reasons for this is that decision makers often lack the necessary experience to give a sufficiently accurate estimation for the cost-benefit ratio.

To counteract the *intrusive* aspect of Direct Load Control mentioned above, technologies have emerged that offer specifying a declarative objective, but the means to optimize that objective is automated. For that purpose, techniques like *Dynamic Programming* [RV08] and *Binary Particle Swarm Optimization* [PSM09] can be used. One example for specifying a declarative objective is when a customer configures a target room temperature for the air conditioner, as well as an allowed deviation from said target temperature, but leaves the actual implementation to the devices as part of *home automation*. In that case these devices may semi-autonomously carry out their assigned task with intervention from the energy provider being limited to the usage restrictions as given by the temperature deviation tolerated by the customer.
However, in order for Demand-Side Management to work in a home automation environment, agreements between smart home devices or energy providers have to take place to some extend. If smart home devices such as air conditioners were to act fully autonomous, the main basis of decisionmaking for when to start or stop heating or cooling is the current room temperature. This leads to an event what the authors of [RV08] have called the *cold load pickup phenomenon*. This anomaly occurs if the devices cause a sudden surge in energy consumption when they turn themselves on after they have been turned off for a while. In general, if no coordination between devices takes place, this phenomenon could cause huge energy load peaks, which is the opposite of what Demand-Side Management methods are trying to achieve. One solution for this problem involves *Dynamic Programming*, meaning that air conditioning devices are being turned on and off in a cyclic and staggered manner, so that the overall energy consumption stays constant or increases / decreases as uniformly as possible, while at the same time preferring devices being turned on where the difference between the target and actual temperature is greatest [RV08].
In addition to the agreements between devices mentioned above by smart home equipment explicitly communicating with the devices of the customer's neighbors, the Intelligent Metering Systems could be designed to allow for measuring voltage fluctuations in addition to the energy consumption [MRT12]. From the point of view of an energy provider, keeping the electricity grid balanced is of utmost importance. With local voltage readings known, this allows for quick and precise diagnoses of potential prob-

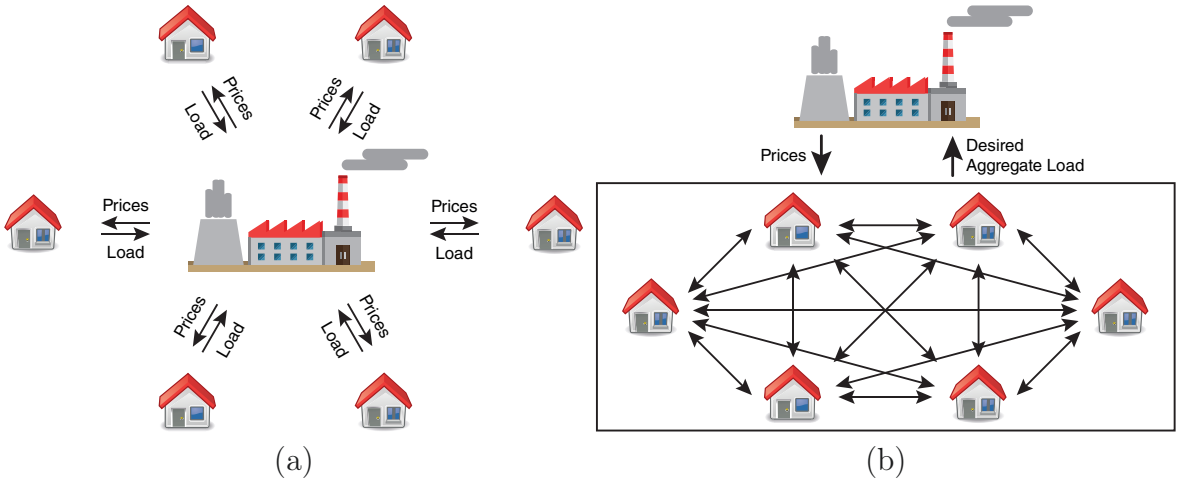


Figure 3.3: Overview of Demand-Side Management strategies focusing on (a) individual interactions between the energy provider and each customer and (b) the Smart Grid with enabled communication between the customers and the energy provider. Adapted from [Moh+10] with some assets taken from [Wik].

lems, which ultimately helps in reducing customer complaints and minimizing customer minutes lost. In addition, knowing local voltage readings enables the energy provider to identify *problem areas*, meaning areas where consumption shaping measures such as Direct Load Control most urgently need to be taken. At the same time, since voltage measurements at the house connection are region-specific, not customer-specific, and thus do not allow to deduce private activities of the consumer, no privacy protection measures such as the ones presented in section 3.1.2 need to be put in place.

Another way for the consumer to specify a declarative intent for smart home devices to implement is an *Energy Consumption Scheduler (ECS)* [Moh+10]. The idea here is that, for the example of electric vehicles, most customers are indifferent to when exactly the car is being charged, as long as it is sufficiently charged the next time it is needed. Thus, ECS allows for the consumer to specify a time interval for when the vehicle can be charged; the actual charging time can then either be dictated by the energy provider or agreed upon by neighboring smart home devices so that charging happens in a staggered manner with the goal for the energy provider being to avoid problematic peak loads and for the end-consumer to benefit from financial incentives due to Demand Response programs. This idea is not limited to electric vehicles, but can also be extended to include other high-power household appliances such as intelligent washing machines and dishwashers [And13].

Overall, assuming the necessary communication infrastructure for both consumers and energy providers is in place, DSM programs are commonly employed using one of two strategies, which are illustrated in figure 3.3. The topology in figure 3.3 (a) is more commonly used in RTP programs, but also puts more stress on the energy provider as the provider has to individually communicate with each customer [Moh+10]. In contrast, the topology in figure 3.3 (b) frees the energy provider of almost all coordination efforts and instead allows the provider to focus on giving financial incentives to shape the aggregated energy load to more directly impact the PAR introduced in equation 3.3. Although the scenario in figure 3.3 (b) requires some interactions between smart home devices of the customer, these interactions can be automated using

bidirectional digital communication.
In order for a set of customers to reach an agreement using a decentralized topology such as in figure 3.3 (b), a distributed mechanism based on game-theory is proposed in [Moh+10]. Here, as a preliminary step, the ECS distinguishes between *shiftable* and *non-shiftable* household appliances. Shiftable appliances, such as electric vehicles or dishwashers, allow for flexible scheduling of their energy consumption, in contrast to non-shiftable appliances, such as refrigerators or lights, of which the user expects to be switched on permanently or on-demand. By knowing which devices are shiftable, as well as their required energy and the scheduling interval that the customer tolerates for each device, it is possible for an Intelligent Metering System to derive a set of consumption scheduling plans $\mathcal{X}_i$ for all connected household appliances.
Finally, the goal for the Intelligent Metering System is to minimize the PAR using the *best* scheduling plan among the set of feasible consumption scheduling plans it has calculated earlier. The *best* schedule $x_i \in \mathcal{X}_i$ for a given customer then corresponds to the solution for the following optimization problem:

$$\underset{x_i \in \mathcal{X}_i}{\text{minimize}} = \sum_{h=1}^{H} C_h \cdot \left(x_i^h + \sum_{i' \in \mathcal{N} \setminus \{i\}} l_{i'}^h \right) \tag{3.4}$$

Here, x_i corresponds to a scheduling plan from the set of all possible consumption scheduling plans $\mathcal{X}_i$ for the i-th customer, with x_i^h describing the scheduled consumption at the time h. Furthermore, C_h describes the cost function announced by the energy provider, meaning the price per energy unit, which is assumed to be only dependent on the total amount of energy the energy provider has to ingest into the electricity grid. Lastly, $\sum_{i' \in \mathcal{N} \setminus \{i\}} l_{i'}^h$ describes the sum of the announced energy schedules by all consumer *other* than the i-th customer.
Since the optimization problem 3.4 only contains local variables of a given customer, its solution can be locally computed by the Intelligent Metering System. As shown in [Moh+10], the solution for the optimization problem 3.4 exists and is unique assuming that C_h is *monotone increasing* and *strictly convex*. In that case, the solution to the optimization problem 3.4 corresponds to the most cost-effective scheduling plan for a given customer and at the same time forms a *Nash equilibrium* with the most cost-effective scheduling plans of the other customers, where no customer benefits if they deviate from their *best* schedule or provide inaccurate information to mislead other consumers.
While the strategy presented in [Moh+10] is an interesting approach to incentivize collaboration of intelligent appliances in a Smart Grid environment in a decentralized way, relieving energy providers in the process, its main disadvantage is that it requires all x_i^h to be predictable accurately and the l_i^h of all customers to be known by all other customers in order to compute $\sum_{i' \in \mathcal{N} \setminus \{i\}} l_{i'}^h$, giving rise to valid concerns with respect to privacy. While the authors of [Moh+10] argue that announcing l_i^h publicly does not disclose details about the energy consumption, this is not the case, as we have discussed in section 3.1.2 of this thesis. One possible solution for this problem is a hybrid of both topologies depicted in figure 3.3, where all customers send their preferred x_i^h as l_i^h to the energy provider using an encrypted communication channel; the energy provider then aggregates all energy schedules by computing $\sum_{i \in \mathcal{N}} l_i^h$, which is then broadcasted to all consumers. Each customer can then individually solve his or her own optimization

problem 3.4 by computing:

$$\sum_{i' \in \mathcal{N} \setminus \{i\}} l_{i'}^h = -x_i^h + \sum_{i \in \mathcal{N}} l_i^h \tag{3.5}$$

Since x_i^h is known only by the i-th customer and the energy provider, no other customer $i' \in \mathcal{N} \setminus \{i\}$ is able to derive l_i^h from the broadcast of the energy provider. While this solution centralizes the problem to some degree and necessitates the energy provider to be a trusted party with regard to privacy, its role would be limited to aggregating the individual l_i^h and broadcasting the result, which still requires less effort from the energy provider than a fully centralized topology.

3.3 Impact on energy providers

The primary ways for energy providers to improve are by achieving a higher predictability for their buy-in of energy as well as lowering costs for electricity grid infrastructure. The first aspect can be accomplished by gaining a better understanding of customer needs or by optimizing existing internal processes. In addition, new technologies such as Smart Metering enable shaping of the consumption behavior using techniques like *Direct Load Control* as introduced in section 3.2, which allows for shifting energy consumption into periods of lower demand, thus making it possible for the electricity grid to not be designed for maximum load at all times, ultimately lowering infrastructure costs. At the same time, legal obligations and requirements of the energy market, such as the ones presented in chapter 2, must be kept in mind. Over the course of this section, we will introduce some of the most important goals of energy providers as well as discuss typical strategies from academic literature on how to achieve these goals.

3.3.1 Extraction of customer insight

One way for energy provider to gain information about customers aside from contractual data consists of conducting customer surveys, in addition to employing a *Customer Engagement Portal*. This often is a central part of marketing strategies to better understand customer requirements which in turn enables the energy provider to adjust tariff offers and internal processes accordingly. In order to yield a representative statistic of the whole customer basis, in most cases it is necessary for the energy provider to incentivize participation in these surveys. Such incentives to increase participation might be discounts for upcoming billing cycles or perks in cooperation with other companies.

However, even when survey participation is incentivized, it is very rare for the majority, let alone the entirety, of customers to partake. Thus, businesses usually seek ways to deduce information on customers who have not partaken in surveys by extrapolating insight gained from users who have voluntarily provided their information. One way to apply this approach on customers of energy providers is presented in [HSK16]. The idea here is to use supervised machine learning to build a model for household characteristics from survey data. As a data basis, the authors have used data of 3.986 customers of an energy provider in Switzerland who have voluntarily entered their data in a Customer Engagement Portal. In total, said energy provider delivers electricity to 10.482 customers, meaning that the overall participation rate is approximately 38

percent. Among the features that have been gathered as part of the survey to train the model are the following [HSK16]:

- **Household Type:** A categorical attribute which describes whether the customer lives in a *house* or an *apartment*.
- **Living Area:** An attribute which corresponds to the living area of the customer. In the dataset, the living area of the customers cover a range from $10m^2$ to $5.443m^2$. Although the living area of customers is given in the dataset as a numerical attribute, the authors have converted it into a categorical attribute by defining the ranges $0m^2 - 95m^2$, $95m^2 - 145m^2$ and $145m^2 - \text{inf}$ as class borders, which are motivated on 33 percent and 66 percent quartiles.
- **Number Of Residents:** An attribute which describes the number of people living in the corresponding house or apartment. Similar to the feature-attribute *Living Area*, this attribute has been converted to a categorical attribute by defining a set of ranges to serve as class borders. During their research, the authors have experimented with multiple set of ranges. Overall, the class border ranges 1, 2, $3 - 5$ and > 5 have yielded the best results.
- **Logarithmic Average Daily Consumption:** This numerical attribute corresponds to the logarithmic value of the ratio of the total amount of energy consumed divided by the number of days where energy consumption took place. The logarithmic transformation is applied to achieve a more symmetric distribution of this attribute in the dataset.
- **Relative Consumption Trend:** This attribute corresponds to the linear regression of the customers consumption over the last 4 years.
- **Z-Score In Neighborhood:** An attribute which describes the *Z-Score* of the customer's attribute *Logarithmic Average Daily Consumption* within the customer's neighborhood. The *Z-Score* corresponds to the number of standard deviations of a data tuple x_i from the expected value of the distribution. The *Z-Score* is calculated as follows: $z_i = \frac{x_i - \overline{x}}{\sigma}$

In addition to the features described above, the authors of [HSK16] have used 66 geographical features derived from *Volunteered Geographic Information (VGI)* sources, namely *OpenStreetMap* and *GeoNames.org*, such as the distance of the home of the customer to the nearest city center with up to 1.000, 5.000 or 15.000 inhabitants. To increase the accuracy of the model, some of the attributes have been pruned using *correlation-based feature selection* [Hal99] to reduce the number of dimensions. The basic idea of correlation-based feature selection is that an attribute is useful for a model if and only if that feature is predictive of or correlates with the class, but the features themselves are uncorrelated among each other.
With this setup, the best results have been achieved using *Random Forest* classification [Bre01], followed up closely by *Support Vector Maschines* [HPK11], using *5-fold cross validation* [Sto74]. In simplified terms, *Random Forest* classification works by independently generating random decision trees with weak correlation and then employing a majority vote on a per-feature-vector basis to assign the class which the feature-vector got most often classified to by the generated decision trees. By including VGI sources

for the evaluation, the authors achieved accuracies from 49,4 and 68,7 percent. For most business cases of energy providers, this accuracies are good enough to then justify targeted marketing campaigns for personalized tariff offerings. If VGI data is left out or not available, the accuracy of the feature *Household Type* drops by 12,7 percent and the accuracy of *Living Area* drops by 7,0 percent.

3.3.2 Installation of intelligent metering systems

One of the major upcoming technical advancements in the energy economy are *Smart Metering* devices, which enable fine-grained measurements of the energy consumption as well as corresponding data processing and transmission. In previous sections of this chapter as well as figure 1.1 in section 1.1, we have outlined the most important changes brought by Smart Metering technology. With historical consumption time series for each customer of the energy provider present, this allows for the employment of time series analysis or other appropriate techniques to help classify a customer or to find reoccurring patterns in the data, but also to assess the influence of plants for the production of renewable energy such as photovoltaic systems (see section 3.1.3). The latter is particularly important since photovoltaic systems make the customer a partially self supporter, this causes deviations between the expected consumption by the energy provider, for which the energy provider has prepared for, and the actual consumption of the customer (see section 2.2).

Due to the fact that the data gathered by Intelligent Metering Systems and possibly other smart home devices are progressively complex and vast, technologies which allow for the extraction of useful knowledge from such data are becoming increasingly significant. Such useful knowledge can include models which allow for a more accurate prediction of the upcoming energy consumption of a customer, thus helping the energy provider to manage the tasks outlined in section 2.2, but also ways to help identify target audiences for which custom tariff offers might be financially attractive while still covering their use-cases as good as possible. Over the course of the following sections, we will introduce some approaches from academic literature to help achieve this goal.

3.3.2.1 Prediction of household properties

The main motivation of energy providers for seeking to know certain properties of their customers is that it allows for targeted ad campaigns and custom tariff offerings, which customers might deem more relevant for their use-cases. Very often, the underlying data for such analyses is gathered by executing expensive customers surveys as we have discussed in section 3.3.1. Due to the emerging Smart Metering technology however, customer properties can also be extracted by examining customer behavior patterns from energy consumption time series. Since time series data is fundamentally different from survey data, which is often in the form of multiple-choice questions, the data requires a different approach to be processed expediently.

One approach to segment private households is presented in [VVS16], which is special in that it uses the *CER-Dataset* [CER] for analysis, as it consists of both Smart Metering time series as well as survey data for most of the corresponding households. Using this dataset, the approach focuses on Data Mining techniques by classifying the customers using features derived from the Smart Metering time series, while using the

survey data as *Ground Truth* to assess the results for households not part of the training data. To train a classification model, [VVS16] relies on *Support Vector Maschines* [HPK11] while employing *10-fold cross validation* [Sto74] using features such as the *ratio of the average daily consumption compared to the daily maximum consumption* or *ratio of the daily minimum consumption compared to the daily maximum consumption* to assign classes to the households in the categories *income* (*high* or *low*), *education* (*superior* or *non-superior*) and *children* (*present* or *not present*). Using this setup, the authors of [VVS16] have achieved accuracies of 63 percent for the category *income*, 63 percent for *education* and 69 percent for *children*. When assessing the quality of these results, it is important to recall that the categories in the experiments are binary, meaning that accuracies from 63 to 69 percent pose only marginal to moderate improvements compared to random guessing.
A similar approach is presented in [Hop+16]. Like [VVS16], the authors aim to predict properties of private households using classification techniques on Smart Metering data, specifically the *CER-Dataset* [CER]. In addition to [VVS16] however, which only uses 8 features for their method, [Hop+16] defines a total of 88 features and employs feature-selection methods, such as *correlation-based feature selection* [Hal99] and *Kolmogorov-Smirnov Test based feature selection* [Lop11], to prune redundant and unnecessary features. Within the scope of the experiments in [Hop+16], the authors have noted that including feature-selection methods improves classification accuracy by approximately 8 percent. A much smaller influence on classification accuracy has been the choice of the classification algorithm itself, with *Support Vector Machines* [HPK11] yielding an on average only 2 percent higher accuracy than *k-Nearest-Neighbor* [ES00; HPK11], which has been the other classification algorithm tested. Overall, the average classification accuracy achieved using Support Vector Machines and feature-selection methods is approximately 60 percent, which is slightly lower than the experiments shown in [VVS16], even though the same dataset is used. However, the main influence on classification accuracy has been the choice of class borders, as the authors have shown by also repeating their experiments using different definitions for said class borders. For example, the experiments have yielded an accuracy of 50 percent for the class *age of the house* when using the definition *old* $\hat{=} >$ 30 years and *new* $\hat{=} \leq$ 30 years, while yielding an accuracy of 80 percent when using the definition *old* $\hat{=} >$ 10 years and *new* $\hat{=} \leq$ 10 years, which indicates a strong influence of the choice of class borders on the accuracy of the results. Another point of criticism is that the experiments in [Hop+16] have been conducted using only 1 week of consumption time series out of the 76 weeks available from the *CER-Dataset*. While the authors claim that the week they used is noise-free and without public holidays as well as average weather conditions, the assumption that this 1 week is representative for all consumers, meaning that no end-consumers are gone on holiday during this period or have other unusual events disturbing their normal daily routines, is still hard to justify.

One aspect that has not been part of the experimental evaluation of the aforementioned research of [VVS16; Hop+16] is the influence of the granularity of the data on the performance of the Data Mining algorithm. With that in mind, it is important to note that, as an energy provider, requiring a particular minimum resolution of the consumption time series from the Intelligent Metering System means carefully balancing business interests with the privacy concerns of customers, as we have mentioned in section 3.1.2.3. This aspect of data resolution has been examined in further research.

In [Sod+17; HSS18] for example, the authors pursue the same goal of predicting private household properties using Smart Metering data, while also assessing the impact of the resolution of the time series on the classification accuracy. In both publications, the authors have used a Smart Metering dataset from a Swiss energy provider with 9.000 customers containing time series data for 1 year, of which 527 customers have participated in a voluntary online survey. However, due to the fact that not all customers have fully answered the online survey, 40 percent of these 527 customers had to be pruned from the dataset, so that approximately only data from 316 customers have been used for the experimental evaluation in both [Sod+17; HSS18]. The basic approach in both publications is very similar to [VVS16; Hop+16] in that the authors define classification features according to the Smart Metering time series, use the categories of the survey data as *Ground Truth*, such as *age of appliances* with the possible values *new / average / old* and *cooking type* with the possible values *electric / not electric*, and utilize *cross validation* [Sto74] to segment the data into training and test datasets. The main novelty of [Sod+17], besides the aforementioned evaluation of the influence of the granularity of the time series data on the performance of the Data Mining algorithm, is that [Sod+17] also assesses the influence of the choice of the classification algorithm by repeating the experiments using *Support Vector Maschines* [HPK11], *k-Nearest-Neighbor* [ES00; HPK11], AdaBoost [HPK11] as well as *Random Forrest* [Bre01] and *Naïve Bayes* [ES00; HPK11] classification. In contrast to [Sod+17], [HSS18] does not mention the usage of feature-selection methods and only uses *Random Forrest* classification, but includes *multiweek*-classification, where the experiments are not only conducted using a single week of Smart Metering data like in [Hop+16; Sod+17], but all weeks in the dataset are classified individually; the resulting class predictions are then aggregated by computing the average of the confidence values over all weeks.
As for the results, the authors of [Sod+17] have achieved an average accuracy of 70 percent in their experiments, with *Random Forrest* classifiers yielding the best results in 7 out of 11 cases. In doing so, the granularity of the data has been a critical factor for the classification performance, with the accuracy significantly worsening when switching from hourly Smart Metering time series to daily measurements. In general, the performance is better the higher the resolution of the time series data is, however with the accuracy only marginally improving beyond hourly resolution. Furthermore, in [HSS18], the authors have shown that employing *multiweek*-classification significantly improves accuracy over only using singular weeks for most classes, although the accuracy strongly depends on the number of weeks used. While heating related attributes, such as *water heating type*, *space heating type*, *heat pump* and *solar installation*, have been found to be very accurately classifiable using Smart Metering time series, the performance of some other attributes, such as *age of appliances*, *cooking type* and *efficiency measures*, has been only marginally better than random guessing. As in [Sod+17], the authors of [HSS18] have found that, for most attributes, higher resolution time series yield better accuracies; the accuracy of some attributes however, such as *solar Installation*, has even got worse when using data at a higher resolution than *daily*.
Overall, while the basic approach in [Sod+17; HSS18] seems very promising, it has to be noted that only approximately 316 out of 9.000 customers have been used during the evaluation in both publications, which suggests a strong *selection bias* due to the small sample size, as the authors themselves have noted. In addition, the experiments

have been conducted using data only from one urban energy provider, which also constitutes a *regional bias*. This causes the conclusions of the experiments to likely not be generally valid and possibly need to be repeated if statements for the classification performance of the customer behavior in more heterogeneous datasets are required. To address the *selection bias* of the dataset, it might have been desirable to first perform clustering on the Smart Metering data of all 9.000 customers, possibly using one of the algorithms presented in section 1.2.2, and then to examine if the customers who have participated in the online survey evenly distribute themselves on all clusters or if survey participants are over- or underrepresented in some clusters. With respect to the fact that less than 4 percent of the customer base of the Swiss energy provider has been considered during experimental analysis by the authors and depending on how misrepresented some clusters would be with regard to survey participants, it would have given the reader a better sense for how strong the *selection bias* and thus how representative the results of the experimental evaluation of the dataset actually are.

3.3.2.2 Creation of target-group-specific tariffs

The creation of target-group-specific tariffs is often carried out by marketing salesmen after sufficient information about potentially interested customers is available. For energy providers, this information about customers typically comes in the form of household properties which can be gathered as part of customer surveys or by analyzing Smart Metering time series as we have outlined in the preceding section. Very often however, it is desirable to have the Data Mining algorithm output the target group segmentation directly, either instead of or in conjunction with outputting discovered and potentially relevant household properties. For this purpose, research has predominantly applied clustering techniques to identify such target groups.

One such clustering-based approach is presented in [Rod+03]. Here, the authors use *K-Means* [Mac+67], a well-known partitioning clustering algorithm which we have introduced in for detail in section 1.2.2, *Kohonen Self-Organized Maps* [Koh89] as well as a combination of both to segment consumption time series as means to identify tariff groups. Since customers may or may not behave very similar in terms of energy consumption, the careful choice of the metric which is used to evaluate the dissimilarity of energy consumers to segment the data tuples is paramount. For this purpose, the authors have opted to use a distance function based on the well-known *Euclidean Distance* and to normalize all energy consumption time series by dividing all measurement by the *peak value* of the corresponding customer to accommodate for customers whose consumption behaviors differ by a scaling factor. To convert consumption time series into tuples in order to be able to apply the Euclidean Distance, the time series of each customer is split at midnight to form a feature-vector $s_{i,a} = \left(s_i^1, s_i^2, \ldots, s_i^H\right)$. Here, H describes the number of measurement per day. Since consumption time series in the energy economy are typically recorded in either 60-, 30- or 15-minute intervals, H equals 24, 48 or 96, respectively. Overall, the number of $s_{i,a}$ feature-vectors for the i-th customer corresponds to the number of calendar days a for which consumption measurements are available. Using a dataset containing 165 customer covering a time span of 6 months of consumption as well as corresponding contractual data of the customers, the authors have found a poor correlation between contractual details and cluster membership, indicating that contractual details are not suitable for tariff group segmentation.

In order to investigate if, in contrast to contractual data, information about household properties has a significant relation with class segmentation, the authors of [VVS15] present a clustering approach to yield seasonal load profiles and a classification approach to assign customers to a load profile depending on their household properties. For this purpose, the authors have used the *CER-Dataset* [CER], which contains Smart Metering data and survey data. As for their experimental setup, the time series consumption data is first segmented into seasons (spring, summer, autumn, winter) and further split at midnight into tuples. The seasonal datasets are then individually clustered by the well-known partitioning clustering algorithm *Fuzzy-C-Means* [Bez81] while testing values from 2 to 7 for the number of clusters c per dataset. What is important to note here is that no artificial segmentation into training and test data has been conducted; instead, the quality of the cluster segmentation has been exclusively evaluated using *Cluster Validity Indices (CVI)*, a mechanism which we will introduce in section 4.2. With all clustering segmentations and their corresponding assessment via Cluster Validity Indices present, the authors of [VVS15] have concluded that the best clustering segmentation is the one when using $c = 2$ as the number of clusters for each season. In each of these seasons, the pair of clustering centroids for $c = 2$, which can be interpreted as a representative consumption for all customers assigned to that cluster, have described roughly the same consumption behavior, with a multiplicative factor being the most noticeable difference between each pair of centroids. Because of this, the consumption patterns have been labeled the *high*- and *low*-profile. By further training classification models using the survey data that have come with the *CER-Dataset*, the authors have tried to determine the most crucial household properties to predict the cluster membership for a given customer. Here, the best results have been achieved using *Support Vector Maschines* [HPK11], with the insight of the *low*-profile correlating strongly with the household properties *employment status*, *social status* and *education*, while the *high*-profile has correlated strongly with the properties *dishwasher*, *tumble dryer*, *number of adults* and *number of children*. Since the authors have declared an accuracy of approximately 75 percent for predicting household properties using their trained model, these results indeed suggest that household properties help to identify tariff groups. At the same time, it is important to note that the authors mention a strong overlap in the memberships of the properties, which according to the authors indicates the need for feature-selection.

Clustering techniques have also been part of other research publications which aim to build an optimal tariff groups segmentation, for example in [Chi12]. Similar to [VVS15], the author segments Smart Metering consumption time series to gain clusters representing typical consumption patterns, which the authors call *traversal grouping*. In contrast to [VVS15] however, the author of [Chi12] does not assume a strict seasonal segmentation on which traversal grouping analysis is performed, but also derive a temporal segmentation by clustering the total energy consumption, which the authors call *longitudinal grouping*. For both grouping methods, either normalized daily consumption vectors can be used as features, analogous to [Rod+03], or *indirectly determined* shape features, such as the ratio of the average consumption during day- and nighttime or the coefficients of the *Fourier transformation* of the consumption time series. The dataset for the clustering analysis then consists of a matrix with one row per customer and one column per feature. As for the actual clustering algorithm, the author of [Chi12] tests a variety of approaches including partitioning and hierarchical clustering.

All these techniques have in common that for the purpose of identifying target groups, only the memberships values are of interest, even if the clustering algorithm used outputs clustering prototypes as it is the case by *K-Means* [Mac+67] or *Fuzzy-C-Means* [Bez81] for example. This combination of *longitudinal grouping* and *traversal grouping* has been tested by the author with a dataset containing consumption time series of 400 customers with a resolution of one measurement every 15 minutes for one representative working day during the *interim period*. The best results, according to the Cluster Validity Indices used to assess the quality of the clustering segmentation, are dependent on the specific goal of the process; in order to identify and exclude outliers, hierarchical clustering in conjunction with the Single-Linkage-Criterion has achieved the best results, whereas K-Means and Fuzzy-C-Means have been deemed suitable if a direct segmentation of consumption data into target groups is desirable.

Another approach, with focuses on a *Mixed Fuzzy Clustering* method [Fer+15] to incorporate both Smart Metering time series and household properties for target group segmentation, is presented in [Sch+15]. Contrary to the previously mentioned research of [VVS15], which uses household properties in a classification mechanism following the clustering step, the authors of [Sch+15] use these household properties directly as part of the clustering step. For this purpose, the authors consider the Smart Metering data as *time-dependent* data, whereas the household properties are considered *time-independent* data, such as consumer income, age and number of residents.
The basic idea of *Mixed Fuzzy Clustering* is to compute clustering prototypes for the *time-independent* attributes and for each timeslot of the *time-dependent* data independently using *Fuzzy-C-Means* [Bez81] and then to deduce a clustering segmentation of the complete dataset from these clustering prototypes using an adequate distance function. More specifically, the r time-independent attributes for the i-th customer are notated as a vector

$$x_i^s = \left(x_{i,1}^s, \ldots, x_{i,r}^s\right) \tag{3.6}$$

where the superscript s indicates the time-independent attributes. Similarly, the time-dependent data is notated as a matrix X_i^t as follows:

$$X_i^t = \begin{pmatrix} x_{i,1,1}^t & x_{i,1,2}^t & \cdots & x_{i,1,p}^t \\ x_{i,2,1}^t & x_{i,2,2}^t & \cdots & x_{i,2,p}^t \\ \vdots & \vdots & \ddots & \vdots \\ x_{i,q,1}^t & x_{i,q,2}^t & \cdots & x_{i,q,p}^t \end{pmatrix} \tag{3.7}$$

Here, p corresponds to the number of time-dependent attributes and q describes the number of measurements for each attribute for the i-th customer. The o-th clustering prototype of the time-independent attributes is then computed as

$$v_o^s = \frac{\sum_{i=1}^n u_{o,i}^m \cdot x_i^s}{\sum_{i=1}^n u_{o,i}^m} \tag{3.8}$$

while the o-th clustering prototype of the time-dependent attributes is given as

$$v_{o,k}^t = \frac{\sum_{i=1}^n u_{o,i}^m \cdot x_{i,k}^t}{\sum_{i=1}^n u_{o,i}^m} \tag{3.9}$$

for the k-th time-dependent attribute. Similar to X_i^t, the clustering prototypes $v_{o,k}^t$ belong to the matrix V_o^t. Once the individual v_o^s and $v_{o,k}^t$ are known, the clustering

segmentation for the complete dataset is then given as

$$u_{o,i} = \frac{1}{\sum_{b=1}^{c} \left(\frac{d_\lambda^2 (v_o^s, V_o^t, x_i)}{d_\lambda^2 \left(v_b^s, V_b^t, x_i\right)} \right)^{\frac{2}{m-1}}} \tag{3.10}$$

using

$$d_\lambda^2 \left(v_o^s, V_o^t, x_i\right) = \left\| v_o^s - x_i^s \right\|^2 + \lambda \sum_{k=1}^{p} \delta \left(v_{o,k}^t, x_{i,k}^t\right) \tag{3.11}$$

where λ corresponds to a user-defined weight for the time-dependent data and δ represents the euclidean distance. Analogous to Fuzzy-C-Means as presented in section 1.2.2, the computation of V_o^t, V^s and the membership matrix $u_{o,i} \in U$ is repeated in a loop until the changes of U compared to the previous iteration are less than a user-defined threshold ϵ.
Overall, [Sch+15] presents an interesting concept using *Mixed Fuzzy Clustering* [Fer+15] to process both time-dependent and time-independent data. However, due to the fact that each clustering is performed individually per timeslot, it requires the dataset to be *complete* and can not be applied unaltered if the data contains *missing values* in either the time-dependent or time-independent attributes. Since time-independent attributes in the form of survey data typically is, if at all, available only in very limited quantities, and since *missing values* can occur in the time-dependent data due to technical failures, only a very small portion of the customer base can be processed using this approach in a real-world scenario. Because of this, Mixed Fuzzy Clustering is likely best suited to find common interests of consumers for target-group-specific tariff offerings since, in contrast to the forecast of the energy consumption of consumers, it is not critical for all customers to be included in this processing technique. Alternatively, it might be desirable for Mixed Fuzzy Clustering to adapt techniques to allow data containing *missing values* to be processed, as the authors of [HB01; SL01] have presented in the case of Fuzzy-C-Means. Another possibility would be to apply methods to predict missing information about household properties using available Smart Metering data, such as the ones presented in section 3.3.2.1, and then to use Mixed Fuzzy Clustering on the original household properties in conjunction with the conclusions drawn by classifying previously unknown properties. Depending on the set of features used to train a classifier to carry out this task, the features can be meaningfully defined even if the Smart Metering time series contain missing values. This idea outlines a potential starting point for further research.

3.3.2.3 Forecast of the energy consumption

The forecast of the total energy consumption of their customers is one of the most important aspects of the task of energy providers to keep the electricity grid balanced in terms of energy injection by the energy producers and withdrawal by the customers. This is due to the fact that energy providers need a bit of lead time to announce the expected energy demand to the energy producers. Knowing the expected energy load in advance enables to ensure that necessary capacities are allocated and the energy is injected into the electricity grid exactly when it is needed by the consumers. We have given a more detailed introduction into this topic in section 2.2; in this section, we

present some approaches from academic literature on how to more accurately predict the energy load of customers in order to plan the energy providers buy-in of energy more efficiently.

One way of improving the forecast of the energy consumption for energy providers is by building upon existing technology for how this task is currently being dealt with by the industry. As of this writing, most energy providers in Germany use the *BDEW Standard Load Profiles* [Mei+99; FT00], a consumption forecast model which we have introduced in section 2.3 in more detail. However, these load profiles, having been compiled during the end of the 1990s, have not been adjusted to recent advances in technology. Because of this, they are considered to be an increasingly bad model to forecast the energy load of customers [Roo+14]. With the upcoming broad availability of Intelligent Metering Systems and corresponding high resolution consumption time series, research has been conducted to build better prediction models to help energy providers maintain the security of the energy supply.
The approach presented in [SM17] tries to achieve this goal by slightly altering the shape of the BDEW Standard Load Profiles by utilizing Smart Metering data. To accomplish this, the authors use two datasets. The first dataset is a total consumption time series of an energy provider for the year 2012 until 2015, from which disturbing factors, such as grid losses, have been manually subtracted. As a result, the total consumption time series only of the customers who have a load profile assigned remains, which corresponds to what we have introduced and referred to as the *SLS* time series in section 2.3. This *SLS* time series can be seen as the sum of an unknown consumption time series containing only private households and an unknown consumption time series containing only industrial customers. The second dataset contains Smart Metering data for the years 2012 until 2014. In their analysis, the authors of [SM17] have kept most of the structures from the BDEW Standard Load Profile in place, including the set of *day-types* and the concept of a *dynamization function*. To derive a new household load profile from the first dataset, the authors have used the original industrial load profile *G0* from the BDEW to yield an approximation for the unknown consumption time series containing only private households. This approximation is then processed according to the original approach of the creators of the BDEW Standard Load Profiles [Mei+99; FT00]. In simplified terms, this means that the time series have been grouped according to the day-type segmentation; for each day-type, the representative consumption pattern for the load profile is then given by computing the average consumption per time of day of the corresponding time series data. Similarly, the load profiles for the second dataset are built by aggregating the individual Smart Metering time series to a household time series and an industrial time series. These two time series are then further processed analogous to [Mei+99; FT00] to yield a new household load profile and a new industrial load profile akin to the BDEW profiles *H0* and *G0*, respectively.
During the experimental evaluation, the authors claim a deviation between the actual energy load and the load predicted using these *new* load profiles of roughly 3 percent, which is an excellent result compared to the BDEW Standard Load Profiles typically achieving a forecast error of 12 to 13 percent as we have shown in figure 2.5. In addition, the authors have demonstrated that their load profiles as well as their dynamization functions significantly differ from the existing BDEW Standard Load Profiles, which could indicate both the consumers having changed their consumption behavior over

the last decades or a *selection bias* caused by the input dataset. In fact, very little information over the dataset is provided apart from the fact that the dataset contains 0,6 percent of all industrial customers, which strongly reinforces the suspicion of a *selection bias* in the input data. If this circumstance would be confirmed, then the algorithm might not have had to struggle to consolidate the requirements of industrial customers of different industry branches, allowing a single load profile to yield excellent results due to the homogeneity of the dataset. With regard to future research, since the basic idea of this approach is very akin to [Mei+99; FT00], testing it with a more heterogeneous dataset might give more insight on whether the customer behaviors have indeed significantly changed over the last decades.

Instead of relying on established methodology for generating load profiles to forecast the energy demand, Data Mining techniques can also be used in a slightly altered fashion to process Smart Metering time series to create a model which is only distantly related to the concept of load profiles. For example, the authors of [ALB13] maintain the concept of customer group and day-types, but make use of an *autoregressive model* $AR(1)$ to be able to long-term forecast the energy load. More specifically, the expected energy demand of a given customer group at a given time c^t is modeled as follows:

$$c^t = \sum_d a_d \cdot D_d^t \sum_m a_{d,m} \cdot D_m^t \sum_h a_{d,m,h} \cdot D_h^t + \epsilon_t \tag{3.12}$$

Here, the coefficients a of the model are dependent on the day-type (subscript d), month (subscript m) and the time of day (subscript h) and are determined by the $AR(1)$ model. The variables $D_d^t \in \{0,1\}$, $D_m^t \in \{0,1\}$ and $D_h^t \in \{0,1\}$ in equation 3.12 are binary and equal 1 if and only if a given time, notated by the superscript t, belongs to the corresponding day-type, month or time of day. While this is a simple model based on statistics, it is a straightforward way for energy providers to make long-term predictions and planning, either instead of, or in conjunction with, existing models based on load profiles.

4

Clustering using Smart Meter Data

With the increasing availability of *Smart Metering* devices, it has become progressively easier for energy providers to optimize business processes by improving customer understanding using techniques from the field of *Knowledge Discovery in Databases (KDD)*. Among the most important business processes of energy providers are marketing strategies, such as tariff optimization and customer loyalty programs, as well as negotiations with market participants to allocate resources from energy producers.
For the latter aspect, energy providers rely on forecast models to plan for their future buy-in of energy. These forecast models employed by energy providers come in the form of *load profiles*. Due to policies by controlling authorities mandating the usage of load profiles, and since changes to the concept of load profiles may, depending on the implementation used by the energy providers, require a high amount of man-power to realize, we have opted to adhere to the existing specification as introduced in section 2.3. Thus, over the course of this chapter, we will present approaches that utilize KDD-techniques and Smart Metering data to build load profiles. By maintaining the existing concept of load profiles, it is not only significantly easier for the energy provider to adopt the resulting models, but it also allows to satisfy regulatory requirements such as the guideline to announce load profiles to all market participants at least 3 months prior to their usage [Bun13], meaning that the forecast models need to yield predictions with acceptable error rates many months in advance.

4.1 Description of datasets

To evaluate the performance of our approach to generate load profiles, which we will present in section 4.3, we have used three real-world datasets containing Smart Metering time series. These datasets have been visualized in figure 4.1.

The first dataset, which we will refer to as the *HHU-Dataset*, contains data for 7793 distinct customers. The data covers a time period of 65 months with a resolution of

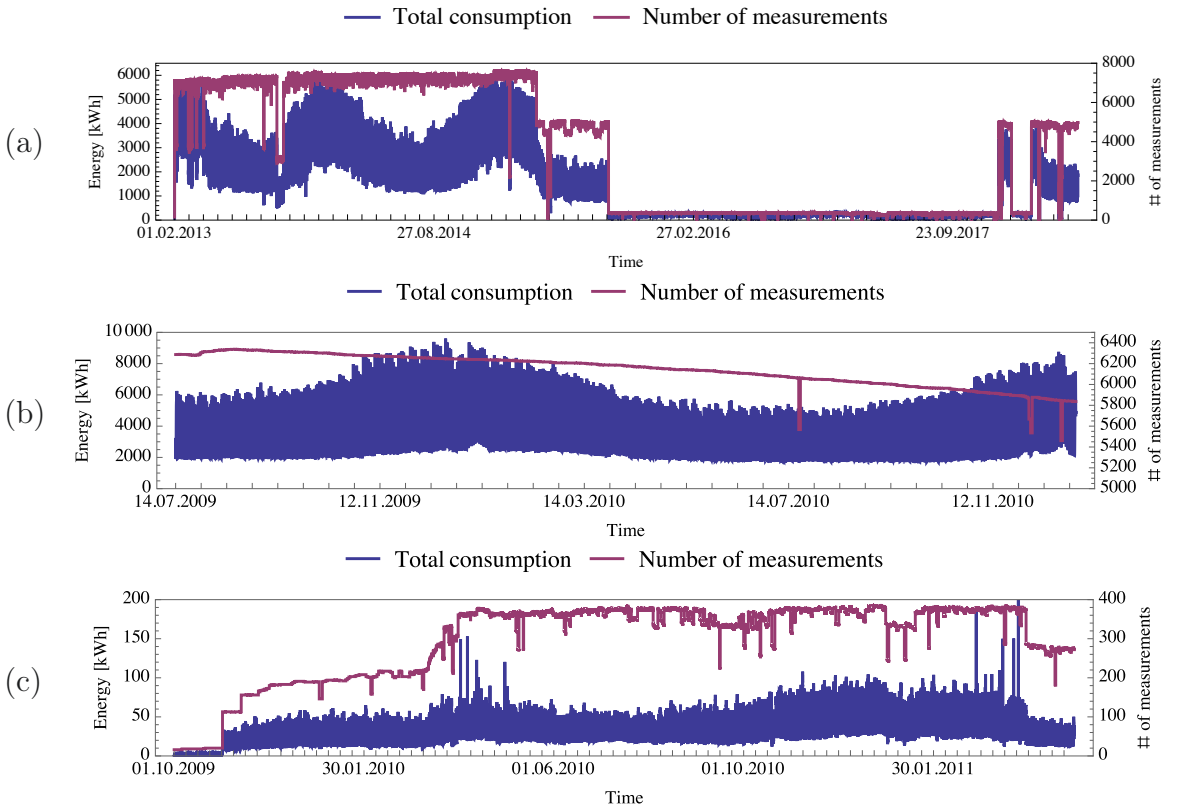


Figure 4.1: Overview of the (a) *HHU-Dataset*, (b) *CER-Dataset* and (c) *IZES-Dataset*. The blue colored graphs show the total energy consumption (primary axis); the purple colored graphs show the number of non-missing values per time slot (secondary axis).

1 measurement per hour. What is noteworthy about this dataset is that it has been made available in cooperation with a German electricity company who already had a complete rollout of smart meters, meaning that it has a customer base with a realistic ratio of end-consumers, industrial customers, agricultural customers, etc., enabling us to evaluate our approach under realistic conditions.

The second dataset, named the *CER-Dataset*, consists of a total of 6445 distinct Irish household customers with 1 measurement every 30 minutes over the course of 16 months. It is provided by the *Irish CER (Commission for Energy Regulation)* and accessed via the *Irish Social Science Data Archive (ISSDA)* [CER]. What makes this dataset special is that it is also accompanied by survey data, thus providing additional information about the consumers such as salary and civil status. We have presented some approaches that utilize the survey data provided by this dataset to extract useful knowledge about customers in section 3.3.2. For our experiments, we have opted to not incorporate this survey data into our analysis as in most real-world scenarios electricity companies are unlikely to have accurate and extensive survey data available to train their models.

The third dataset, which we will call the *IZES-Dataset*, covers the energy demand of 416 distinct household consumers with 1 measurement every 15 minutes over the course of 19 months. The dataset was gathered as part of the field test *"Moderne Energiesparsysteme im Haushalt" (modern energy saving systems in household environments)* and is made available by the *IZES institute* [Hof+12]. Similar to the *CER-*

Dataset, this dataset does only contain household customers who voluntarily participated in the study, meaning it does not represent a complete customer base which also includes industrial and agricultural customers, potential leading to a *selection bias* when trying to generalize results from the *CER-Dataset* or the *IZES-Dataset*.

One important property of all three datasets is that, for each customer, we have a single time series describing the energy consumption of the customer. According to [Zha+14], these datasets would therefore be classified as *None-Instrusive Load Monitoring (NILM)* as opposed to *Instrusive Load Monitoring (ILM)*, where an individual consumption time series is recorded for each appliance the customer owns.

For our evaluation, we have split each dataset into two disjoint sets, where one set has served as *training data* and the other set as *test data*. As for the *HHU-Dataset*, the first 23 months of Smart Metering data have been used as *training data*, while in the case of the *CER-Dataset* and *IZES-Dataset* the first 13 and 15 months of data have been used as *training data*, respectively. The remaining data of each dataset has been used as *test data*. This segmentation of *training data* and *test data* has remained static for all experiments presented in this chapter. More specifically, we have opted to not use *Cross-Validation* [Sto74] as part of our evaluation. This is because of the fact that we process Smart Metering time series, which means that on an application level using *Cross-Validation* would imply to use *future* data to predict *past* events, which is not a realistic use-case. Furthermore, as we will discuss in section 4.3 in more detail, the *training data* need to encompass at least 1 complete year, hampering the employment of *Cross-Validation* for the rather small *CER-Dataset* and *IZES-Dataset* in particular. Additionally, *Cross-Validation* makes it harder to account for the problem of *evolving distributions* in the dataset, an aspect which will discuss in more detail in chapter 5.

Since all of the datasets presented in this section contain real-world data, they are also subject to technical failures, such as temporary network transmission failures or any of the Smart Metering becoming faulty, thus requiring manual maintenance. In any of these cases, a *missing value* is introduced into the dataset. In figure 4.1, these *missing values* are indicated by a non-constant value for the graph representing the number of measurements on the secondary axis.
As we have stated in section 1.2.2, the mechanism behind *missing values* can have a significant influence on the process of *Knowledge Discovery in Databases* as a whole. Because of this, unless known through application knowledge, tests based on statistics can be used to ascertain which failure mechanism is present. In the case of the Smart Metering datasets used as part of our evaluation, we can suspect due to application knowledge that the MCAR failure mechanism is present, meaning that if a value in a given dataset is missing, the absence of the value is caused by technical deficiencies that are unrelated to the value itself and all other values of the corresponding datset. In the reverse case, assuming NMAR as the failure mechanism would indicate that Intelligent Metering Systems are more likely to fail depending on the value of the electric current measured by the metering device, while assuming MAR would imply that the likelihood of the device to fail is dependent on the value of the electric current measured at a different time, which might be earlier or even later than the missing value. To check for MCAR, a test based on the χ^2 test is presented in [Lit88]. Since this test requires the dataset to be a set of tuples rather than a set of time series, we perform the test on the processed datasets as discussed in section 4.3. The results of the test for the three datasets are given in table 4.1. Given the degrees of freedom and the value of

Dataset	degrees of freedom	χ^2 statistic
HHU	7.624	14.383,21
CER	20	3,12
IZES	19.438	189.709,26

Table 4.1: Overview of the results of the χ^2 test for the MCAR failure mechanism according to [Lit88].

the χ^2 statistic, we can conclude that the test in [Lit88] strongly rejects the hypothesis of the mechanism behind the missing values in the datasets being MCAR for the *HHU-Dataset* and the *IZES-Dataset*, while strongly accepting the hypothesis for the *CER-Dataset*. These results for the *HHU-Dataset* and the *IZES-Dataset* can likely be explained by real-world circumstances surrounding Smart Metering devices: unless the missing values are caused by temporary transmission failures due to networking problems, which might be automatically resolved until the next scheduled measurement, a missing value caused by a Smart Metering device becoming faulty is most likely to also affect immediately consecutive upcoming measurements. This causes the consumption time series to contain a block range of missing values, which typically end with a technician completing his or her maintenance work for the device. As such, missing values do not always occur completely at random, but are often lumped into blocks, which can be verified by manually looking at samples of missing values in the datasets. This is presumably the reason why the datasets failed the MCAR test. Nevertheless, since these missing value are neither reliant on the value of the missing nor observed values, the mechanism is still classified as MCAR according to the definition in equation 1.3.

4.2 Assessment of clustering quality

One of the most important properties of clustering is that it belongs to the group of techniques classified as *Unsupervised Learning*. This means that the clustering algorithm semi-automatically segments the elements of the given dataset into groups, with no *a priori* knowledge, such as *labels* or precategorized data, being incorporated into the clustering process. However, many clustering algorithms require an input parameter specifying the number of clusters to build. Since the number of clusters often has a significant influence on the quality of the clustering segmentation, it is typically desirable to have the *optimal* number of clusters being known in advance, for example due to application knowledge about the real-world state of affairs encoded in the dataset. Often however, only limited information about the real-world circumstances are available, translating to the optimal number of clusters being usually unknown. Combined with the fact that determining the optimal number of clusters is a difficult problem in and of itself, this has led to this issue gaining a lot of attention from researchers.
One of the ways to determine the optimal number of clusters and to make the quality of clustering segmentations more comparable is by the usage of so called *Cluster Validity Indices (CVI)*. By employing this technique, clustering becomes an iterative process nested within the iterative process of *Knowledge Discovery in Databases (KDD)*, where the clustering step of KDD is repeated for different values for the number of clusters; the resulting clustering segmentations are then evaluated and consolidated into one

key figure per clustering segmentation. By tracking the evolution of the key figure of a given Cluster Validity Index for different values for the number of clusters, the *optimal* number of clusters can be determined as the number of clusters where the key figure of the CVI has reached a global optimum over all tested numbers of clusters. In doing so, each CVI focuses on different aspects of the clustering segmentation, for example rewarding uniqueness in cluster assignment or incentivizing the creation of compact and well-separated clusters. As such, each CVI has different strengths and weaknesses in coping with overlapping, heterogeneous shaped or hierarchically structured clusters. Over the course of this section, we will introduce some common Cluster Validity Indices for Fuzzy Clustering often used to assess the quality of a clustering segmentation.

4.2.1 Partition Coefficient

The *Partition Coefficient (PC)* [Bez74] evaluates a given clustering segmentation based on the membership degrees $u_{o,i}$ of each data tuple to each clustering prototype and is defined as follows:

$$V_{PC}(U,c) = \frac{1}{N}\sum_{o=1}^{c}\sum_{i=1}^{N} u_{o,i}^2 \tag{4.1}$$

According to this CVI, the *best* clustering partitioning is the one where all tuples have been unambiguously assigned to a given clustering prototype, which translates to the memberships of all tuples being equal to 1 for one clustering prototype and 0 for all other prototypes since $u_{o,i} \in [0,1]$. In that case, $V_{PC} = 1$.
However, the *worst* clustering segmentation, according to this CVI, is present if all tuples belong to all clustering prototypes equally, which due to the Fuzzy Clustering property $\forall i : \sum_{o=1}^{c} u_{o,i} = 1$ is the case if $\forall o, i : u_{o,i} = \frac{1}{c}$ and thus $V_{PC} = \frac{1}{c}$. Because of this, V_{PC} has a range of $\left[\frac{1}{c}, 1\right]$.

Due to the fact that the range of V_{PC} is dependent on the number of clusters c, the CVI is biased towards a smaller value for c. To overcome this issue, the *Normalized Partition Coefficient (NPC)* has been proposed [Bac78; Rou78]:

$$V_{NPC}(U,c) = 1 - \frac{c}{c-1}\left(1 - V_{PC}(U,c)\right) = 1 - \frac{c}{c-1}\left(1 - \frac{1}{N}\sum_{o=1}^{c}\sum_{i=1}^{N} u_{o,i}^2\right) \tag{4.2}$$

Unlike V_{PC}, the range of V_{NPC} is $[0,1]$ regardless of the value of c. Like V_{PC}, a higher value of V_{NPC} indicates a better cluster segmentation.

While both V_{PC} and V_{NPC} are rather simple CVIs, possibly their biggest flaw is the property to yield good results only when the dataset consists of spherical and clearly separated clusters, or else the optimal number of clusters is very often underestimated [BWS06; Him16].

4.2.2 Compactness & Separation by Xie & Beni

The *Xie-Beni (XB)* Index [XB91] works by consulting the original objective of generic clustering algorithms, which instructs the algorithm to find a clustering segmentation so that data tuples belonging to the same cluster are as similar as possible while data tuples belonging to different clusters are as dissimilar as possible. For this purpose, the Xie-Beni Index uses two key figures to evaluate how good the clustering algorithm has

solved the objective criterion. The first key figure, called the *Compactness*, describes how similar data tuples are compared to their assigned clustering prototype:

$$Comp_{XB}(U, X, V) = \frac{J}{N} = \frac{\sum_{o=1}^{c} \sum_{i=1}^{N} u_{o,i}^2 \left\| x_i - v_o \right\|^2}{N} \tag{4.3}$$

The connection of the *Compactness* to the objective function of clustering algorithms is also illustrated by the fact that $Comp_{XB}(U, X, V)$ incorporates $J(\cdot)$, which corresponds to the objective function of Fuzzy-C-Means as described in equation 1.1. Because clustering algorithms strive to minimize the objective function, a good clustering segmentation according to this CVI is achieved when the *Compactness* is minimized as well.
The second key figure expresses the dissimilarity of clusters and is called the *Separation*:

$$Sep_{XB}(V) = \min_{\substack{1 \leq o,o' \leq c \\ o \neq o'}} \left\| v_o - v_{o'} \right\|^2 \tag{4.4}$$

Since clustering algorithms aim to maximize the dissimilarity of each pair of clusters, a high value for the *Separation* signalizes a good clustering segmentation.

To aggregate both key figures, the Xie-Beni Index V_{XB} is defined as the ratio of the *Compactness* and the *Separation*:

$$V_{XB}(U, X, V) = \frac{Comp_{XB}}{Sep_{XB}} = \frac{\sum_{o=1}^{c} \sum_{i=1}^{N} u_{o,i}^2 \left\| x_i - v_o \right\|^2}{N \min_{\substack{1 \leq o,o' \leq c \\ o \neq o'}} \left\| v_o - v_{o'} \right\|^2} \tag{4.5}$$

For the experiments presented in section 4.4 however, the definition of V_{XB} as given in equation 4.5 can not be used because the value of the CVI is dependent on the values of the data tuples x_i, which are partially unavailable if the dataset contains *missing values*. To handle this issue, we employ a modified definition of V_{XB} where the euclidean distance between x_i and v_o is replaced by their partial distance as described by [HHC11]. The resulting modified definition of V_{XB} is as follows:

$$V_{XB}(U, X, V) = \frac{\sum_{o=1}^{c} \sum_{i=1}^{N} u_{o,i}^2 \frac{h \sum_{n=1}^{h} I_{i,n} (x_{i,n} - v_{o,n})^2}{\sum_{n=1}^{h} I_{i,n}}}{N \min_{\substack{1 \leq o,o' \leq c \\ o \neq o'}} \left\| v_o - v_{o'} \right\|^2}$$
$$\text{with } I_{i,n} = \begin{cases} 1 & \text{if } x_{i,n} \text{ is not a missing value} \\ 0 & \text{else} \end{cases} \tag{4.6}$$

Although the design of V_{XB} is well thought out due to its definition being closely inspired by the objective function of clustering algorithms, it can be formally shown that the value of the CVI is monotonously decreasing when increasing the number of clusters c, which the authors of the CVI themselves have pointed out [XB91].

4.2.3 Compactness & Separation by Bouguessa, Wang & Sun

The idea behind the *Bouguessa-Wang-Sun (BWS)* Index [BWS06] is very similar to the Xie-Beni Index in that the Bouguessa-Wang-Sun Index also assesses the quality of the clustering segmentation by evaluating the ratio between two key figures called

Compactness and *Separation*. However, in contrast to V_{XB}, V_{BWS} is defined as the ratio of *Separation* divided by *Compactness* as opposed to V_{XB} where it is the other way around:

$$V_{BWS}(U, V, X) = \frac{Sep_{BWS}}{Comp_{BWS}} \tag{4.7}$$

Thus, a good clustering partition according to this CVI is characterized by a value for V_{BWS} which is as large as possible. The *Separation* of V_{BWS} is based on the *between-cluster fuzzy scatter matrix* S_B:

$$\begin{aligned} Sep_{BWS}(U, X, V) &= trace(S_B) \\ \text{with } S_B &= \sum_{o=1}^{c} \sum_{i=1}^{N} u_{o,i}^{m} (v_o - \overline{x}) (v_o - \overline{x})^{\top} \end{aligned} \tag{4.8}$$

Here, $\overline{x}$ is the means of all data tuples x_i. A large value for Sep_{BWS} indicates that the fuzzy clusters are well-separated. Similar to the *Separation* Sep_{BWS}, the *Compactness* $Comp_{BWS}$ is also based on the trace of matrices, more specifically the fuzzy covariance matrices for each cluster:

$$\begin{aligned} Comp_{BWS}(U, X, V) &= \sum_{o=1}^{c} trace(Cov_o) \\ \text{with } Cov_o &= \frac{\sum_{i=1}^{N} u_{o,i}^{m} (x_i - v_o) (x_i - v_o)^{\top}}{\sum_{i=1}^{N} u_{o,i}^{m}} \end{aligned} \tag{4.9}$$

Due to the CVIs definition of *Compactness* and *Separation*, particularly by its inclusion of the fuzzy covariance matrices, V_{BWS} aims to perform well if the clusters partially overlap or differ in their size, shape and density. In experiments with both synthetic and real-world datasets, V_{BWS} has achieved above-average results [BWS06; Him16].
However, because V_{BWS} is dependent on the value of the data tuples x_i, the original definition of V_{BWS} is undefined if there are *missing values* present in the dataset. Because of this, for our experiments, we employ a modified definition of this CVI according to [HCC12]:

$$\begin{aligned} Cov_{o(p,l)} &= \frac{\sum_{i=1}^{N} u_{o,i}^{m} \cdot I_{i,p} \cdot I_{i,l} (x_{i,p} - v_{o,p}) (x_{i,l} - v_{o,l})^{\top}}{\sum_{i=1}^{N} u_{o,i}^{m} \cdot I_{i,p} \cdot I_{i,l}} \qquad , 1 < p, l < h \\ \text{with } I_{i,n} &= \begin{cases} 1 & \text{if } x_{i,n} \text{ is not a missing value} \\ 0 & \text{else} \end{cases} \end{aligned} \tag{4.10}$$

In equation 4.10, $Cov_{o(p,l)}$ describes a modified version for the fuzzy covariance matrices used to compute the *Compactness* in equation 4.9. To compute the *Separation* of V_{BWS}, only the definition of $\overline{x}$ in equation 4.8 is altered:

$$\overline{x} = \frac{\sum_{i=1}^{N} I_{i,n} \cdot x_{i,n}}{\sum_{i=1}^{N} I_{i,n}} \qquad \text{with } I_{i,n} = \begin{cases} 1 & \text{if } x_{i,n} \text{ is not a missing value} \\ 0 & \text{else} \end{cases} \tag{4.11}$$

4.2.4 Fuzzy Hypervolume and Partition Density

Fuzzy Hypervolume (FH) [GG89] is a CVI that incentivizes the creation of clusters with minimal volumes. In doing so, the volume of a cluster is measured based on the fuzzy covariance matrix:

$$V_{FH}(U, X, V) = \sum_{o=1}^{c} \sqrt{det(Cov_o)} \tag{4.12}$$

The definition for the fuzzy covariance matrices we have employed is given in equation 4.10, which converges to the definition of Cov_o as in equation 4.9 in the absence of *missing values* in the dataset. Similar to V_{BWS}, by incorporating the fuzzy covariance matrices of each cluster, V_{FH} is able to recognize non-spherical, partially overlapping clusters with varying size, shape and density. Since the idea of V_{FH} is to build clusters with minimal volumes, a good clustering segmentation is achieved when this CVI is minimized.

One disadvantage of V_{FH} is that, by design, clusters are rated solely on their volume, discouraging the formation of *large* clusters. Thus, the authors of [GG89] propose the concept of *Partition Density (PD)*:

$$\begin{gathered} V_{PD}(U, X, V) = \frac{Z}{V_{FH}(U, X, V)} \\ \text{with } Z = \sum_{o=1}^{c} \sum_{i=1}^{N} u_{o,i} \ \forall x_i \in \left\{ x_i \,\middle|\, (x_i - v_o)^\top Cov_o^{-1} (x_i - v_o) < 1 \right\} \end{gathered} \tag{4.13}$$

Here, the term Z is described by the authors of [GG89] as the *sum of central members*. A *central member* as in equation 4.13 is a data tuple whose radius to the corresponding clustering prototype is less than one standard deviation of the overall size of the cluster as given by the fuzzy covariance matrix. The more *massive* each cluster is, that is, the more *central members* each cluster has and the higher the membership degree of each *central member* to said cluster is, the higher the value Z. Using Z, the authors of [GG89] aim to overcome the weakness of V_{FH} favoring clusters with minimal volumes and instead reward the formation of large clusters if they are sufficiently *massive*.
As in the case of V_{XB} and V_{BWS}, the definition of V_{PD} might be undefined if the dataset contains *missing values* as equation 4.13 accesses the values of x_i directly. Because of this, we have used an adaptation for our experiments according to an idea introduced in [HHC11]:

$$Z = \sum_{o=1}^{c} \sum_{i=1}^{N} u_{o,i} \ \forall x_i \in \left\{ x_i \,\middle|\, \frac{h \begin{pmatrix} I_{i,1}(x_{i,1} - v_{o,1}) \\ \vdots \\ I_{i,h}(x_{i,h} - v_{o,h}) \end{pmatrix}^\top Cov_o^{-1} \begin{pmatrix} I_{i,1}(x_{i,1} - v_{o,1}) \\ \vdots \\ I_{i,h}(x_{i,h} - v_{o,h}) \end{pmatrix}}{\sum_{n=1}^{h} I_{i,n}} < 1 \right\} \tag{4.14}$$

$$\text{with } I_{i,n} = \begin{cases} 1 & \text{if } x_{i,n} \text{ is not a missing value} \\ 0 & \text{else} \end{cases}$$

4.2.5 Silhouette Coefficient

Another well-known indicator for evaluating the quality of a clustering segmentation is the *Silhouette Coefficient (SC)* [Rou87; KR90; ES00]. The idea behind this index is to incorporate the *distance function* used during the clustering process to rate how unambiguous each data tuple belongs to their assigned cluster. For this purpose, the *Silhouette Coefficient* V_{SC} defines two utility functions:

$$a(i) = \frac{1}{|C_o| - 1} \sum_{\substack{x_i, x_{i'} \in C_o \\ i \neq i'}} dist(x_i, x_{i'}) \tag{4.15}$$

$$b(i) = \min_{o' \neq o} \frac{1}{|C_{o'}|} \sum_{\substack{x_i \in C_o \\ x_{i'} \in C_{o'}}} dist(x_i, x_{i'}) \tag{4.16}$$

Here, C_o is derived from the fuzzy memberships $u_{o,i}$ and describes the set of data tuples x_i which have been assigned to the o-th cluster. While $a(i)$ can be vividly described as the average distance of x_i to other members of the same cluster, $b(i)$ describes the average distance of x_i to members of its *"neighbor" cluster*.
With both equation 4.15 and 4.16, the *Silhouette Coefficient* V_{SC} is defined as follows:

$$V_{SC}(U, X) = \frac{1}{\sum_{o=1}^{c} |C_o|} \sum_{o=1}^{c} \sum_{x_i \in C_o} \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{4.17}$$

The range of V_{SC} is $[-1, 1]$, whereas a value closer to 1 means that, for most data tuples, the average distance of each data tuple to members of its *"neighbor" cluster*, expressed by $b(i)$, is much larger than the average distance each data tuple to other members of their assigned cluster, expressed by $a(i)$. Thus, a value of approximately 1 signifies a good cluster segmentation, where most data tuples unambiguously belong to the cluster they have been assigned to according to the dissimilarity measure of the clustering process.

4.2.6 Average Clustering Uniqueness

One difficulty in using the *Silhouette Coefficient (SC)* [Rou87; KR90; ES00] introduced in section 4.2.5 arises when applying V_{SC} on increasingly large datasets. Due to the fact that V_{SC} is defined as the average of the $\frac{b(i)-a(i)}{\max(a(i),b(i))}$ for all data tuples x_i, it is required to compute the distances between each pair of data tuples $dist(x_i, x_{i'})$ with $i \neq i'$. This is because for all $i \neq i'$, the term $dist(x_i, x_{i'})$ is either part of $a(i)$ as in equation 4.15 if both x_i and $x_{i'}$ belong to the same cluster or part of $b(i)$ as in equation 4.16 to determine which of the other clusters is the *"neighbor" cluster*. Overall, this results in a time complexity of $O(\mathcal{N}^2)$ to compute V_{SC}, where $\mathcal{N}$ is the number of data tuples in the dataset. This causes V_{SC} to become infeasible to compute when the dataset is sufficiently large.
To counteract this issue, we introduce a new CVI in this section, which we will refer to as *Average Clustering Uniqueness (ACU)*, which is based on the idea behind V_{SC}. For this purpose, we redefine the utility functions $a(i)$ and $b(i)$ as follows:

$$a(i) = dist(x_i, v_o) \qquad , x_i \in C_o \tag{4.18}$$

$$b\left(i\right) = \min_{o' \neq o} dist\left(x_i, v_{o'}\right) \qquad , x_i \in C_o \tag{4.19}$$

As in section 4.2.5, C_o describes the set of data tuples belonging to the o-th cluster, represented by the clustering prototype v_o. We then define the *Average Clustering Uniqueness* V_{ACU} as follows:

$$V_{ACU}\left(U, X, V\right) = \frac{1}{\sum_{o=1}^{c} |C_o|} \sum_{x_i} \frac{b\left(i\right) - a\left(i\right)}{\max\left(a\left(i\right), b\left(i\right)\right)} \tag{4.20}$$

Though very similar to V_{SC}, the V_{ACU} is based on the distance to the clustering prototypes v_o instead of the set of data tuples assigned to the o-th cluster. This causes the time complexity of V_{ACU} to be $O\left(\mathcal{N} \cdot c\right)$, where $\mathcal{N}$ corresponds to the total number of data tuples and c is equal to the number of clustering prototypes. Since one has $c \ll \mathcal{N}$ in a typical clustering scenario, V_{ACU} can be used as an efficient approximation of V_{SC} in cases where the clusters are known to be of spherical shape.

4.3 Framework for generating load profiles

Due to the design of load profiles as introduced in section 2.3, the construction of such load profiles consists of several steps which are processed in sequence. On a high level, these steps conform to our previous work [Boc16; Boc17; Boc18] and can be summarized as follows:

1. **Day-type segmentation:** As the first step in the construction of load profiles, the optimal number of day-types as well as their segmentation onto the individual calendar days are determined.

2. **Identification of typical consumption patterns:** For each day-type, identify customer groups and derive one representative consumption pattern for each customer group.

3. **Compilation of load profiles:** Compile the load profiles by combining the results from the previous steps and assign a load profile to each customer.

4.3.1 Day-type segmentation

For determining a good day-type segmentation, the most crucial aspect is to identify groups of calendar days on which the total energy consumption of customers is sufficiently similar. Since load profiles are used by energy providers to forecast the total energy demand as accurately as possible, the careful choice of an optimal day-type segmentation is arguably the most crucial factor in achieving that goal. That is, if it is plausible to assume that the total energy demand of customers does not significantly differ on two or more given calendar days, there is no reason to use a different daily forecast, even if that forecast does not satisfy the actual consumption of individual customers.
To derive a good day-type segmentation during our experimental evaluation introduced in section 4.4, we have opted to first construct a new time series $X = \{x_1, x_2, ..., x_T\}$

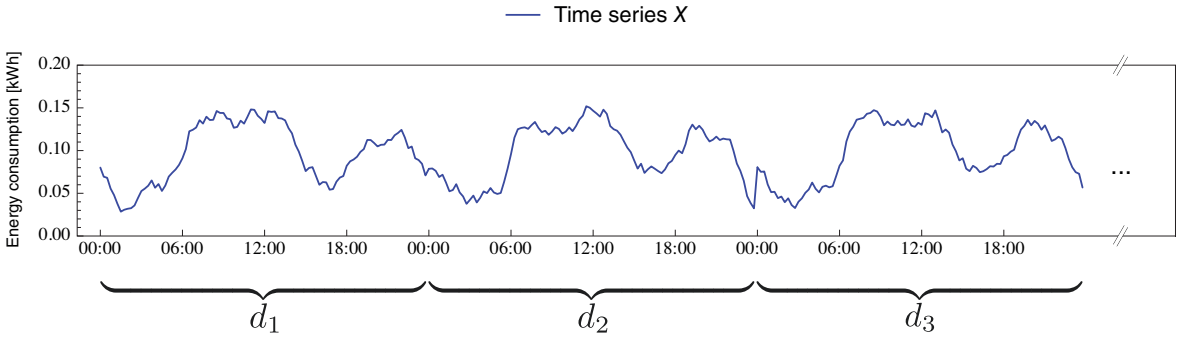


Figure 4.2: Visualization of how the Dataset D is constructed from the time series X by splitting the time series into tuples at the start of a new calendar day.

for each point in time $t_j, 1 \leq j \leq T$ using the individual customer's Smart Metering time series $s_{i,j} \in S_i, 1 \leq i \leq N, 1 \leq j \leq T$ as follows:

$$x_j := \frac{1}{N_j \cdot DynFactor\,(t_j)} \sum_{i=1}^{N} \frac{s_{i,j}}{YCF_{i,j}} \tag{4.21}$$

Here, N corresponds to the total number of distinct customers and N_j is equal to the number of measurements available for t_j. The distinction between N and N_j is necessary so as to not distort X when *missing values* are present in the data or when customers leave or join to or from other energy providers. The term $YCF_{i,j}$ stands for the *Year Consumption Forecast* assigned to the i-th customer, represented by his or her Smart Metering time series S_i, for the year that t_j belongs to. The optional term $DynFactor\,(t_j)$ corresponds to the *dynamization function* evaluated at t_j; in the case that no *dynamization function* is used, the term is omitted. What is important to note here is that $s_{i,j}$ describes the amount of energy consumed since the previous measurement $s_{i,j-1}$, not the cumulative consumption such as the meter reading. Although it is technically possible to record and process $s_{i,j}$ as power (*kilowatts*) instead of energy (*kilowatthours*), this requires correction terms in order to work properly with the concept of the *Year Consumption Forecast*. Thus, for equation 4.21 and following, we notate $s_{i,j}$ and related terms as energy. In a nutshell, X can be described as an average time series of the normalized Smart Metering time series of all customers.

We then construct dataset D using X as follows:

$$D := \left\{ d_l := (x_j, ..., x_{j+H-1}) \;\middle|\; \begin{array}{c} \forall a \text{ with } 1 \leq j \leq a \leq j+H-1 \leq T: \\ t_a \text{ belongs to the } l\text{-th calendar day} \end{array} \right\} \tag{4.22}$$

Here, H corresponds to the number of measurements per day. In general, each tuple d_l corresponds to a slice of the time series X containing data for one calendar day. In doing so, we account for application knowledge, which states it is reasonable to think of Smart Metering time series not as an atomic data source, but as a set of tuples, where each tuple describes the consumption time series of a single calendar day. Since Smart Metering time series typically have a resolution of one measurement every 15, 30 or 60 minutes, each d_l corresponds to a 96-, 48- or 24-tuple, respectively. Figure 4.2 visualizes how the d_l are derived using X.

As the second to final step in deriving a good day-type segmentation, we apply clustering on the dataset D. In principle, an arbitrary clustering algorithm can be employed for this task, such as the one presented in section 1.2.2. However, it is important

to recall that the goal of this step is to group calendar days where the customer base as a whole behaves as similar as possible. Even if the consumption behaviors of consumers vary widely only in a few number of dimensions, these differences contribute to the financial risk associated with *imbalance energy*. Because of this, even if ellipsoid clusters or clusters of other shapes were to be recognizable in the data, a further segmentation of such ellipsoid clusters into multiple spherical clusters is more desirable when incorporating application knowledge. Thus, the chosen clustering algorithm should discourage the formation of non-spherical clusters. This requirement removes some algorithms from being considered, for example density-based clustering algorithms such as *DBSCAN* [Est+96], which are able to identify clusters of arbitrary shape. For our experiments, we have opted to use *Fuzzy-C-Means* [Bez81], a partitioning clustering algorithm with the tendency to recognize only spherical clusters [BWS06]. Additionally, Fuzzy-C-Means has the advantage of being a well-known and comparatively simple algorithm, which helps us in creating an approach with low computational requirements that is easy for energy providers to adopt and can be feasibly employed locally without relying on cloud computing service providers where compliance with data privacy policies are often a concern.

Once a good clustering segmentation is available, a day-type segmentation can be deduced by knowing which d_l got assigned to the same cluster and which calendar days each d_l represents. Starting out from this, an analyst examines regularities in the clustering segmentation, therefore deriving rulesets to classify future calendar days. At the same time, the analyst has to keep the strict periodicity of load profiles of 1 year in mind. That is, if for example multiple years of Smart Metering time series are being analyzed and a cluster only appears in a single year, the analyst has to take a close look at the corresponding calendar days. If the suspicion of these calendar days being an outlier day-type solidifies, for example caused by major sport events such as the *FIFA World Cup*, the analyst has to prune these calendar days from the evaluation and decide on appropriate replacement day-types. For the same reason, the size of the training data has to cover at least 1 year of continuous data to ensure that each day of year is present during analysis.

4.3.2 Identification of typical consumption patterns

The identification of typical consumption patterns is carried out after the day-type segmentation is finalized. With the set of day-types known, the goal of this process is to find representative consumption patterns for groups of customers for each day-type. Since load profiles expect the consumption pattern to represent all customers associated to the same customer group as a whole rather than individually, the same consumption patterns can be used as long as it matches the total energy consumption as accurately as possible. As a result of the identification of typical consumption patterns taking place after the day-type segmentation, it is already known which calendar days exhibit a sufficiently similar total energy consumption, namely all calendar days assigned to the same day-type. Because of this, our goal for this step is to look at the Smart Metering time series for each day individually in order to deduce day-type-specific consumption patterns for each customer. For this purpose, let $K_n, 1 \leq n \leq L$ be the sets of day-types from section 4.3.1 where each K_n consists of the corresponding t_j.

Algorithm 2 Compiling load profiles

Input: $S_i, P_n, K_n, V_{o,n}, U_n$

Output: set of all load profiles G, set of profile assignments Z

1: $G \leftarrow \emptyset$
2: $Z \leftarrow \emptyset$
3: **for** $i = 1$ **to** N **do** // for each customer
4: **for** $n = 1$ **to** L **do** // for each day-type
5: $A[n] \leftarrow V_{o,n}$ with $o := \arg\max_{o'} \left| \left\{ p_{e,n} \;\middle|\; \begin{array}{c} \exists j : (y_{i,j}, ..., y_{i,j+H-1}) = p_{e,n} \\ \wedge \nexists o'' : u_{o'',e,n} > u_{o',e,n} \end{array} \right\} \right|$
6: **end for**
7: $G \leftarrow G \cup A$
8: $Z \leftarrow Z \cup (S_i, A)$
9: **end for**
10: **return** G, Z

We then construct the disjoint sets $P_n, 1 \leq n \leq L$ as follows:

$$P_n := \left\{ p_{e,n} := (y_{i,j}, ..., y_{i,j+H-1}) \;\middle|\; \begin{array}{c} \forall a, b \text{ with} \\ 1 \leq j \leq a, b \leq j + H - 1 \leq T : \\ t_a, t_b \in K_n \text{ and } y_{i,a}, y_{i,b} \text{ belong to} \\ \text{the same calendar day} \end{array} \right\} \tag{4.23}$$

$$\text{with } y_{i,j} := \frac{s_{i,j}}{YCF_{i,j} \cdot DynFactor(t_j)}$$

Each dataset P_n contains all normalized Smart Metering data belonging to the n-th day-type as an H-dimensional tuple, similar to the dataset D in equation 4.22. We then process with clustering each of the P_n individually with a centroid-based clustering algorithm. The reason for requiring a centroid-based clustering algorithm is that the optimal clustering prototypes $V_{o,n}, 1 \leq o \leq c_{n,optimal}$ for each P_n directly correspond to the desired typical consumption patterns for the day-type K_n. Note that each P_n contains all corresponding meter readings of customers; this is in contrast to the methodology of other researchers like [VVS15] who compute an aggregated value, such as the *arithmetic mean* or the *median*, per customer and per time of day prior to applying clustering on this aggregated data.

4.3.3 Compilation of load profiles

With both the day-type segmentation and customer groups known, these results can be combined to form load profiles as introduced in section 2.3. While doing so, it is important to recall that the Year Consumption Forecast as well as the dynamization function are managed separately from the actual load profiles and are thus not considered during this step. By themselves, however, the load profiles are not very useful, as energy providers also require a means to assign *the best* load profile to each customer. For this purpose, we propose to employ a *majority vote* where for each day-type each customer gets the consumption pattern $V_{o,n}$ designated to which he or she has gotten most often assigned to during clustering. This procedure is outlined in algorithm 2 as pseudocode. In the process of constructing the set of all load profiles G and the set of all load profile assignments Z, algorithm 2 uses a L-dimensional helper variable A

where $A[n]$ stores the optimal consumption pattern for the current customer for the day-type K_n.

4.3.4 Assessment of load profiles

For the most part, we evaluate the load profiles generated during our experiments in section 4.4 according to the real-world requirements and policies as outlined in section 2.3. Even so, minor adjustments still need to be made in order to accommodate for the fact that our experiments are synthetic.

One of those adjustments concerns the customer-specific *Year Consumption Forecast*. Usually, energy providers assign each customer their total energy consumption of the past year as the estimated energy consumption for the upcoming year. In cases where no past year consumption is available, for example as it is the case for new customers, a default value is chosen for the first year. For our experiments, adhering to this procedure would result in rendering the first year of each of the datasets unusable as the Year Consumption Forecast would be undefined, severely reducing the available training and test data. Because of this, for our experiments, we have opted to change equation 2.1 so that the Year Consumption Forecast of a given year for a given customer is equal to his or her actual total energy consumption for the same year:

$$\begin{aligned} YCF_{2016} &= \text{actual total consumption in 2016} \\ &= \text{reading}\,(01.01.2017) - \text{reading}\,(01.01.2016) \end{aligned} \tag{4.24}$$

Though this handling with regard to the Year Consumption Forecast introduces a minor inaccuracy compared to the real-world usage of load profiles, it helps us in taking full advantage of the complete data that is made available to us. Similarly, if the consumption time series for a given customer does not start at the beginning of the calendar year or terminates before the end of the year, for example because the consumer has joined another energy provider as a new customer in the middle of the year or because the dataset itself starts on a date other than the beginning of the year, we extrapolate the Year Consumption Forecast for a given customer by computing his or her total energy consumption and multiply it by the number of calendar days where measurements should be available divided by the number of calendar days where measurements are actually available for said customer.

Another adjustment when evaluating the performance of the load profiles is accounted for by the fact that our datasets contain *missing values*. In a real-world scenario, by monitoring the load on the electricity grid, the total energy consumption is accurately known by the energy provider in hindsight. Due to customers consuming energy continuously even if some individual Smart Meter readings are unavailable due to transmission errors or technical failures in the Intelligent Metering Systems, the actual total energy consumption is higher than the sum of available Smart Metering measurements per time slot. In contrast, for our experiments the total energy consumption time series is simulated to be equal to the sum of available Smart Metering measurements per time slot. Because of this, to prevent the presence of *missing values* from skewing the accuracy of our evaluation, we omit the value of a consumption forecast if we were to compare the forecasted value to a *missing value* in the Smart Metering time series for a given customer and time slot; in other words, we assess the accuracy of load profiles only for customers and time slots where the actual data does

not contain a *missing value*. This procedure also elegantly allows us to cope with the problem of consumers leaving or joining the energy provider by assuming *missing values* before the customer joins as a new customer or after the customer has left the energy provider. As such, if there is no consumption data available for a given customer in the training data and is only observed in the test data, for example because said customer has only recently joined the energy provider as a new customer, the customer is omitted from the assessment of the load profiles completely. In a real-world scenario, new customers could possibly be accounted for by legacy load profiles until enough Smart Metering data is available to classify them to an existing load profile that better suits their consumption behavior.

With both the load profile forecast known, we are able to assess the accuracy of the forecast by comparing them to the actual total energy consumption. To make the performance of load profiles comparable when evaluating energy providers of different sizes, it is expedient to look at the *relative forecast error*, defined as the ratio of the absolute difference between the actual energy consumption and the predicted load yielded when using the load profiles to the actual consumption:

$$\text{relative forecast error} = \frac{|\text{forecast} - \text{actual demand}|}{\text{actual demand}} \tag{4.25}$$

During our experimental evaluation, we grade the performance of the load profiles according to their *relative forecast error*. As a baseline, the *BDEW Standard Load Profiles* typically achieve an average *relative forecast error* of approximately 12 to 14 percent as shown in figure 2.5.

Due to the method of how the load profiles are built, they serve two purposes. On the one hand, they can be used to forecast the actual energy demand in accordance with the requirements of energy providers as discussed in section 2.3. On the other hand, knowing which customer has gotten assigned to which load profile aids in deriving target-group-specific tariff offers, akin to the approaches presented in section 3.3.2 where this is the main goal. The reason why load profile assignments can be useful for compiling such tariff offers is because, by definition, all customers which have gotten assigned to the same load profile have exhibited roughly the same consumption behavior, meaning that one can assume those customers have the very similar requirements regarding energy consumption and might be interested in the same tariff offers. As such, an energy provider might decide to conduct a poll only for a small sample of customers from each set of customers per load profile and extrapolate the poll results to gather data for decision making processes. In the case of the experiments presented over the course of this chapter, we will focus on assessing the load profiles according to the *relative forecast error* of the forecast of the actual total energy demand.

4.4 Evaluation

Over the course of the following sections we present the evaluation results for the framework presented in section 4.3 as a means to predict the actual total energy consumption of the customers. In doing so, each section introduces and evaluates concretizations or slight modifications to the base framework.

4.4.1 Evaluation using the Euclidean distance

4.4.1.1 Experimental Setup

To evaluate the performance of generating load profiles using Smart Metering time series, we have employed the framework outlined in section 4.3 with some additions according to our previous work [Boc16; Boc17]. Specifically, when using real-world data as part of a Data Mining process, the data is only rarely available completely; often, some measurements are at least partially unobserved. In the case of Smart Metering time series, possible causes for *missing values* include temporary network transmission errors or Intelligent Metering Systems becoming faulty. In section 1.2.2, we have given a brief introduction to possible strategies that enable a clustering algorithm to cope with *missing values* in the data [HB01; SL01]. In general, the three approaches to account for incomplete data presented in that section are *Adaption of analysis methods*, *Complete-Case Analysis* and *Imputation of missing values* [LR02]. For the experiments presented in this section, we have opted to use an *adaption of analysis methods* in order to be able to handle Smart Metering data containing *missing values*. More specifically, we have decided to employ the *Partial Distance Strategy* for Fuzzy-C-Means to incorporate processing data with *missing values*, which uses the following distance function [HB01]:

$$dist_{PED}(a,b) = \frac{H}{I} \cdot \sqrt{\sum_{n=1}^{H} (a_n - b_n)^2 \cdot I_n}$$
$$\text{with } I_n = \begin{cases} 1 & \text{if neither } a_n \text{ nor } b_n \text{ are missing values} \\ 0 & \text{else} \end{cases} \quad \text{and } I = \sum_{n=1}^{H} I_n \tag{4.26}$$

The reason we have chosen the *Partial Distance Strategy* is that this approach has achieved the overall best results next to *Optimal Completion Strategy* in experiments [Him16].

Another important aspect to point out is the premise to use a distance function based on the euclidean distance for our evaluation. Since the underlying type of the datasets we experiment on is that of time series fundamentally, an intuitive approach would be to employ dissimilarity measurements more commonly associated with processing time series data, for example *Dynamic Time Warping (DTW)* [BK59] or one of the many extensions and approaches based on DTW such as *Windowing* [BC94], *Derivative Dynamic Time Warping (DDTW)* [KP01], *Slope Weighting* [SC78; KL83], *Step Patterns* [Ita75; MRR80] or *Fast-DTW* [SC07]. However, on an application level this could cause meter readings from 8 o'clock to be compared to meter readings from 9 o'clock, which is what we aim to prevent. For instance, consider two office employees with the same consumption behavior where the only difference between the two customers is that one employee starts working at 8 o'clock while the other employee starts working at 9 o'clock; when a dissimilarity measurement based on DTW were to be used, the algorithm would likely assign a distance of almost 0 between both employees with a corresponding warping path for measurements between 8 o'clock and 9 o'clock, causing both customers to likely be assigned to the same cluster. Due to the application knowledge outlined in section 2.3 however, the time slots where the energy consumption is measured are distinctive attributes, meaning it is undesirable for the energy forecast for a given time of day to be influenced by meter readings measured at different times

of day. To accommodate for this fact, we recommend to choose a distance function for the clustering process which compares values only against values of the same dimension, for example by using a distance function based on the L_p-norm, sometimes also referred to as the *Minkowski distance*, which we have introduced in equation 1.4 in section 1.2.3. For $p = 2$, equation 1.4 produces the well-known euclidean distance on which the *Partial Euclidean Distance* $dist_{PED}(a, b)$ in equation 4.26 is based upon.

Though the Fuzzy-C-Means algorithm in conjunction with the *Partial Distance Strategy* works deterministically once the starting configuration of the clustering prototypes is given, said starting configuration of the clustering prototypes, along with other parameters such as the number of clusters and the termination condition, can have a major impact on the quality of the clustering result. We have outlined this set of facts in section 1.2.2. A common approach to generate the initial clustering prototypes, as stated in line 1 of algorithm 1, is to choose random values in the feature space for the coordinates for the clustering prototypes. In this thesis, we refer to this strategy as *Random Coordinates*. To avoid a negative impact on the final clustering segmentation by an unfortunately chosen starting configuration, the clustering process is typically repeated multiple times with a different starting configuration in addition to the clustering process being repeated multiple times for different values for the number of clusters c. However, instead of relying only on *Random Coordinates*, other streategies to generate a starting configuration, such as *K-Means++* [AV07], are conceivable. Contrary to what the name might suggest, *K-Means++* is not a modified version of the *K-Means* [Mac+67] clustering algorithm. Instead, *K-Means++* aims to choose better initial clusters for *K-Means* and similar clustering methods by repeatedly nominating a data object for the position of a new initial clustering prototype until c cluster have been chosen, with the probability of choosing a given data object as the position for the next clustering prototype being weighted according to the square of the minimum distance to an already chosen clustering prototype. For our experiments, we have opted to test our approach using *K-Means++* for the day-type segmentation while using *Random Coordinates* as well as *K-Means++* separately for the identification of typical consumption patterns.

4.4.1.2 Results

The results of our experiments using the *Partial Euclidean Distance* $dist_{PED}(a, b)$ as given in equation 4.26 are shown in figure 4.3. These graphs show the *relative forecast errors* in percent as defined in equation 4.25 for the load profiles yielded when using different values for the number of day-types and consumption patterns according to the framework outlined in section 4.3.1 and 4.3.2, respectively. The load profiles themselves have been based on the consumption patterns which the Cluster Validity Indices have considered *optimal* among 20 independent iterations of Fuzzy-C-Means per either *K-Means++* and *Random Coordinates* and per combination of *number of day-types* and *number of consumption patterns*.

For the *HHU-Dataset*, this approach has achieved relative forecast errors roughly in the range from 8 percent up to 18 percent. Compared to the *BDEW Standard Load Profiles*, which usually yield a relative forecast error of between 12 and 14 percent (see figure 2.5 and [SM17]), this shows that the resulting load profiles might perform worse than the baseline if their parameters, that is the number of day-types and the number of consumption patterns, are not chosen carefully. When using the *K-Means++*

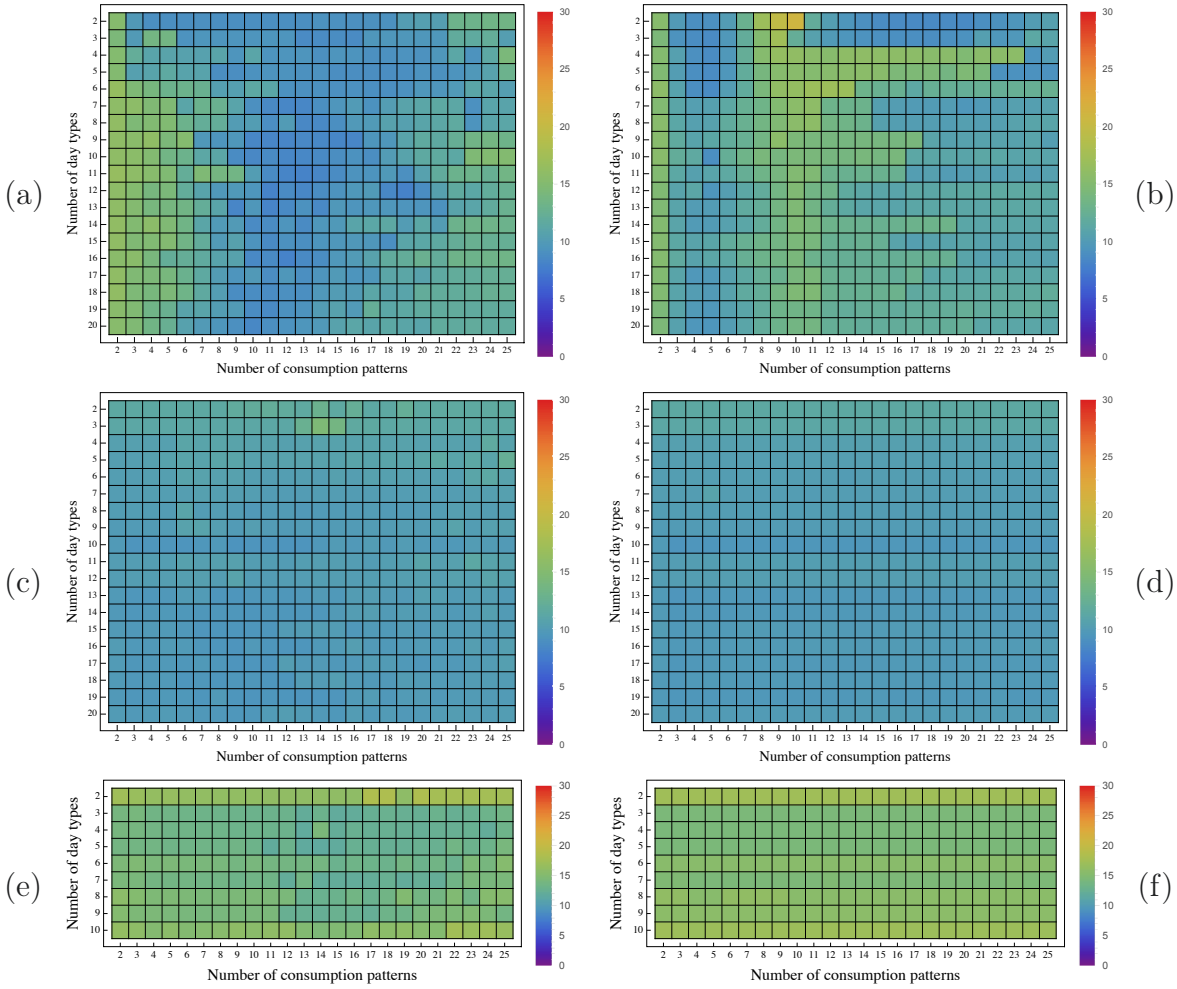


Figure 4.3: Ratio of the deviations and the actual consumption in percent yielded by the load profiles generated using different values for the number of day-types and consumption patterns using the experimental setup described in section 4.4.1.1. The graphs visualize the results for (a)(b) the *HHU-Dataset*, (c)(d) the *CER-Dataset* and (e)(f) the *IZES-Dataset* using (a)(c)(e) *K-Means++* and (b)(d)(f) *Random Coordinates* to generate the starting configuration of the clustering process.

cluster initialization method, the results in conjunction with a value for the number of consumption patterns in the range from 9 to 13 have been among the best of all tested values, whereas when using *Random Coordinates*, the same settings have yielded comparatively bad results.

For the *CER-Dataset* and the *IZES-Dataset*, this phenomenon could not be observed as those results have remained relatively constant regardless of the number of day-types, the number of consumption patterns or the cluster initialization method. Here, the *relative forecast errors* are roughly in the range from 9 to 11 percent for the *CER-Dataset* and 13 to 15 percent for the *IZES-Dataset*. Among the possible reasons for this behavior are the fact that both the *CER-Dataset* and the *IZES-Dataset* consist of only household customers, whereas the *HHU-Dataset* also contains non-household customers. Another possibility is conveyed when looking at figure 4.4 and 4.5, which visualize a comparison between the actual consumption and the forecast as predicted by the load profiles. From these graphs it can be seen that, especially by the example

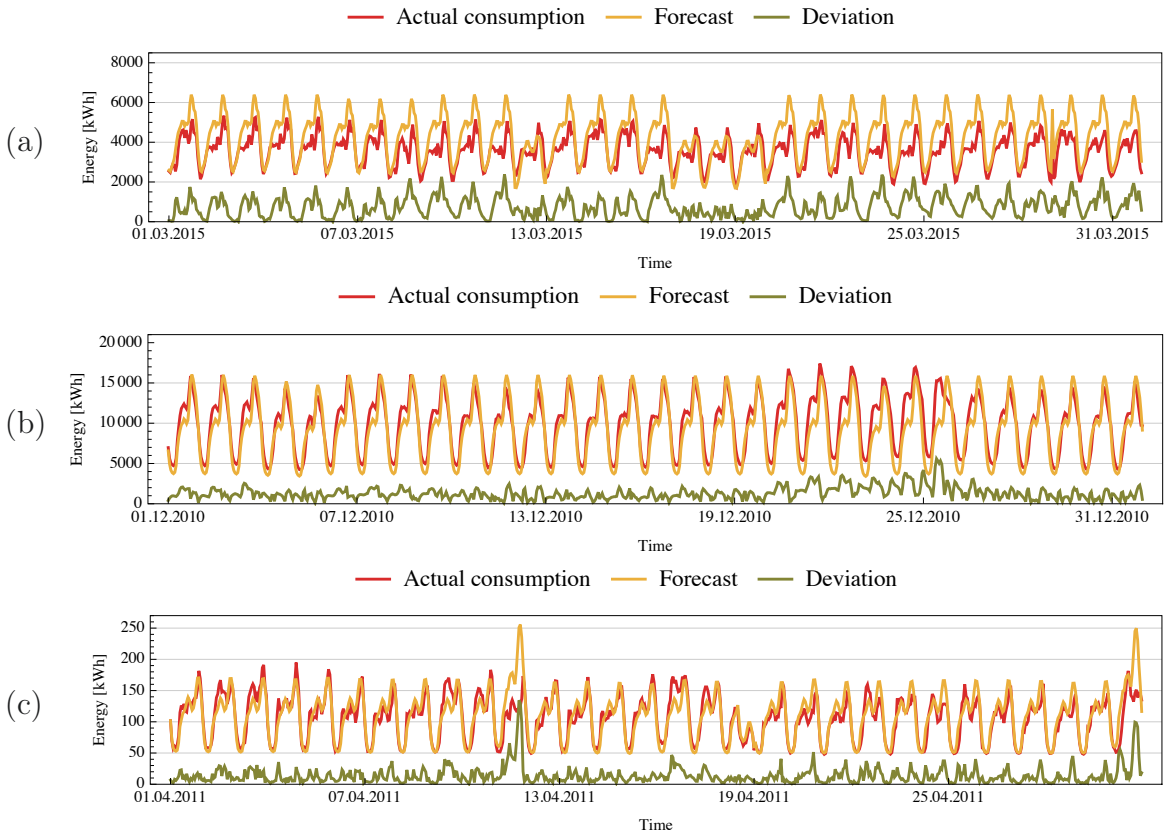


Figure 4.4: Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on 2 day-types and 2 consumption patterns per day-type and using *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.1.1 for (a) the *HHU-Dataset*, (b) the *CER-Dataset* and (c) the *IZES-Dataset*. The green graph shows the absolute deviation of the forecast from the actual consumption.

of the *CER-Dataset*, household customers tend to exhibit approximately the same consumption behavior every day on a large scale, regardless of whether it is a weekend or a working day. For all tested datasets, the generated load profiles have broadly recognized the shape of the typical consumption behavior correctly. In the case of the *HHU-Dataset* however, the load profiles seem to have slightly overestimated the actual consumption. Since this anomaly could not be detected for neither the *CER-Dataset* nor the *IZES-Dataset*, it can be assumed that overlapping clusters and clusters of unequal size in the dataset are among the possible reasons for this observation.

Overall, though the results of our approach can compete with the accuracy of the *BDEW Standard Load Profiles*, we have yet to discuss the mechanisms behind the day-type segmentation, which have to be used by an energy provider to use the correct consumption pattern on a given calendar day. Using the example of the *HHU-Dataset* and the *CER-Dataset*, the day-type segmentations for 4 and 8 day-types are visualized in figure 4.6. One outstanding property of these segmentations is that when the number of day-types is small, the segmentation roughly resembles seasonal segmentations that have been manually chosen by other researchers, for example in [VVS15; Sch+15]. However, by further increasing the number of day-types, the day-type segmentation of

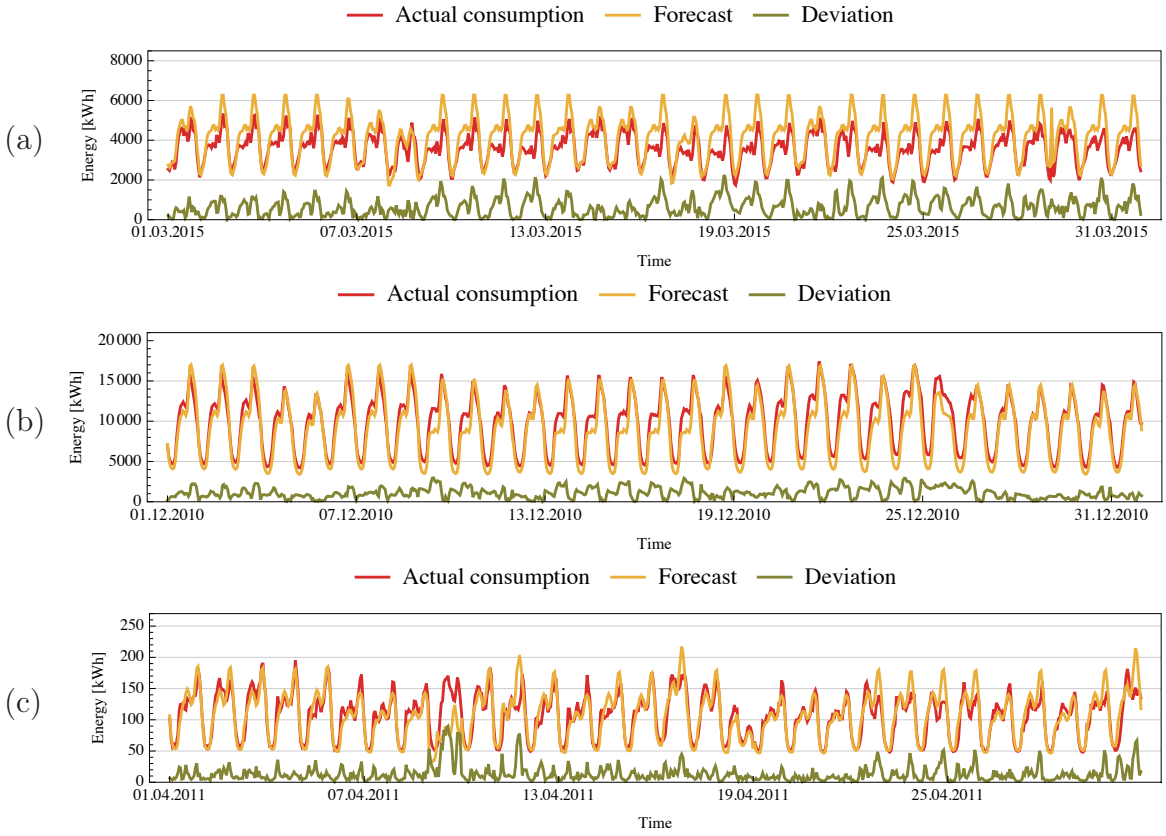


Figure 4.5: Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on 10 day-types and 25 consumption patterns per day-type and using *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.1.1 for (a) the *HHU-Dataset*, (b) the *CER-Dataset* and (c) the *IZES-Dataset*. The green graph shows the absolute deviation of the forecast from the actual consumption.

the datasets start to resemble that of a threshold filter, where the total energy consumed on a given calendar day has a more significant impact on the day-type classification than the shape of the consumption as a result of the behavior of customers. Since real-world consumption time series, such as the ones presented as part of our evaluation, tend to have seasonal patterns with a recognizable peak near the end of a year, modifying our approach to take advantage of the regularity of these patterns is a promising optimization which we will pursue in the following sections.

4.4.2 Evaluation using the Manhattan distance

4.4.2.1 Experimental Setup

Some of the main aspects we want to take advantage of in order to further improve upon the approach presented in section 4.4.1 are the regular and often sine-like patterns that become visible when looking at the total energy consumption of real-world Smart Metering datasets. For this purpose, we can make use of the fact that the specification of load profiles includes the concept of a *dynamization function*, a tool that allows to model and thus exclude the recurrent patterns during the analysis. The idea behind this

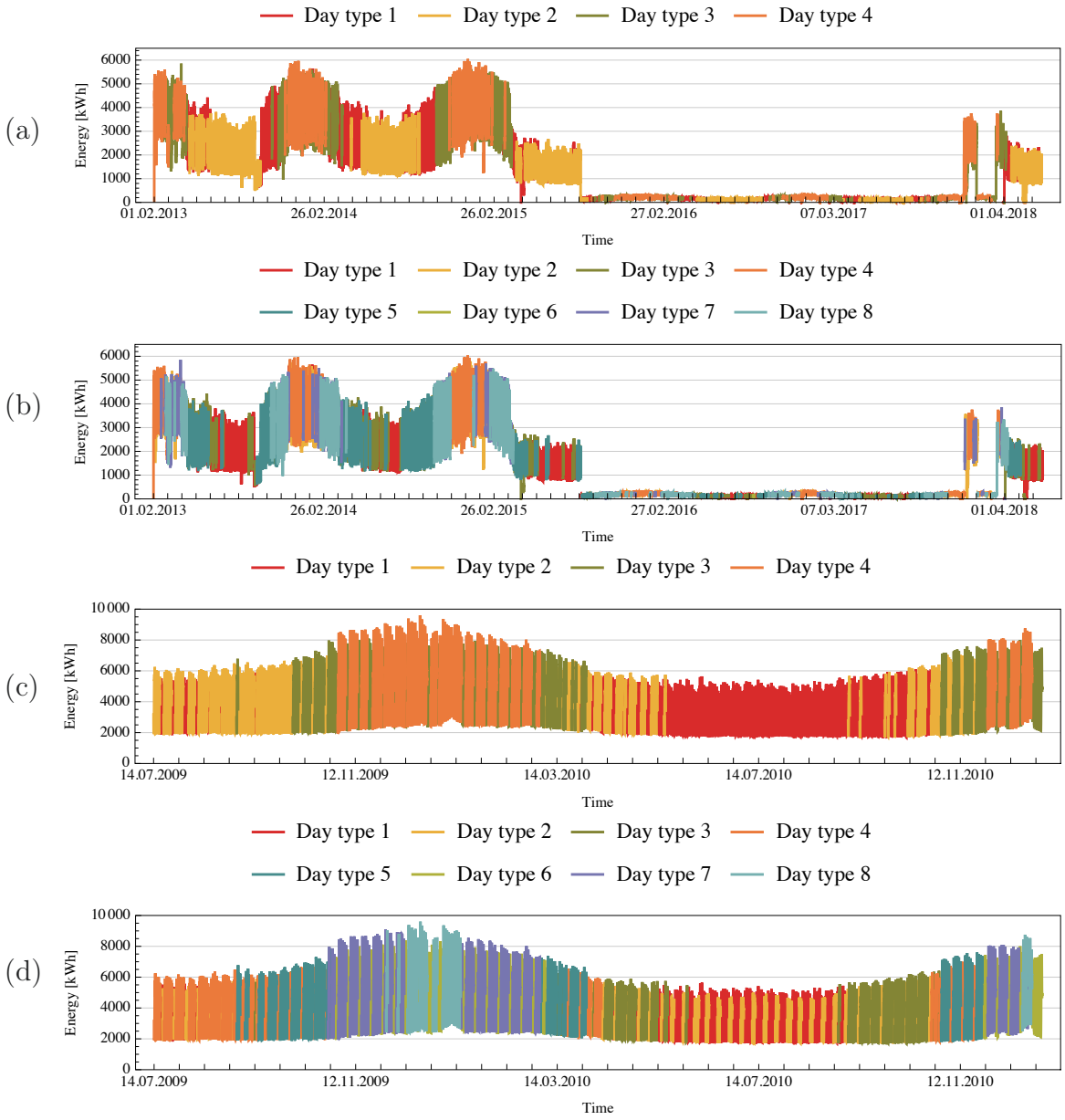


Figure 4.6: Overview of the day-type segmentations for the (a) (b) *HHU-Dataset* and (c) (d) *CER-Dataset* yielded by using (a) (c) 4 day-types and (b) (d) 8 day-types. The graphs have been colorized depending on which cluster the total energy consumption time series has been assigned to on a given calendar day.

process complies with our previous work [Boc18]. To accomplish this, a *high-pass filter* is applied on the total energy consumption time series, for example by using the *Fourier transformation*, where the lowest-frequency terms from the total energy consumption time series are used to notate the recurring patterns embedded in the dataset. In doing so, those lowest-frequency terms are consolidated as the dynamization function for that dataset. We then divide each Smart Metering measurement of the dataset by the value of said dynamization function to exclude the recurring patterns during analysis; when the load profiles are assessed, the seasonal patterns are reapplied as per equation 2.4 before the forecast is compared to the original, unmodified energy consumption. This

allows the clustering process to focus on the shape of the daily consumption time series as opposed to being mostly influenced by the energy offset. As a consequence of this procedure, each energy provider may extract a unique dynamization function from the data of their customer base. However, since the dynamization function is already part of the normal market communication the interoperability between energy providers is guaranteed where necessary [Bun06].
Though the usage of a *high-pass filter* to derive a dynamization function has the potential to improve the clustering process, the resulting load profiles are more likely to be region-specific since the dynamization function might encode recurring patterns unique to the underlying customer base. As experiments such as [SM17] have shown, region-specific load profiles can differ from the *BDEW Standard Load Profiles* quite significantly, tough more extensive evaluation is required to distinguish whether these differences are caused mainly due to region-specific factors or changes in the customer behavior over time. Complementary to this, [HRR14] mentions that region-specific load profiles have the potential to significantly improve upon non-region-specific load profiles. Because of this, energy providers seeking to minimize *imbalance energy* may want to acknowledge a dynamization function as well as associated load profiles as being region-specific, causing them to use the forecasting models of a different region if they gain a new customer from said foreign region. For the datasets used in our evaluation, no regional information about the customer is present. Thus, we have opted to assume that all customers of a given dataset belong to the same region.
In addition to the adoption of a *dynamization function*, we also modified the original *Partial Distance Strategy* to be based on the manhattan distance rather than the euclidean distance:

$$dist_{PMD}(a,b) = \frac{H}{I} \cdot \sum_{n=1}^{H} |a_n - b_n| \cdot I_n$$
$$\text{with } I_n = \begin{cases} 1 & \text{if neither } a_n \text{ nor } b_n \text{ are missing values} \\ 0 & \text{else} \end{cases} \quad \text{and } I = \sum_{n=1}^{H} I_n \tag{4.27}$$

We refer to this distance measure as the *Partial Manhattan Distance* $dist_{PMD}(a,b)$. Our motivation behind this change is to make the clustering process more sensitive to the *imbalance energy* which we have introduced in section 2.2, as imbalance energy can be modeled as the manhattan distance between the actual consumption and the forecast time series.
Additional, similar to our setup described in section 4.4.1.1, we have opted to use *K-Means++* [AV07] for the day-type segmentation while employing *Random Coordinates* as well as *K-Means++* separately for the identification of typical consumption patterns.

4.4.2.2 Results

The first step in conducting the experiments as outlined above involved applying the *Fourier transformation* on each dataset. The resulting *dynamization functions* are visualized in figure 4.7. For the *HHU-Dataset*, the dynamization function has been trained according to the Smart Metering data for the year 2014 and can be expressed as

$$DynFactor_{HHU}(doy) = 1 + 0{,}23654 \cos\left(\frac{2\pi \cdot doy}{365}\right) + 0{,}05169 \sin\left(\frac{2\pi \cdot doy}{365}\right) \tag{4.28}$$

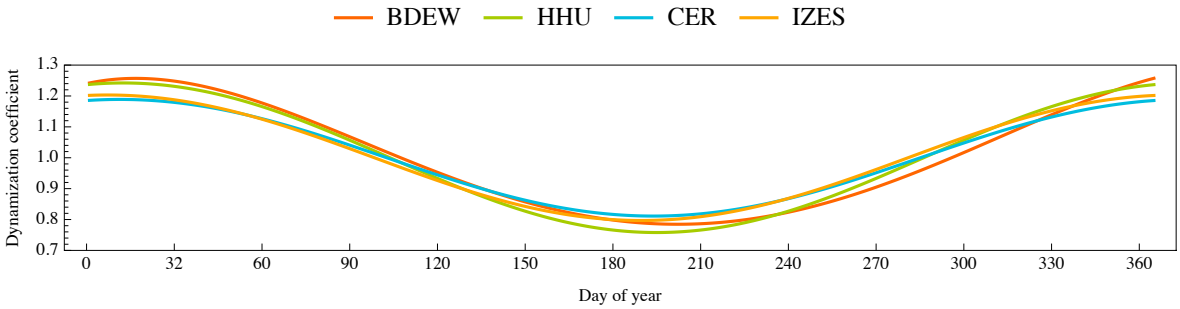


Figure 4.7: Visualization of the dynamization functions discovered by applying the *Fourier transformation* on the *HHU-Dataset* (green graph), the *CER-Dataset* (blue graph) and the *IZES-Dataset* (orange graph). For comparison, the dynamization function for the *BDEW Standard Load Profiles* is depicted in this diagram as the red graph.

while for the *CER-Dataset*, the Fourier transformation has yielded

$$DynFactor_{CER}(doy) = 1 + 0{,}18522 \cos\left(\frac{2\pi \cdot doy}{365}\right) + 0{,}03678 \sin\left(\frac{2\pi \cdot doy}{365}\right) \quad (4.29)$$

as the optimal the optimal representation of the recurring patterns in the dataset for the year 2010. Lastly, for the *IZES-Dataset*, the discovered dynamization function is

$$DynFactor_{IZES}(doy) = 1 + 0{,}2016 \cos\left(\frac{2\pi \cdot doy}{365}\right) + 0{,}02503 \sin\left(\frac{2\pi \cdot doy}{365}\right) \quad (4.30)$$

when using the available data for the year 2010. In the case of a given calendar year being a leap year, the denominator in the sine and cosine argument is set to 366 instead of 365. When compared to the original *BDEW dynamization function*, the overall shape of the dynamization function is very similar as can be seen in figure 4.7, though the maximum and minimum value of the BDEW dynamization function occurs slightly later than for the dynamization functions derived by applying the Fourier transformation. Yet, their shape is sufficiently distinct that one may conjecture regional differences between the datasets. This would back up the argument of [HRR14], where the authors state that region-specific forecast models yield significantly better results than non-region-specific models. Since all dynamization functions are purposefully normalized such that their average value over the course of a year is 1, they do not distort the estimated total energy consumption of a customer as given by the *Year Consumption Forecast* when reapplying the recurring patterns according to equation 2.4. Dividing the individual Smart Metering measurements by the value of the dynamization function yields the desired behavior of a *high-pass filter* as seen in figure 4.8. The overall much more linear appearance helps to resolve the issue we mentioned earlier about the clustering primarily being influenced by seasonal fluctuations rather than deviations in the daily routines of customers.

Our results when using the aforementioned dynamization functions as well as the *Partial Manhattan Distance* $dist_{PMD}(a, b)$ are shown in figure 4.9. For the *HHU-Dataset*, our approach has achieved relative forecast errors in the range from 8 to 33 percent, while for the *CER-Dataset* and *IZES-Dataset*, the results are roughly in the range from 9 to 23 percent and from 9 to 13 percent, respectively. Generally, the load profiles performed better when the number of consumption patterns was small. This

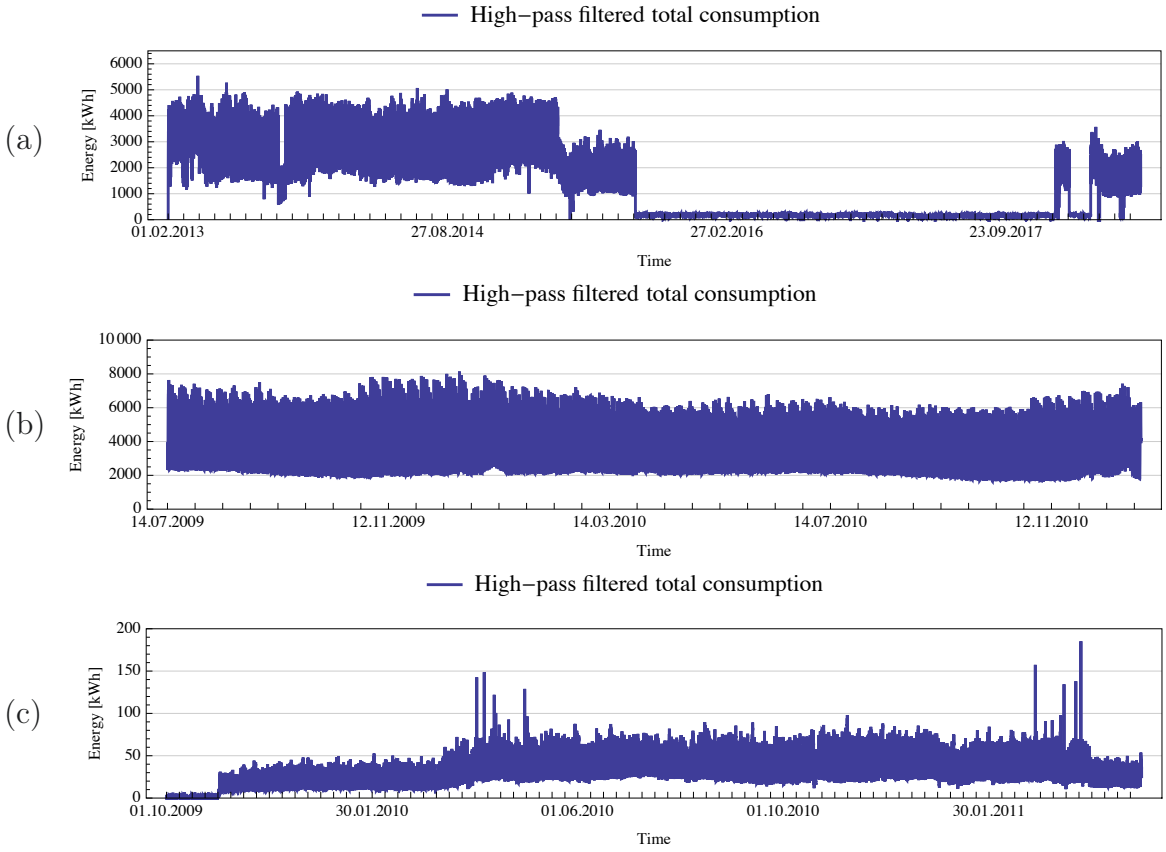


Figure 4.8: Overview of the (a) *HHU-Dataset*, (b) *CER-Dataset* and (c) *IZES-Dataset*, where each measurement has been divided by the corresponding value of $DynFactor_{HHU}$, $DynFactor_{CER}$ and $DynFactor_{IZES}$, respectively.

conclusion is most noticeable for the *CER-Dataset* in conjunction with the *K-Means++* cluster initialization, where a large jump for the values of the relative forecast error can be observed when using more than 14 consumption patterns. One possible reason for the experiments with a smaller number of consumption patterns performing better can be seen when looking at the comparison between the actual consumption and the forecast time series as visualized in figure 4.10 and 4.11. Similar to the experiments presented in section 4.4.1, the load profiles generally recognized the overall shape of the actual consumption successfully. However, as one can see for the *HHU-Dataset* in figure 4.11, when the number of consumption patterns is sufficiently large, the clustering prototypes become susceptible to overlapping clusters in the Smart Metering data. In the case of the *HHU-Dataset*, this has led to the forecast for some day-types to worsen overall, with the deviation graph showing a recurring pattern for those day-types. The same behavior, albeith to a slightly lesser extend, can also be observed for the *CER-Dataset* and the *IZES-Dataset*, where for most calendar days, the actual consumption has been slightly underestimated by the load profiles. At the same time, especially for the week from the 13-th of december until the 17-th of december, the forecast matches the actual consumption of the *CER-Dataset* very good, however with a noticeable spike in overestimating the actual consumption by a large margin during the evening hours.

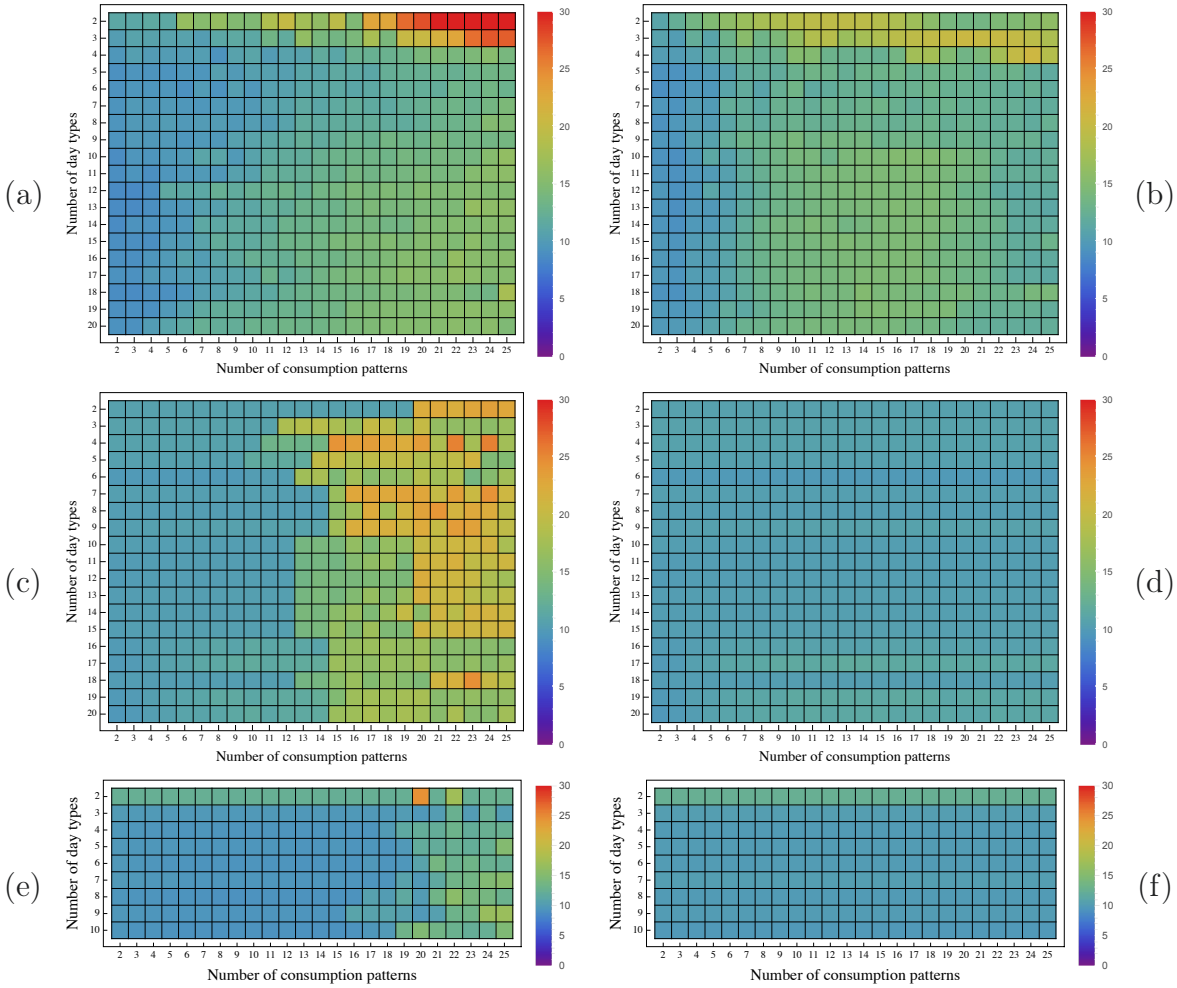


Figure 4.9: Ratio of the deviations and the actual consumption in percent yielded by the load profiles generated using different values for the number of day-types and consumption patterns using the experimental setup described in section 4.4.2.1. The graphs visualize the results for (a)(b) the *HHU-Dataset*, (c)(d) the *CER-Dataset* and (e)(f) the *IZES-Dataset* using (a)(c)(e) *K-Means++* and (b)(d)(f) *Random Coordinates* to generate the starting configuration of the clustering process.

4.4.3 Evaluation using the exponential Manhattan distance

4.4.3.1 Experimental Setup

As mentioned in section 2.2, one of the most important criterions for energy providers when assessing the performance of forecast models is minimizing the amount of imbalance energy caused by them. Notably, unexpected spikes in energy consumption are more undesirable than misvalued uniform consumption due to the financial risk involved. The evaluation of our experiments in the previous sections has shown that while the algorithm successfully identified the overall shape of the consumption, the deviation graphs have shown recurring patterns as well as imbalance energy spikes in the evening hours. One possible explanation for imbalance energy spikes in the evening hours is that the tuples in the dataset match during daytime and the deviations during the evening hours do not warrant splitting the tuples into separate clusters. Because of

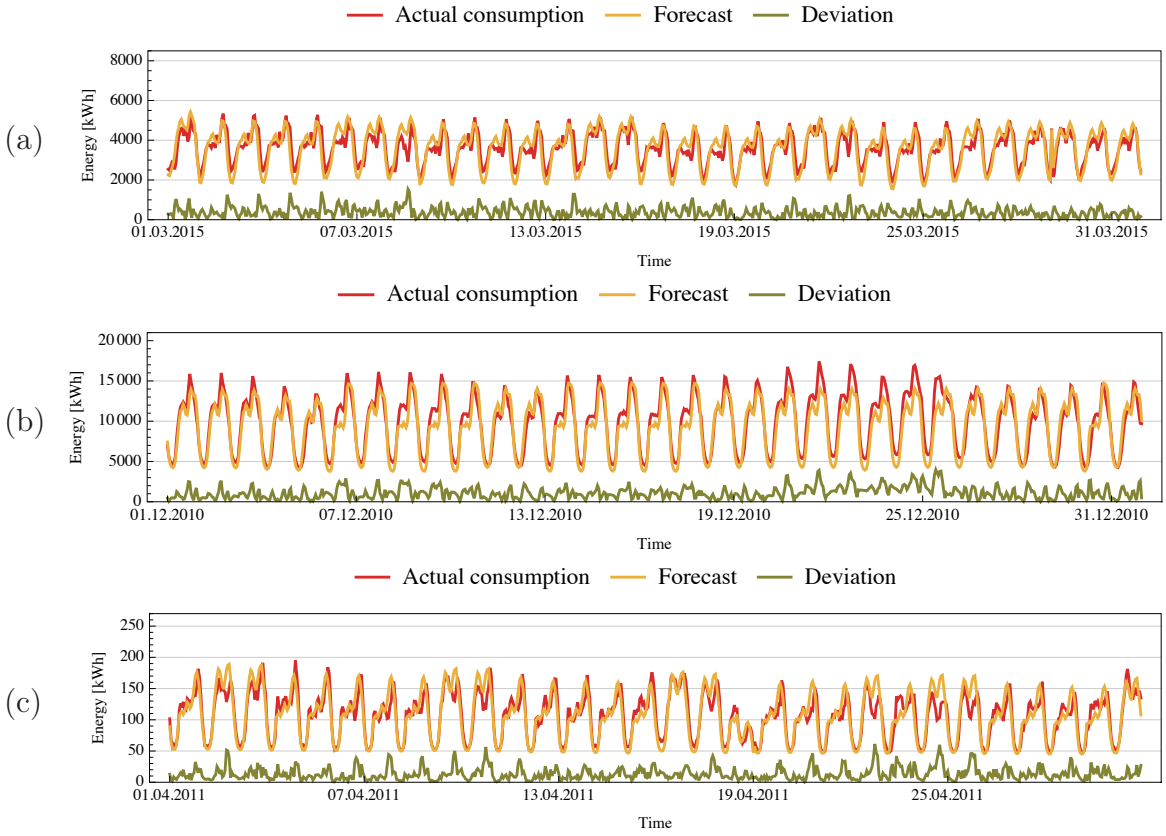


Figure 4.10: Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on 2 day-types and 2 consumption patterns per day-type and using *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.2.1 for (a) the *HHU-Dataset*, (b) the *CER-Dataset* and (c) the *IZES-Dataset*. The green graph shows the absolute deviation of the forecast from the actual consumption.

this, we aim to modify the clustering process so that it rates high deviations in a few dimensions as being worse than low deviations in a large number of deviations in an effort to minimize imbalance energy spikes and to further incentivize the formation of spherical clusters. To accomplish this, we present the *Partial Exponential Manhattan Distance* $dist_{PEMD}(a,b)$ as follows:

$$dist_{PEMD}(a,b) = \frac{H}{I} \cdot \sum_{n=1}^{H} \left(\left(e^{|a_n - b_n|} \right) - 1 \right) \cdot I_n$$
$$\text{with } I_n = \begin{cases} 1 & \text{if neither } a_n \text{ nor } b_n \text{ are missing values} \\ 0 & \text{else} \end{cases} \quad \text{and } I = \sum_{n=1}^{H} I_n \tag{4.31}$$

The idea of using a distance measure based on the exponential function and thus clustering datasets in non-metric feature spaces is not entirely new. For example, [GG89] presented a clustering algorithm known as *Fuzzy-Maximum-Likelihood-Estimation (FMLE)*, which employs the *fuzzy covariance matrix* [DAY69] to account for clusters of different sizes and shapes when assigning a membership degree to each data object. In the case of $dist_{PEMD}(a,b)$, we aim to put more emphasis on the formation of spherical clusters,

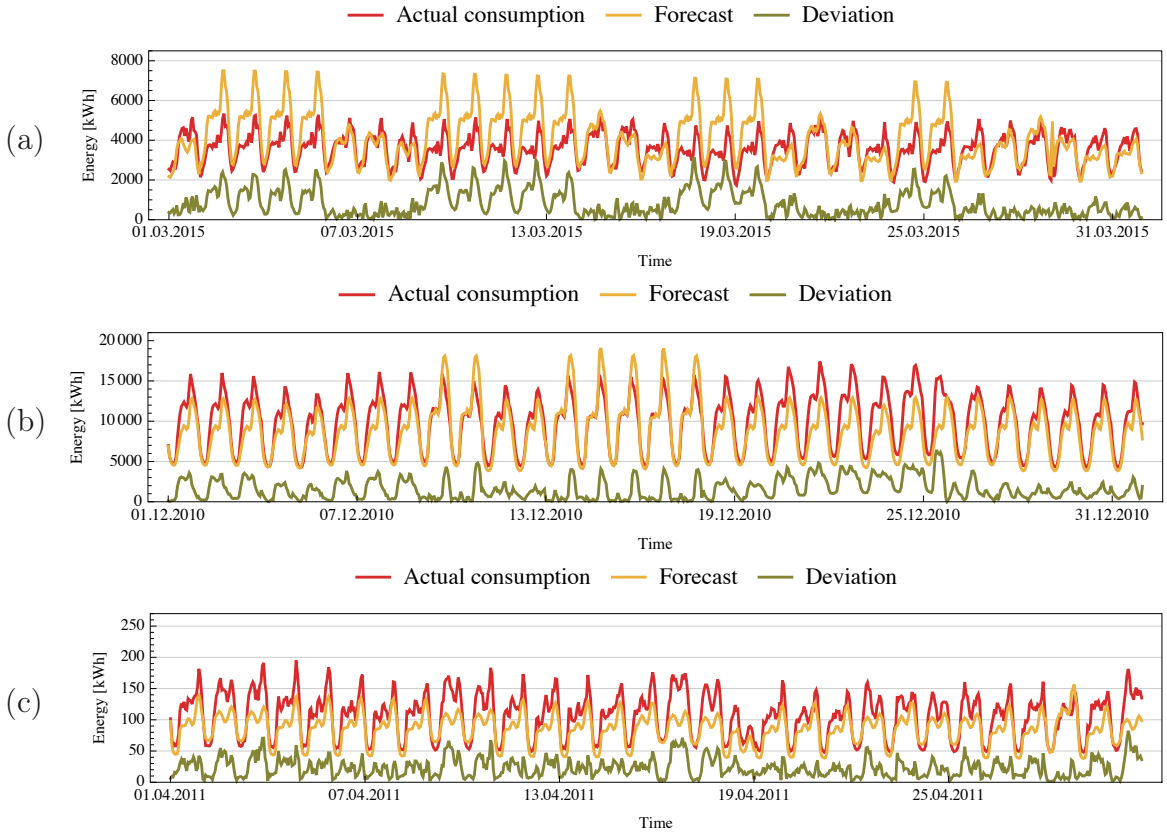


Figure 4.11: Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on 10 day-types and 25 consumption patterns per day-type and using *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.2.1 for (a) the *HHU-Dataset*, (b) the *CER-Dataset* and (c) the *IZES-Dataset*. The green graph shows the absolute deviation of the forecast from the actual consumption.

which is one of the reasons why we have not adopted FMLE directly. Additionally, employing Fuzzy-C-Means instead of FMLE has the advantage of keeping computational requirements low, as having to update and compute the distances with the fuzzy covariance matrix during each iteration of FMLE would make our approach infeasible for most energy providers, while data privacy policies make external services like cloud computing impractical. However, it is noteworthy that having the clustering process to operate on $dist_{PEMD}(a, b)$ no longer defines a metric feature space. This is because while the properties of *symmetry*

$$dist(a, b) = dist(b, a) \tag{4.32}$$

and *positive definiteness*

$$dist(a, b) \geq 0 \qquad \text{and} \qquad dist(a, b) = 0 \Leftrightarrow a = b \tag{4.33}$$

are satisfied, the *triangle inequality*

$$dist(a, b) \leq dist(a, c) + dist(c, b) \tag{4.34}$$

is violated. As a one-dimensional counterexample for this, consider the case of $a = 1$, $b = 7$ and $c = 4$:

$$dist_{PEMD}(a, b) = e^6 - 1 \not\leq 2 \cdot e^3 - 2 = dist_{PEMD}(a, c) + dist_{PEMD}(c, b) \quad (4.35)$$

One may suspect that, due to both methods using a distance measure based on the exponential function, a lot of properties that are applicable to FMLE also apply to a clustering process based on $dist_{PEMD}(a, b)$. For example, [GG89] stresses the importance of a good starting configuration for FMLE as the algorithm quickly converges to a local optimum in a narrow region.
In addition to the *Partial Exponential Manhattan Distance* $dist_{PEMD}(a, b)$, we have employed the same dynamization functions as in section 4.4.2 as well as both *K-Means++* [AV07] and *Random Coordinates* to generate the starting configuration of the clustering process when identifying typical consumption pattern, while using *K-Means++* during the day-type segmentation.

4.4.3.2 Results

The results of our experiments when using the *Partial Exponential Manhattan Distance* $dist_{PEMD}(a, b)$ are visualized in figure 4.12.

One apparent property when comparing these results with the ones presented in figure 4.9 is that the overall performance of the load profiles is roughly the same: for the *HHU-Dataset*, the approach based on the *Partial Exponential Manhattan Distance* $dist_{PEMD}(a, b)$ has achieved relative forecast errors from 8 to 33 percent. Similarly, for the *CER-Dataset* the results are roughly in the range from 9 to 16 percent, while for the *IZES-Dataset*, the results for the relative forecast errors range from 9 to 13 percent. Though the results have improved only marginally for the *HHU-Dataset* and the *IZES-Dataset*, the results for the *CER-Dataset* have improved noticeable when using a large value for the number of consumption patterns. As such, the stronger incentive of the clustering process to build spherical clusters appears to be a plausible explanation for the increase in performance. Aside from that, the results fall in line with the load profiles from the experiments presented in section 4.4.2 as can be seen when comparing figure 4.13 and 4.14 with the results from figure 4.10 and 4.11. However, since the load profiles presented in this section have not eliminated some of the problems with the *Partial Manhattan Distance* $dist_{PMD}(a, b)$ such as the peaks in the relative forecast error during the evening hours as visualized in figure 4.11, it can be assumed that a stronger emphasis on the formation of spherical clusters is not sufficient to improve the overall quality of the resulting load profiles.

4.4.4 Evaluation using weighted clustering

4.4.4.1 Experimental Setup

Some of the areas where the load profiles presented in the previous sections can be improved the most are the peculiar offsets, where the overall shape of the consumption has been predicted accurately, but multiplicative scalars between the actual consumption and the forecast either during whole calendar days or during short time periods have caused a large amount of imbalance energy. These offsets for whole calendar days can be seen in figure 4.14 (a), for example. Since these offsets only occur during

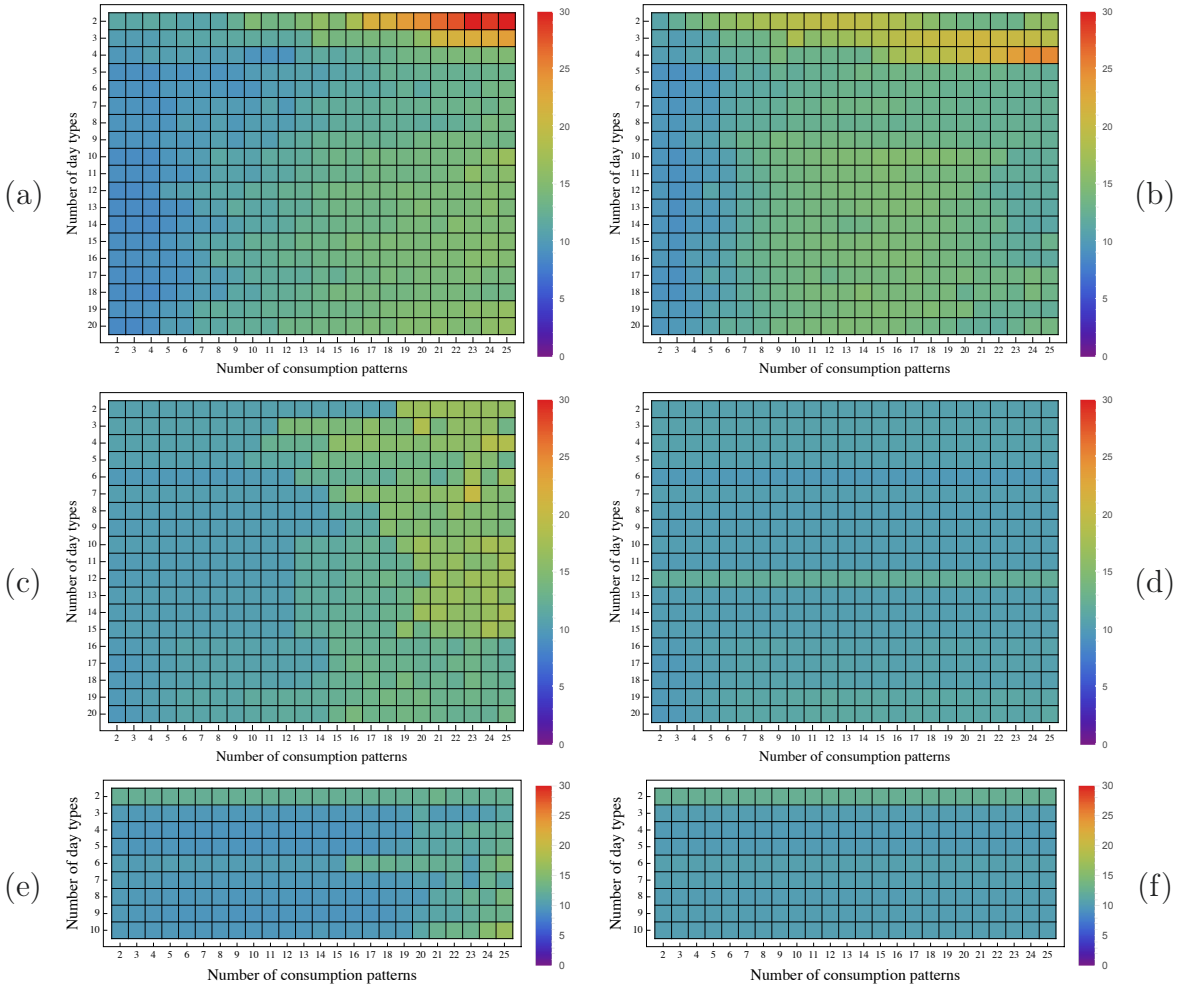


Figure 4.12: Ratio of the deviations and the actual consumption in percent yielded by the load profiles generated using different values for the number of day-types and consumption patterns using the experimental setup described in section 4.4.3.1. The graphs visualize the results for (a)(b) the *HHU-Dataset*, (c)(d) the *CER-Dataset* and (e)(f) the *IZES-Dataset* using (a)(c)(e) *K-Means++* and (b)(d)(f) *Random Coordinates* to generate the starting configuration of the clustering process.

certain day-types, some plausible explanations are overlapping clusters and clusters of unequal size in the *training data*. Another cause for this, which we will investigate over the course of this section, are unevenly distributed values for the *Year Consumption Forecast (YCF)*. To explain our motivation for this, consider the following thought experiment which is illustrated in figure 4.15: let there be the feature-vectors of two distinct customers in a one-dimensional feature space. If both customers are assigned to the same cluster, the clustering prototype vector $\vec{v}$ will most likely be positioned in the middle between the feature-vector of customer A and customer B. In that case, the resulting forecast will be computed as

$$Forecast = (YCF_{customer\ A} + YCF_{customer\ B}) \cdot \vec{v} \tag{4.36}$$

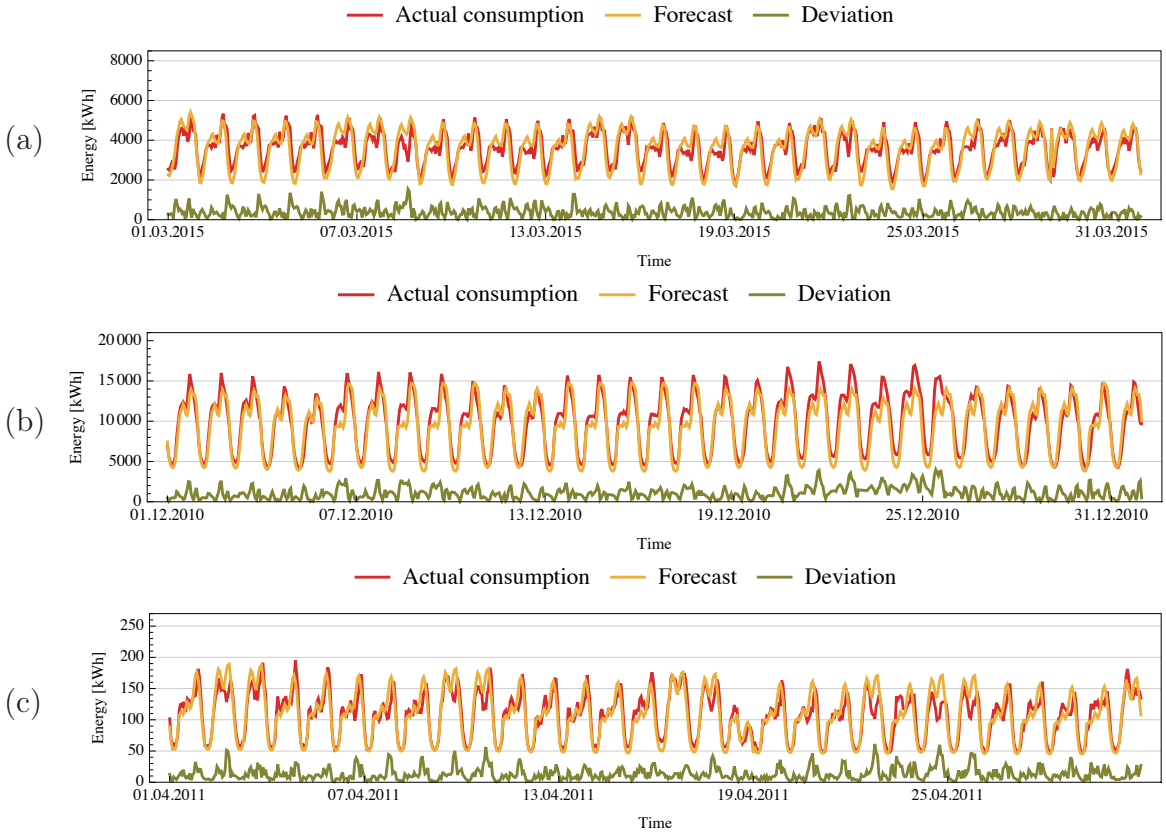


Figure 4.13: Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on 2 day-types and 2 consumption patterns per day-type and using *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.3.1 for (a) the *HHU-Dataset*, (b) the *CER-Dataset* and (c) the *IZES-Dataset*. The green graph shows the absolute deviation of the forecast from the actual consumption.

If both customers have approximately the same YCF, this forecast most likely matches the actual consumption which is given as

$$\begin{aligned} Actual\ Consumption = & YCF_{customer\ A} \cdot Featurevector_{customer\ A} \\ & + YCF_{customer\ B} \cdot Featurevector_{customer\ B} \end{aligned} \tag{4.37}$$

If, however, the YCF of the customers are not radially symmetric with the clustering prototype as the center, equation 4.36 as yielded by the load profiles is no longer a good approximation for equation 4.37 since even though customer B in figure 4.15 has a much higher YCF than customer A, the feature-vectors of both customers are weighted the same. In the illustrated scenario however, it makes much more sense to weight the feature-vector of customer B stronger when computing the position of the clustering prototype $\vec{v}$, yielding a clustering prototype $\vec{v'}$ and thus forecast that is more sensitive to the large consumption customers. As such, we propose to weight the feature-vector so that the position of the clustering prototype is computed according to the product of the membership degree and the YCF of the corresponding customer, where a higher YCF corresponds to having a higher influence of the positioning of the clustering prototype, rather than only the membership degree. Specifically, we

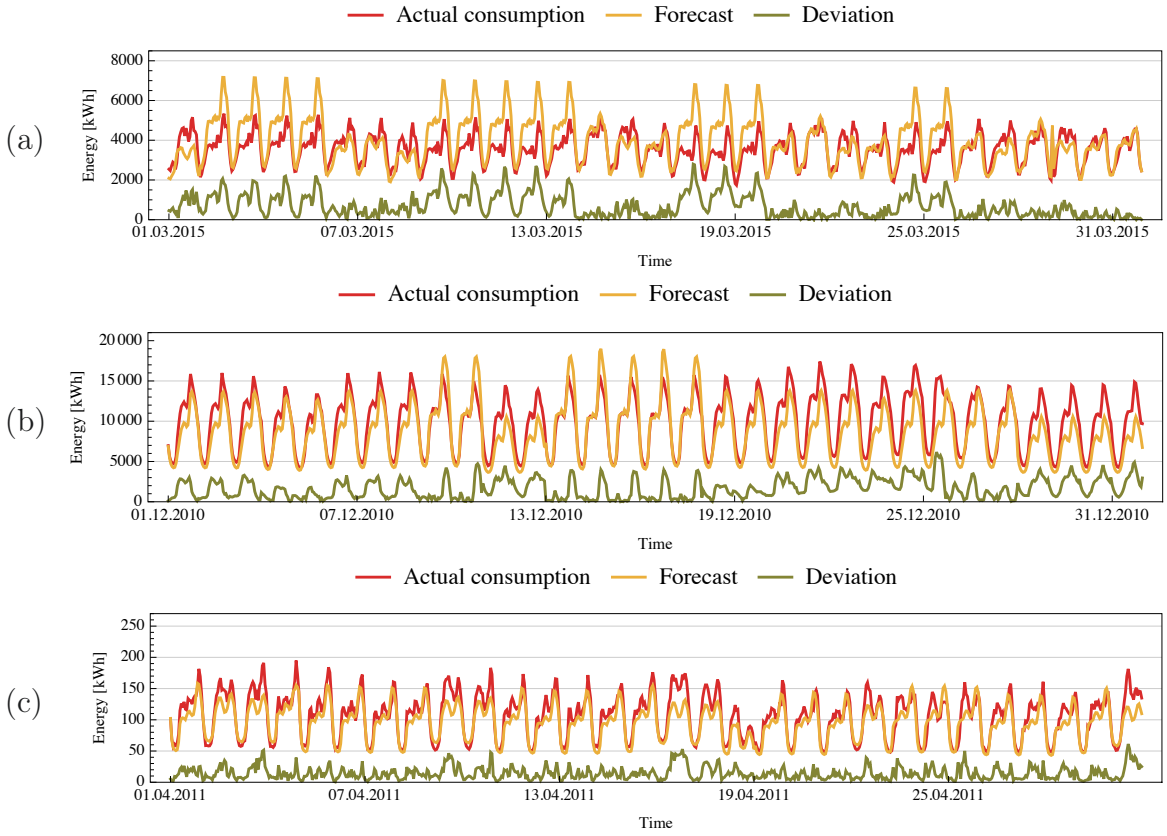


Figure 4.14: Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on 10 day-types and 25 consumption patterns per day-type and using *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.3.1 for (a) the *HHU-Dataset*, (b) the *CER-Dataset* and (c) the *IZES-Dataset*. The green graph shows the absolute deviation of the forecast from the actual consumption.

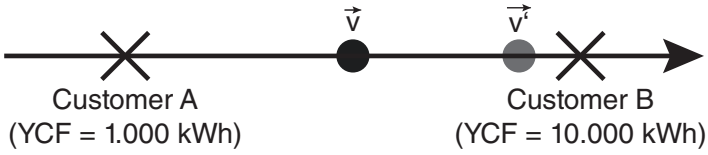


Figure 4.15: Visualization of two customers and the clustering prototype that the clustering process would yield if the position of the clustering prototype was based solely on the distance in the feature space (black dot) and weighted according to the product of *Year Consumption Forecast* of the data tuple and the distance in the feature space (gray dot).

have opted to accomplish this by adopting some ideas from *Weighted Fuzzy-C-Means (WFCM)* [HHG07a], specifically by modifying the computation of the clustering prototypes $v_o^{(r+1)}$ in line 5 of algorithm 1 as follows:

$$v_o^{(r+1)} := \frac{\sum_{i=1}^{N} w_i \cdot \left(u_{o,i}^{(r+1)}\right)^m \cdot x_i}{\sum_{i=1}^{N} w_i \cdot \left(u_{o,i}^{(r+1)}\right)^m} \tag{4.38}$$

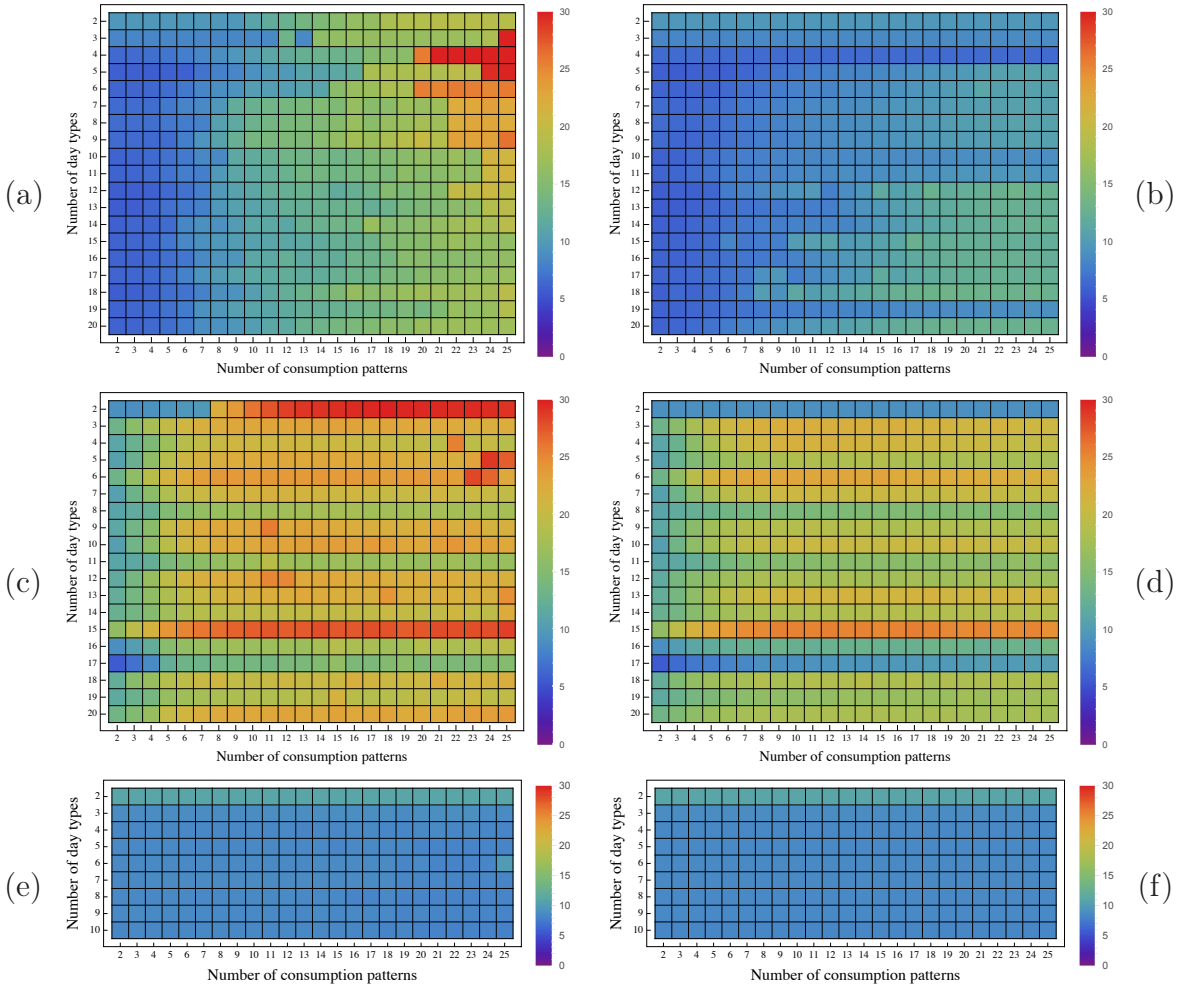


Figure 4.16: Ratio of the deviations and the actual consumption in percent yielded by the load profiles generated using different values for the number of day-types and consumption patterns using the experimental setup described in section 4.4.4.1. The graphs visualize the results for (a)(b) the *HHU-Dataset*, (c)(d) the *CER-Dataset* and (e)(f) the *IZES-Dataset* using (a)(c)(e) *K-Means++* and (b)(d)(f) *Random Coordinates* to generate the starting configuration of the clustering process.

In our case, w_i is equal to the YCF of the corresponding customer.
Aside from the adoption of Weighted Fuzzy-C-Means and similar to section 4.4.2, we have used the *Partial Manhattan Distance* $dist_{PMD}(a, b)$ including the corresponding dynamization functions as well as *K-Means++* [AV07] for the day-type segmentation while utilizing *Random Coordinates* and *K-Means++* separately for the identification of typical consumption patterns.

4.4.4.2 Results

The results of our experiments based on WFCM as described in section 4.4.4.1 are shown in figure 4.16.
For the *HHU-Dataset*, this approach has achieved a major progression in the quality of the load profiles with relative forecast errors below 6 percent for both *K-Means++* and *RandomCoordinates* when using a small number of consumption patterns. This con-

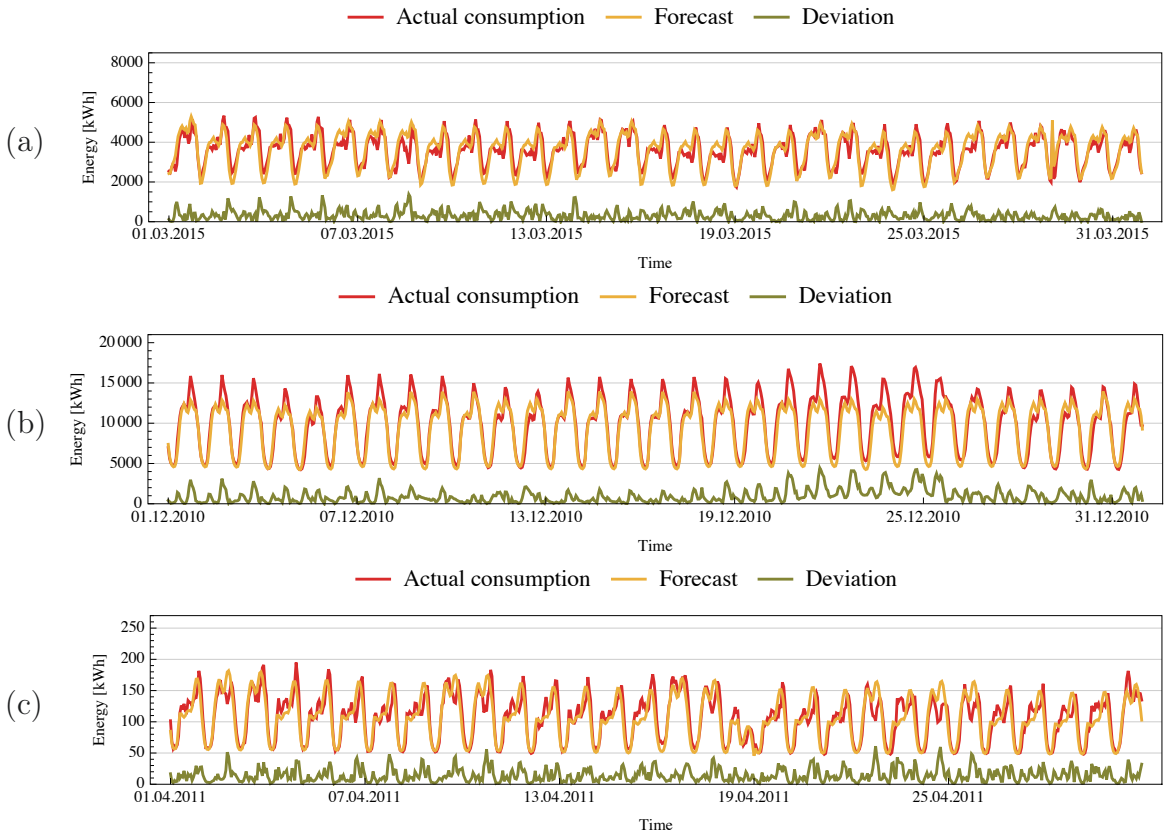


Figure 4.17: Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on 2 day-types and 2 consumption patterns per day-type and using *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.4.1 for (a) the *HHU-Dataset*, (b) the *CER-Dataset* and (c) the *IZES-Dataset*. The green graph shows the absolute deviation of the forecast from the actual consumption.

stitutes a major improvement compared to the *BDEW Standard Load Profiles*, where the expected relative forecast error is between 12 and 14 percent [SM17] as shown in figure 2.5. However, similar to the experiments presented in section 4.4.2 and 4.4.3, the accuracy of the load profiles worsens as the number of consumption patterns used increases.

For the *CER-Dataset*, the WFCM approach has shown to be a significant performance regression for any number of day-types when not using a very small number of consumption patterns. One noteworthy exception to this behavior has occurred when using 17 day-types; in this case, the performance of the load profile for both *K-Means++* and *RandomCoordinates* is comparable to the accuracy of the load profiles for the *HHU-Dataset*.

The performance of the load profiles for the *IZES-Dataset* has also improved compared to the experiments from the previous sections. Here, the relative forecast errors are roughly in the range from 8 to 9 percent. In contrast to the *HHU-Dataset* and the *CER-Dataset*, neither the number of day-types nor the number of consumption patterns seem to have a significant influence on the performance of the resulting load profiles.

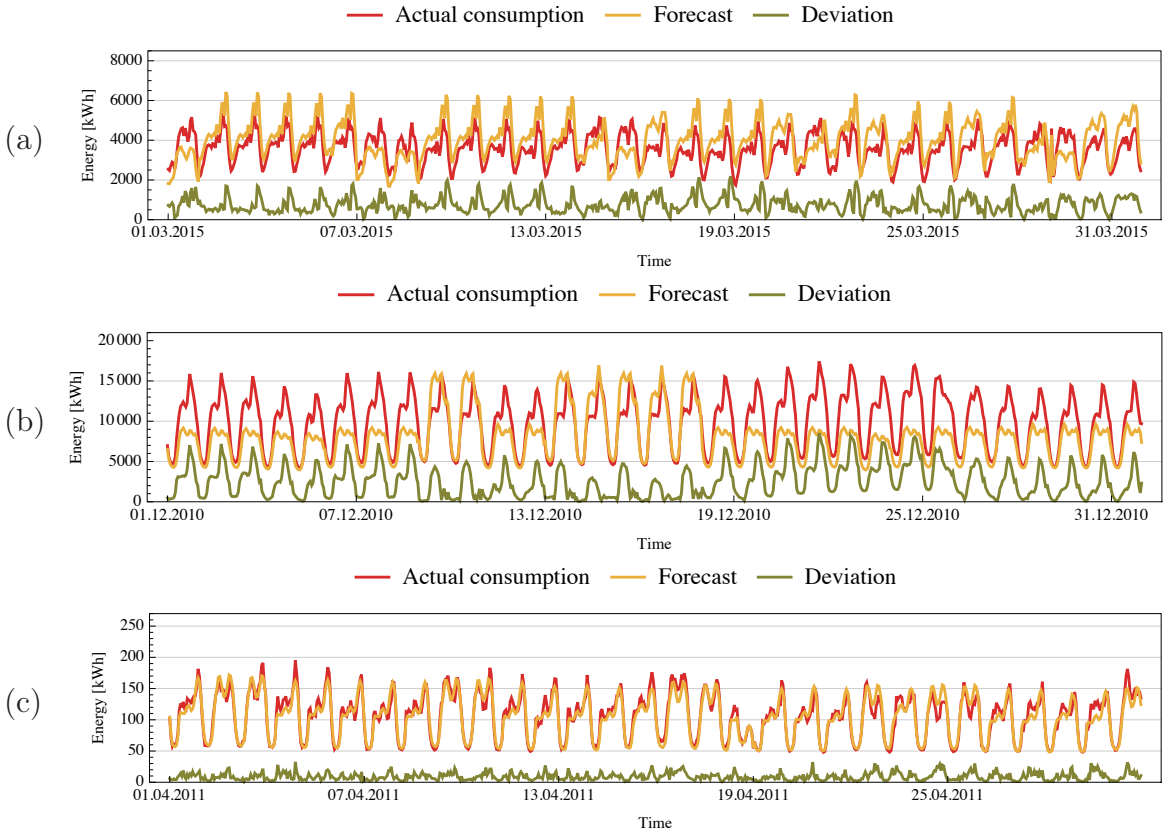


Figure 4.18: Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on 10 day-types and 25 consumption patterns per day-type and using *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.4.1 for (a) the *HHU-Dataset*, (b) the *CER-Dataset* and (c) the *IZES-Dataset*. The green graph shows the absolute deviation of the forecast from the actual consumption.

Overall, most of the deductions made during our experiments presented in the previous sections also apply to the approach based on WFCM. In particular, figure 4.17 shows that for a small number of consumption patterns, the resulting load profiles are well-suited to accurately forecast the total energy consumption. As the number of consumption patterns increase however, the deviations between the actual total consumption and the forecast become larger as shown by figure 4.18.
The results for the load profiles yielding very high relative forecast errors for the *HHU-Dataset* and the *CER-Dataset* are visualized in figure 4.19. Here, the graphs illustrate a considerable, yet systematic deviation between the actual total energy consumption and the forecast. As is has been the case for the experiments presented in the previous sections, these deviations occur only for certain day-types. This indicates that the dataset has certain properties that significantly hampers the clustering performance on these day-types. However, since these results show that not all day-types are affected by this problem and that the actual total energy consumption and the forecast match closely on some day-types, it can be assumed that the quality of the clustering segmentation worsening as the number of consumption patterns increases is no systematic weakness of our framework described in section 4.3. Instead, these observations could

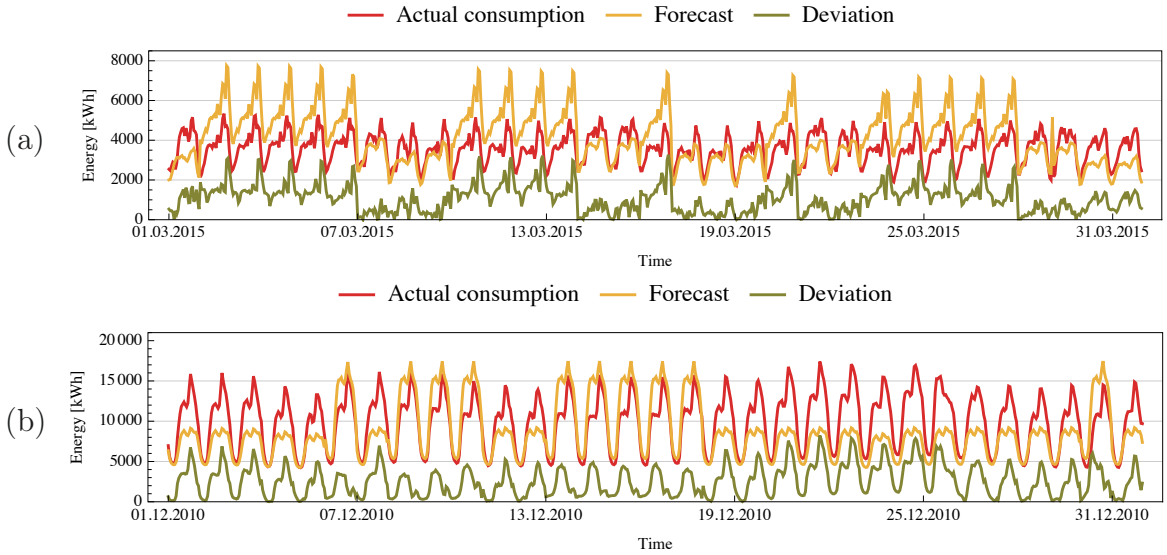


Figure 4.19: Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.4.1 for (a) the *HHU-Dataset* when using 4 day-types and 25 consumption patterns per day-type and (b) the *CER-Dataset* when using 2 day-types and 25 consumption patterns per day-type. The green graph shows the absolute deviation of the forecast from the actual consumption.

be attributed to currently unknown properties of the dataset for these problematic day-types. Possible candidates for the cause of these observations are overlapping clusters or clusters of unequal size.

5

Online Clustering using Smart Meter Data

One important property of load profiles is the ability to predict the energy consumption of customers many months in advance. If the consumption behavior changes over time, maintaining acceptable forecast results is a non-trivial task. In this chapter, we address the problem of keeping load profiles up-to-date in the event of the consumption behavior of customers not being constant.

5.1 Motivation

When utilizing Smart Metering data to build load profiles as the foundation to plan the future buy-in of energy, one important aspect for energy providers is that, in a real-world scenario, there is a constant stream of new Smart Metering data. This new Smart Metering data possibly hints at a change in customer behavior or allows to fine-tune existing forecast models. Unless load profiles are kept up-to-date, their performance can worsen over time as it has been the case for the gas economy [Roo+14]. Additionally, experimental evaluations such as in [SM17] have shown that region-specific load profiles may significantly differ from the existing *BDEW Standard Load Profiles* even though the same methodlogy has been used to build them. This can indicate that either the energy consumption is influenced by region-specific factors or that the consumption behavior of customers have changed over time. Since the technological advancements of society are not expected to slow down anytime soon, it can be assumed that the customer behavior will also change over time, thus creating the necessity to keep the forecast models up-to-date to enable energy providers to fulfill their duties outlined in section 2.2. However, frameworks such as the one presented in chapter 4 usually require to rebuild the forecast models from scratch, which can be a time consuming process depending on the computational equipment available to energy providers. Because of this, energy providers would need to strike a balance between their interest in modernizing the load profiles and waiting until enough new Smart Metering data is

available so that rebuilding the forecast models is worth the effort. Thus, there is a profound interest to process old and new data once in a way that allows to incrementally integrate new data as it made available into existing results with minimal effort. For this purpose, a large variety of *Online Clustering* algorithms exist and have been discussed in academic literature. The prospect of Online Clustering is also especially interesting for energy providers with a very large customer base as the performance of most traditional clustering algorithms degrades drastically if the dataset does not fit into memory completely. This is because *very large* datasets can artificially be considered as *Streaming Data* and processed accordingly [Hav+12].
However, simply using an Online Clustering algorithm instead of a traditional clustering method does not enable the framework presented in section 4.3 to update existing load profiles. This is because building load profiles according to section 4.3 is a multistep process where the optimal day-type segmentation is used as the input for the identification of typical consumption patterns; changing the day-type segmentation as part of a naive Online Clustering approach would also change the input datasets P_n according to equation 4.23, invalidation results that have been achieved with the old definition of P_n. Over the course of this chapter, we aim to address this problem by first giving a short introduction for common Online Clustering algorithms. Subsequently, we will outline an approach that utilizes *Online Data Mining* techniques to generate updated load profiles by performing computations on new data and reusing results from old data. We will then proceed by presenting an experimental evaluation based on our approach.

5.2 Related Work

Overall, Online Clustering algorithms can be categorized into two groups [MLB]:

- **General Clustering:** Algorithms belonging to this group do not assume any ordering on the stream of data. This category includes the algorithms presented in section 1.2.2.

- **Clustering algorithms for time series:** Clustering methods belonging to this category take advantage of the *ordering* that naturally comes with time series data by assuming that data objects that are close in time are highly related. Thus, they do not require the number of clusters to be predefined by the analyst. Instead, they use *change-point detection techniques* to identify the number of clusters, allowing them to be more tailored towards time series, losing some generality in the process. Some representatives for this category are *EROLSC* [IDK18], *eClass* [AZ08] and *OEC* [MLB].

One important aspect when discussing Online Clustering algorithms is the concept of *Evolving Distributions*. This term describes the notion of the *true* cluster centers of a given dataset as being non-static, meaning that the position and even the number of *true* cluster centers is time-dependent. As [Agg+03] points out, Online Clustering algorithms that view the dataset as an one-pass stream of data, where each data object can only be read once in chronological order, are oblivious to such Evolving Distributions. This causes the clustering results of one-pass algorithms to be dominated by *outdated history*. Because of this, if taking the possibility of Evolving Distributions into account

when clustering Smart Metering data, for example caused by current or upcoming technological advancements such as electric vehicles, one-pass clustering algorithms are not the best choice. In addition, many one-pass clustering algorithms such as *Single-Pass Fuzzy-C-Means (spFCM)* [HHG07b] or *LOCALSEARCH* [OCa+02] require the data to be processed in a random order, notably meaning the tuples are scrambled non-chronologically prior to the analysis. Other algorithms, such as *random sampling plus extension Fuzzy-C-Means (rseFCM)* [Hav+12], process only a random subset of the data and generalize the findings onto the complete dataset. Because of this, the applicable use-case of one-pass clustering algorithms is to scale clustering algorithms to *very large* datasets, rather than clustering streaming data [Agg+03].
In the case of analysing Smart Metering time series, processing the data chronologically is highly desirable as this represents a realistic use-case of how an energy provider will want to handle the data in an effort to adapt to new trends that have a significant impact on the consumption behavior of customers in a timely manner. To counteract the drawbacks of one-pass clustering algorithms, [Agg+03] introduces the concept of *micro-cluster* as part of the *CluStream* framework to aggregate data chunks partitioned by time and compare them within a user-specified time window. The basic idea of this approach has also been used in [Cao+] as part of the *DenStream* algorithm, a method that uses the density-based *DBSCAN* [Est+96], allowing the approach to identify clusters of arbitrary shapes.

Another approach for clustering streaming data is presented in [HHG07a]. Here, the authors assume that the data arrives in chunks, meaning that n_j data objects get recorded at the time t_j. In particular, even though data objects may be recorded at an arbitrary point in time in a real-world scenario, all data objects recorded at the time $t \in (t_{j-1}, t_j]$ are assigned to the j-th chunk of data, treating them as if all data objects simultaneously arrived at time t_j. The basic idea is to then apply clustering on the first chunk of data and compute the clustering weights for each clustering prototype according to

$$w_o = \sum_{i=1}^{n_1} u_{o,i} \qquad , 1 \leq o \leq c \tag{5.1}$$

Afterwards, the j-th chunk of data is processed by using a dataset comprised of the clustering prototypes which have resulted from the clustering process of the $(j-1)$-th chunk of data in addition to the n_j data objects of the j-th chunk of data. Here, *Weighted Fuzzy-C-Means (WFCM)* is used to cluster the data, which is equal to algorithm 1 except for line 5 which is replaced by equation 4.38. While the *weight* of the clustering prototypes from processing the previous chunk of data, which are now treated as normal data objects, is given in equation 5.1, the actual data objects get assigned a weight of 1. In doing so, WFCM optimizes a different objective function than the original Fuzzy-C-Means, which is given as follows:

$$J(\cdot) = \sum_{o=1}^{c} \sum_{i=1}^{N} w_i \cdot u_{o,i}^m \cdot \|x_i - v_o\|^2 \tag{5.2}$$

Since the cluster weights are computed according to the membership degrees, their weights do not accumulate indefinitely over several chunks of data, meaning that WFCM will not become insensitive to Evolving Distributions over time. As [HHG07a] emphasizes, this is one of the main differences compared to *LOCALSEARCH* [OCa+02].

Though the above description of the algorithms specifies to use only the clustering prototypes of the previous chunk of data as *history*, it is also possible to use the clustering prototypes of the previous k chunks of data as history, by which the analyst indirectly specifies the sensitivity of the algorithm towards Evolving Distributions. Depending on the application, the analyst may need to strike a balance between utilizing *more history*, which translates to a more precise clustering result but worse performance on rapidly Evolving Distributions, and *less history*, which means the clustering result will become less precise but more accurate when dealing with Evolving Distributions.
An approach similar to [HHG07a] is presented in [Hor+08] called *Online Fuzzy-C-Means (OFCM)*. Here, the authors also assume that the data arrives in chunks. However, unlike [HHG07a], the first step of this approach consists of all chunk being clustered independently of one another using the original Fuzzy-C-Means as presented in algorithm 1, meaning that no *history* data is used. Once clustering has completed, the weights of the clustering prototypes are computed according to equation 5.1. Afterwards, the clustering prototypes that have been yielded from clustering all chunks are merged to form a new dataset on which WFCM is applied. Some of the main advantages of using OFCM are the huge potential for parallelism since all chunks are clustered independently and that the analyst does not need to commit him- or herself to using a certain amount of history beforehand. Instead, the amount of history only needs to be fixed once all clustering prototypes are merged and clustered using WFCM; since the number of clusters is usually much smaller than the number of data objects, the step involving WFCM is typically much faster to compute than the amount of computational power required to cluster all chunks of data individually. Because of this, this approach allows to quickly test out different amounts of *history*, which may be dataset-specific, yet have a significant impact on the quality of the final clustering segmentation [HHG07a; Hor+08].

5.3 Building load profiles using Online Clustering

In order to be able to build load profiles that adhere to the specification outlined in section 2.3 using Online Data Mining techniques, one important requirement is to process the data chronologically. In particular, clustering methods that require *scrambling* of the data, such as *rseFCM* [Hav+12] or *spFCM* [HHG07b], can not be used, as their primary use-case is to process *very large* datasets that are static rather than *streaming* data. The chosen method should also be sensitive to Evolving Distributions as reacting to changes in consumer behavior is the main motivation for adopting Online Data Mining. One candidate that fits all these criteria is *OFCM* [Hor+08] which we have presented in section 5.2. Over the course of this section, we will present slight modifications to OFCM to adapt it to the special use-case of building load profiles.

One of the main difficulties of using OFCM for building load profiles is that our approach as presented in section 4.3 is a multistep process, where input datasets for the identification of typical consumption patterns are dependent on the final segmentation yielding the optimal day-types. While Online Clustering for the day-type segmentation is desirable, a naive adaptation of OFCM would imply that all results regarding typical consumption patterns are obsolete once the day-type segmentation changes as the underlying datasets for the typical consumption patterns change as well. Because of

this, the naive approach would be to keep the day-type segmentation static and use Streaming Clustering algorithms only to fine-tune the expected consumption behavior of customers towards emerging trends by processing newly arriving Smart Metering data. In this scenario, the *chunks* of the dataset as described by the OFCM approach [Hor+08] would be predefined by gathering all data consecutively belonging to a given day-type, similar to P_n as given in equation 4.23. As the final step of OFCM, for each day-type the weighted clustering prototypes of all chunks within a user-specified time horizon are clustered using WFCM to yield the clustering prototypes to build the load profiles with. By varying the time horizon used for WFCM, the algorithm can be tuned by an analyst to include the optimal amount of history.
Though the naive approach as described above has the disadvantage of not being able to alter the day-type segmentation afterwards, exactly this property can be desirable for energy providers in some scenarios. This is because, depending on the financial resources available to said energy provider, a steady change in the clustering prototypes of the day-type segmentation means perpetual effort for the analyst to derive new rules to classify future calendar days. Especially for very small energy providers this is typically not affordable personnel-wise in the long-term.

In order to be able to both have the day-type segmentation as well as the characteristics of the consumption patterns change over time while using OFCM, we propose the following approach:

1. As the first step, we propose to split the entire dataset into partitions, where each partition contains the consumption data for a given calendar day. This means that after this step, we have an amount of dataset partitions equal to the number of distinct calendar days for which Smart Metering data is available. Similar to the original OFCM algorithm, each of these chunks is then independently clustered using the original Fuzzy-C-Means algorithm. In particular, these clustering processes take place without regard to any day-type segmentation. When new Smart Metering data becomes available, the new data also gets split into partitions and is clustered independently using the original Fuzzy-C-Means algorithm while keeping the already clustered partitions unchanged.

2. Since the day-type segmentation utilizes only the aggregated total energy consumption for all calendar days as input, its computational requirements are negligible and can thus be performed unaltered according to section 4.3.1. This is done as the second step of our proposed approach. In particular, this means that each time new Smart Metering data is made available and the load profiles are to be reevaluated, the process of finding an optimal day-type segmentation is started from scratch. To prevent *outdated history* from having an influence on the day-type segmentation, it is reasonable to only use data from the most recent years as the input dataset.

3. For each day-type identified during the second step, all weighted clustering prototypes of the independently clustered dataset partitions during the first step belonging to the corresponding calendar days are consolidated and clustered using WFCM; this corresponds to the final step of the OFCM algorithm. Similar to the day-type segmentation, only data from recent years should be used to prevent the end result from being influenced by *outdated history*. The exact amount of

history to use is for the analyst to decide. In addition, since WFCM is applied only on clustering prototypes instead of complete datasets, the computational requirements are insignificant compared to section 4.3.2 and can thus be performed from scratch each time the load profiles are to be updated.

4. Once both the day-type segmentation and the consumption patterns are known, the load profiles are compiled and assigned to customers according to section 4.3.3.

One important aspect to think about is how the algorithm as described above will deal with new customers joining or existing customers leaving the energy provider. In the case of an existing customer leaving the energy provider for another, the forecast generated for that customer by the load provider is no longer used, adequately reducing the total energy forecast in the process. Over time, the influence of the lost customer on the characteristics of the remaining load profiles will diminish since they are purposefully designed to phase out *outdated history*. If, on the other hand, a new customer joins the energy provider, the new customer can not immediately get an optimal load profile assigned since the necessary Smart Metering data to make that decision is not available. In that case, one possible approach would be to use the legacy *BDEW Standard Load Profiles* for the new customer for the time being until at least one year of Smart Metering data has been accumulated. During that time, the algorithm as described above is executed normally, meaning that the consumption patterns and the day-type segmentation become progressively influenced by the consumption behavior of the new customer, but the load profile assignment for the new customer is ignored. Once enough representative data has been gathered for the new customer, the new load profiles for the new customer is activated. From that point onward, the customer is then treated as an existing customer.

5.4 Evaluation

In this section we will present the evaluation of the approach outlined in section 5.3. We will start by introducing the dataset used for this evaluation, which is a subset of the *HHU-Dataset* introduced in section 4.1, as well as by describing the segmentation of the dataset into training data and test data. We then proceed by explaining how we plan to simulate the passing of time, that is, making *new* Smart Metering data available, and how the segmentation of the dataset into training data and test data is changed as a result. Afterwards, we show and discuss the results of our experimental setup.

5.4.1 Experimental Setup

To evaluate the performance of the approach as presented in section 5.3, we have taken a subset of the dataset we have introduced as the *HHU-Dataset* in section 4.1. This subset, which is visualized in figure 5.1, consists of a total of 383 distinct customers and, like the original *HHU-Dataset*, covers a period of 65 months with a resolution of 1 measurement per hour. We have specifically decided for this dataset because the time periods that the *CER-Dataset* and the *IZES-Dataset* cover are too small to be used for evaluating this approach. At the same time, the subset of the *HHU-Dataset* we have chosen contains customers for which data is continuously available most of the time.

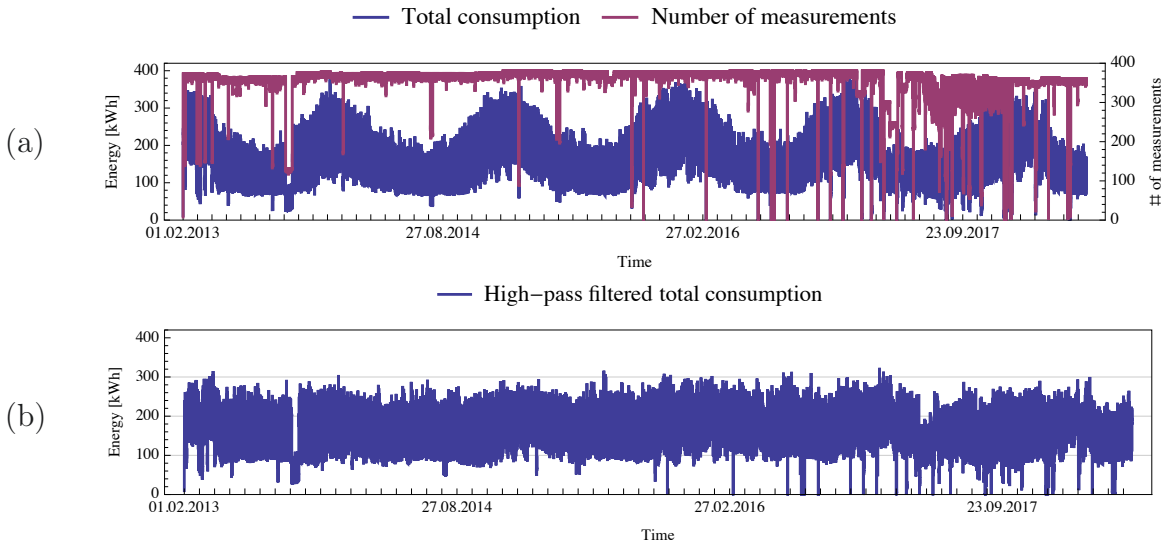


Figure 5.1: Overview of a subset of the *HHU-Dataset* used for the experimental evaluation of the approach described in section 5.3 (a) *as-is* and (b) normalized using the dynamization function described in equation 4.28. The blue colored graphs show the total energy consumption (primary axis); the purple colored graphs show the number of non-missing values per time slot (secondary axis).

Experiment No.	training data	test data
1	[01.02.2013; 01.01.2015)	[01.01.2015; 30.06.2018]
2	[01.07.2013; 01.07.2015)	[01.07.2015; 30.06.2018]
3	[01.01.2014; 01.01.2016)	[01.01.2016; 30.06.2018]
4	[01.07.2014; 01.07.2016)	[01.07.2016; 30.06.2018]
5	[01.01.2015; 01.01.2017)	[01.01.2017; 30.06.2018]
6	[01.07.2015; 01.07.2017)	[01.07.2017; 30.06.2018]

Table 5.1: Overview of the segmentations of the dataset shown in figure 5.1 into training data and test data for the experimental evaluation discussed in section 5.4.

In order to simulate the passing of time and the availability of *new* Smart Metering data to be incorporated into the analysis, we have opted to perform multiple experiments according to the approach outlined in section 5.3, each time with a different segmentation for the training data and test data. These segmentations are shown in table 5.1. While all experiments share the same results for the first step of the approach presented in section 5.3, all remaining steps are supposed to only include *recent* data, which we have implemented by removing half a year of *outdated* training data for the remaining steps of the approach as we simulate the availability of half a year of *new* training data.

The remaining details of our experiment are the same as in section 4.4.4 as this approach has yielded the overall best results of all experiments discussed in section 4.4, meaning that we have used WFCM with the YCF of the corresponding customer as the weight instead of the original Fuzzy-C-Means algorithm, we have used the dynamization function as given by equation 4.28 as well as *Partial Manhattan Distance* $dist_{PMD}(a, b)$ to measure the dissimilarity between data objects during clustering.

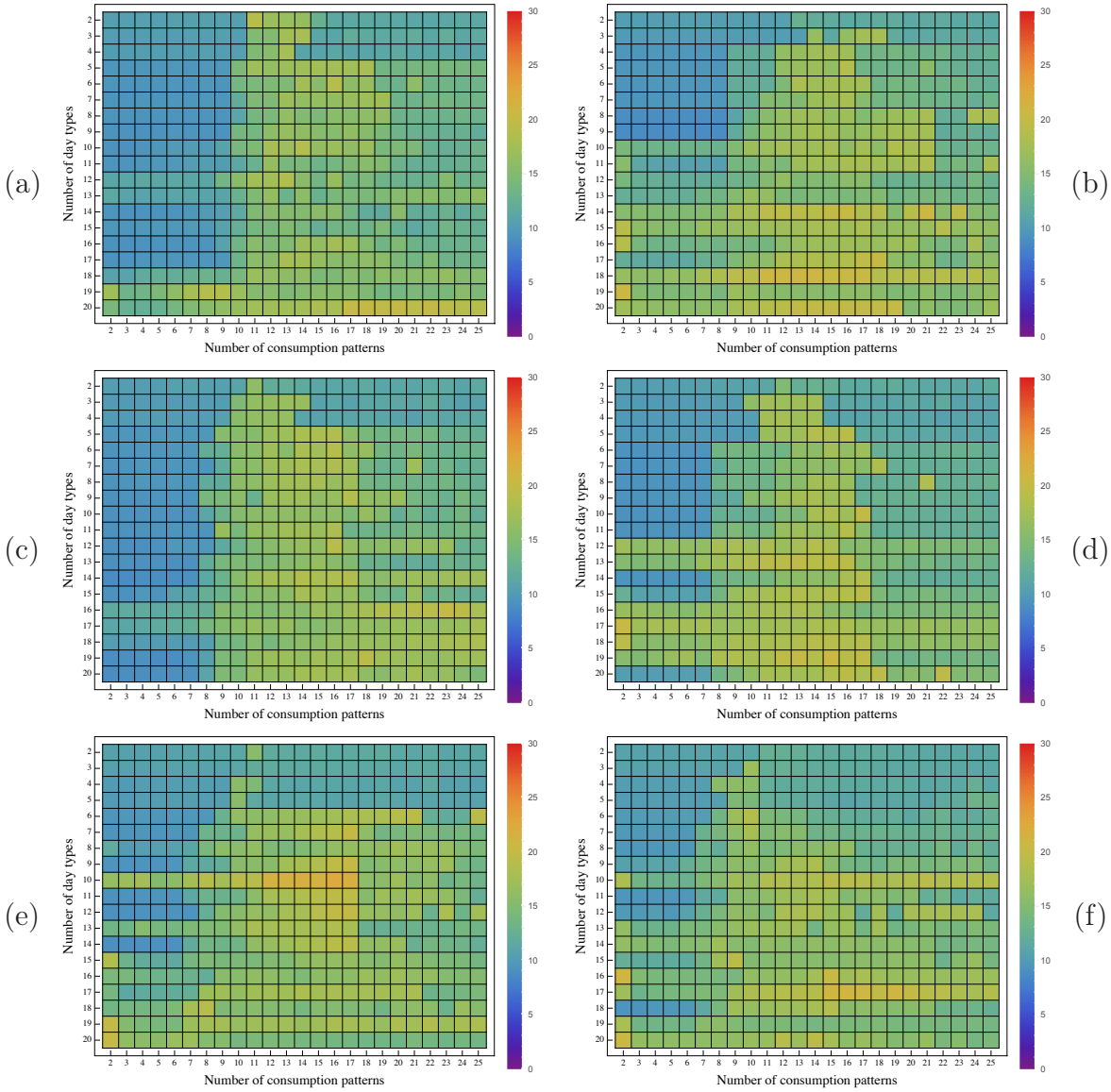


Figure 5.2: Ratio of the deviations and the actual consumption in percent yielded by the load profiles generated using different values for the number of day-types and consumption patterns using the experimental setup described in section 5.3. The graphs visualize the results for the dataset depicted in figure 5.1 using *K-Means++* to generate the starting configuration of the clustering process. The results shown correspond to (a) Experiment No. 1, (b) Experiment No. 2, (c) Experiment No. 3, (d) Experiment No. 4, (e) Experiment No. 5 and (f) Experiment No. 6 as outlined in table 5.1.

5.4.2 Results

The results of our experiments using Online Data Mining techniques as described in section 5.3 are visualized in figure 5.2 and 5.3.
One striking observation when looking at the results is the influence of the strategy to generate the starting configuration of the clustering prototypes. When using *K-Means++* [AV07], the resulting load profiles perform much better in terms of the relative forecast errors, which are roughly in the range from 10 up to 20 percent for the tested dataset. In contrast, the relative forecast errors when using *Random Co-*

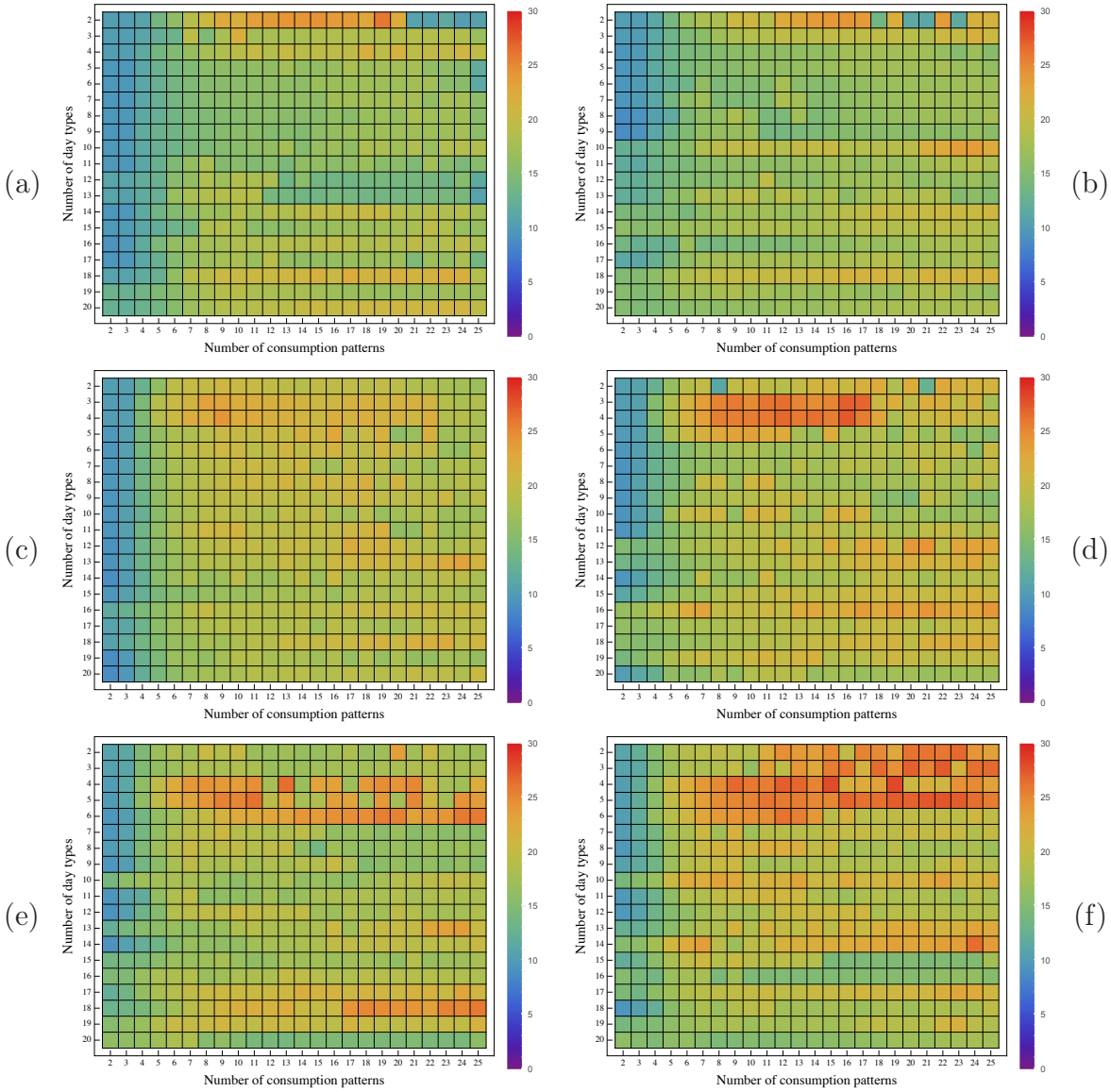


Figure 5.3: Ratio of the deviations and the actual consumption in percent yielded by the load profiles generated using different values for the number of day-types and consumption patterns using the experimental setup described in section 5.3. The graphs visualize the results for the dataset depicted in figure 5.1 using *Random Coordinates* to generate the starting configuration of the clustering process. The results shown correspond to (a) Experiment No. 1, (b) Experiment No. 2, (c) Experiment No. 3, (d) Experiment No. 4, (e) Experiment No. 5 and (f) Experiment No. 6 as outlined in table 5.1.

ordinates are roughly in the range from 10 up to 25 percent, although here the lower relative forecast errors have only been achieved for very small values for the number of consumption patterns.

Though the results between the two strategies to generate the starting configuration of the clustering prototypes differ greatly, the differences between the experiments for a given strategy are only marginal. This is likely due to the tested dataset not containing significant changes in consumer behavior for the tested segmentations in training and

test data.
Overall, the results show a slight regression in the quality of the resulting load profiles compared to the experimental evaluation discussed in section 4.4. Similar to the results for the *HHU-Dataset* on which the dataset for this experimental evaluation is based on, overlapping clusters and clusters of different sizes are among the possible reasons for the performance of the load profiles. For this reason, more research is necessary to achieve relative forecast errors similar to the experimental evaluation outlined in section 4.4.4.

6

Summary

This chapter represents the conclusion of the thesis. In section 6.1, we summarize the contributions of this thesis and describe the results of addressed problems. We then discuss starting points for future research in section 6.2.

6.1 Conclusions

The energy economy is one of the most crucial preconditions for enabling the lifestyle of modern society. As such, structural changes to the energy market face the challenge of increasing efficiency while maintaining uptime and fault tolerance of the security of the energy supply. With data analysis methods having gained increased importance following the possibility to store and process huge amounts of data, new technologies such as *Intelligent Metering Systems* have made it conceivable to employ techniques from the research area *Knowledge Discovery in Databases* to optimize business processes.

In this thesis, we have laid the focus on extracting useful knowledge from Smart Metering time series. For this purpose, in chapter 2, we have given a small introduction into some basic concepts relevant to market participants in the energy economy. In doing so, we have put emphasis on how the security of the energy supply is being organized, particularly with respect to forecasting the energy demand of customers using load profiles. As it has turned out, load profiles are a very simple yet effective means to forecast the total energy demand of customers. They work by segmenting the customer base into groups, where it is assumed that the differences in individuals consumption behavior of the same customer group will balance themselves out compared to the representative consumption pattern specific for the given customer group.

Furthermore, in chapter 3, we have presented some approaches from academic literature on how *Knowledge Discovery in Databases* techniques can be used in an *Intelligent Metering System* environment and discussed their implications concerning consumer privacy. While pilot projects within the context of the studies shown in this thesis so far have all resulted in a positive net benefit for the end consumer, other research work has revealed attack vectors on the lifestyle habits of consumers even if the data

is partly pseudonymized, raising valid concerns for data privacy. To counteract these concerns, data gathered by *Intelligent Metering Systems* should be kept locally as much as possible and privacy techniques such as *data reduction* should be employed where the transmission of information can not be avoided.
Additionally, technologies such as *Demand-Side Management* or *Demand Response* are likely to play an essential role in enabling the widespread usage of electric vehicles and other high-power household appliances such as intelligent washing machines and dishwashers. At the same time, Smart Metering devices with *Demand Response* capabilities can also help to semi-automatically reduce the energy bills of end consumers by being aware of volatile energy prices and scheduling household appliances accordingly.
For energy providers, *Intelligent Metering Systems* pose the possibility of learning more characteristics about their consumers. This information in turn can then be used to optimize business processes, for example by identifying common traits within groups of customers and creating new target-group-specific tariffs as a consequence.

Another very important task of energy providers that can be improved by employing *Intelligent Metering Systems* is to forecast the total energy demand of customers in order to plan the future buy-in of energy, which we have addressed in chapter 4 and 5 of this thesis. In chapter 4, we have introduced and evaluated several approaches based on a common framework to construct load profiles, an established model to predict customer consumption behavior. With the baseline *relative forecast error* of existing load profiles approximately being in the range from 12 to 13 percent, we have managed to achieve *relative forecast errors* below 6 percent, a significant increase in the forecasting quality of the models. By further combining the ability to generate load profiles with *Online Data Mining* techniques as discussed in chapter 5 of this thesis, we have been able to show that it is possible to keep load profiles updated and maintain low *relative forecast errors* even in the event of the consumption behavior of customers changing over time.

6.2 Future Work

In this thesis we have proposed multiple clustering approaches for building load profiles using Smart Metering data. While the experimental evaluation has shown promising results under certain conditions that can help to significantly improve on the existing *Standard Load Profiles* currently in use by the industry, there is still potential for the *relative forecast errors* of the load profiles constructed using our framework to be improved upon. For example, our experiments have shown that increasing the number of consumption patterns actually worsens the performance of the corresponding load profiles. One possible explanation for this observation might be the datasets containing clusters with varying sizes and shapes, which is a situation that Fuzzy-C-Means does not handle very well. In order to help process datasets that have data with this property, it might be worth considering to use *Fuzzy-Maximum-Likelihood-Estimation (FMLE)* [GG89], a clustering approach which incorporates the *fuzzy covariance matrix* [DAY69] to handle clusters of different sizes and shapes.

Another possible area for future research concerns the number of dimensions used during analysis. During our experimental evaluation, we have always used the highest available resolution of the time series per dataset since the load profiles have to have a quarter-hourly resolution in a real-world scenario as mentioned in chapter 2. However,

dimensionality reduction might further help in constructing better performing load profiles by counteracting the effects of the *curse of dimensionality* [Bey+99; FWV07], a term we have introduced in section 1.2.2. Approaches that might help with this problem include:

- To reduce the number of dimensions during analysis, it might be desirable to downscale the resolution of the time series, e.g. to an hourly resolution. Once the analysis is complete and load profiles have been constructed, the consumption patterns used can be scaled back up to quarter-hourly resolution using techniques such as *cubic spline interpolation*.

- Another way to reduce the number of dimensions without downscaling the time series is to split the tuples used during clustering into several groups, for example by having the first group containing only the data from midnight to 8 o'clock, the second group containing only data from 8 o'clock to 16 o'clock and the last group containing data from 16 o'clock to midnight. The framework as described in section 4.3 could then be applied to each of these groups individually so that in the end, the load profile for a given customer consists of a set of smaller load profiles, where one of these smaller load profiles describes the expected consumption behavior from midnight to 8 o'clock, the second smaller load profiles describes the consumption behavior from 8 o'clock to 16 o'clock and the last smaller load profiles describes the behavior from 16 o'clock to midnight. A similar idea in association with tariff groups segmentation has been described in [Chi12; Chi+03].

If the energy provider has detailed knowledge about which customers have a photovoltaic system installed, including its size and degree of efficiency, the energy provider might decide to use historic weather data to compute a theoretical *SLS* time series for each customer that would have arisen as a result of the customer not having a photovoltaic system installed. Depending on the capabilities of the installed Smart Metering device, this feature might already be available to an energy provider without resorting to error-prone estimates based on weather data. If possible, the energy provider might want to build the load profiles used to forecast the energy demand according to those consumption-only *SLS* time series, while at the same time using weather prognosis data to compute a correction term to be applied onto the forecast generated by the resulting load profiles in order to yield the *actual* energy the customer is likely to extract from the electricity grid.

Bibliography

[AE08] M.H. Albadi and E.F. El-Saadany. "A summary of demand response in electricity markets". In: *Electric Power Systems Research* 78.11 (2008), pp. 1989–1996. ISSN: 0378-7796. DOI: 10.1016/j.epsr.2008.04.002.

[Age14] International Energy Agency. *Policy uncertainty threatens to slow renewable energy momentum*. https://www.iea.org/newsroom/news/2014/august/policy-uncertainty-threatens-to-slow-renewable-energy-momentum.html. Aug. 2014.

[Agg+03] Charu C. Aggarwal et al. "A Framework for Clustering Evolving Data Streams". In: *Proceedings 2003 {VLDB} Conference*. Ed. by Johann-Christoph Freytag et al. San Francisco: Morgan Kaufmann, 2003, pp. 81–92. ISBN: 978-0-12-722442-8. DOI: 10.1016/B978-012722442-8/50016-1.

[Agr+95] Rakesh Agrawal et al. "Querying Shapes of Histories". In: *Proceedings of the 21th International Conference on Very Large Data Bases*. VLDB '95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 502–514. ISBN: 1-55860-379-4. URL: http://dl.acm.org/citation.cfm?id=645921.673157.

[Agr+98] Rakesh Agrawal et al. "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications". In: *SIGMOD Rec.* 27.2 (June 1998), pp. 94–105. ISSN: 0163-5808. DOI: 10.1145/276305.276314.

[ALB13] F.M. Andersen, H.V. Larsen, and T.K. Boomsma. "Long-term forecasting of hourly electricity load: Identification of consumption profiles and segmentation of customers". In: *Energy Conversion and Management* 68 (2013), pp. 244–252. ISSN: 0196-8904. DOI: 10.1016/j.enconman.2013.01.018.

[Alg+15] Hugo Algarvio et al. "Electricity Usage Efficiency in Large Buildings: DSM Measures and Preliminary Simulations of DR Programs in a Public Library". In: *Highlights of Practical Applications of Agents, Multi-Agent Systems, and Sustainability - The PAAMS Collection: International Workshops of PAAMS 2015, Salamanca, Spain, June 3-4, 2015. Proceedings*. Ed. by Javier Bajo et al. Cham: Springer International Publishing, 2015, pp. 249–259. ISBN: 978-3-319-19033-4. DOI: 10.1007/978-3-319-19033-4_21.

[And13] Andreas Kießling. *Modellstadt Mannheim (moma) - Abschlussbericht*. July 2013. URL: https://www.ifeu.de/projekt/modellstadt-mannheim/.

[Ank+99] Mihael Ankerst et al. "OPTICS: Ordering Points to Identify the Clustering Structure". In: *SIGMOD Rec.* 28.2 (June 1999), pp. 49–60. ISSN: 0163-5808. DOI: 10.1145/304181.304187.

[AV07] David Arthur and Sergei Vassilvitskii. "K-means++: The Advantages of Careful Seeding". In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '07. New Orleans, Louisiana: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035. ISBN: 978-0-898716-24-5. URL: http://dl.acm.org/citation.cfm?id=1283383.1283494.

[AZ08] Plamen P. Angelov and Xiaowei Zhou. "Evolving Fuzzy-Rule-Based Classifiers From Data Streams". In: *IEEE Transactions on Fuzzy Systems* 16.6 (Dec. 2008), pp. 1462–1475. ISSN: 1063-6706. DOI: 10.1109/TFUZZ.2008.925904.

[Bac78] Eric Backer. "Cluster analysis by optimal decomposition of induced fuzzy sets". PhD thesis. TU Delft, Delft University of Technology, 1978.

[Bar07] Gregory V. Bard. "Spelling-error Tolerant, Order-independent Pass-phrases via the Damerau-levenshtein String-edit Distance Metric". In: *Proceedings of the Fifth Australasian Symposium on ACSW Frontiers - Volume 68*. ACSW '07. Ballarat, Australia: Australian Computer Society, Inc., 2007, pp. 117–124. ISBN: 1-920-68285-X. URL: http://dl.acm.org/citation.cfm?id=1274531.1274545.

[BC94] Donald J Berndt and James Clifford. "Using dynamic time warping to find patterns in time series". In: *KDD workshop*. Vol. 10. 16. Seattle, WA. 1994, pp. 359–370.

[BD75] J. C. Bezdek and J. C. Dunn. "Optimal Fuzzy Partitions: A Heuristic for Estimating the Parameters in a Mixture of Normal Distributions". In: *IEEE Transactions on Computers* C-24.8 (Aug. 1975), pp. 835–838. ISSN: 0018-9340. DOI: 10.1109/T-C.1975.224317.

[Bey+99] Kevin Beyer et al. "When Is "Nearest Neighbor" Meaningful?" In: *Database Theory — ICDT'99*. Ed. by Catriel Beeri and Peter Buneman. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 217–235. ISBN: 978-3-540-49257-3.

[Bez74] J. C. Bezdek. "Numerical taxonomy with fuzzy sets". In: *Journal of Mathematical Biology* 1.1 (May 1974), pp. 57–71. ISSN: 1432-1416. DOI: 10.1007/BF02339490.

[Bez81] James C. Bezdek. "Objective Function Clustering". In: *Pattern Recognition with Fuzzy Objective Function Algorithms*. Boston, MA: Springer US, 1981, pp. 43–93. ISBN: 978-1-4757-0450-1. DOI: 10.1007/978-1-4757-0450-1_3.

[BH65] GH Ball and DJ Hall. "ISODATA, a novel technique for data analysis and pattern classification". In: *Standford Res. Inst., Menlo Park, CA* (1965).

[BK59] R. Bellman and R. Kalaba. "On adaptive control processes". In: *IRE Transactions on Automatic Control* 4.2 (Nov. 1959), pp. 1–9. ISSN: 0096-199X. DOI: 10.1109/TAC.1959.1104847.

[Boc16] Christian Bock. "Clustering-Ansatz zur Erstellung von Lastprofilen zur Vorhersage des Stromverbrauchs". In: *Proceedings of the 28th GI-Workshop Grundlagen von Datenbanken (GvDB 2016)*. Ed. by Lena Wiese, Hendrik Bitzmann, and Tim Waage. Nörten Hardenberg, Germany: CEUR-Workshop Proceedings, Vol. 1594, May 2016, pp. 21–26. URL: http://ceur-ws.org/Vol-1594/.

[Boc17] Christian Bock. "Generating Load Profiles Using Smart Metering Time Series". In: *Advances in Fuzzy Logic and Technology 2017: Proceedings of: EUSFLAT-2017 – The 10th Conference of the European Society for Fuzzy Logic and Technology, September 11–15, 2017, Warsaw, Poland IWIFSGN'2017 – The Sixteenth International Workshop on Intuitionistic Fuzzy Sets and Generalized Nets, September 13–15, 2017, Warsaw, Poland, Volume 1*. Ed. by Janusz Kacprzyk et al. Cham: Springer International Publishing, 2017, pp. 211–223. ISBN: 978-3-319-66830-7. DOI: 10.1007/978-3-319-66830-7_20.

[Boc18] Christian Bock. "Forecasting Energy Demand by Clustering Smart Metering Time Series". In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*. Ed. by Jesús Medina et al. Cham: Springer International Publishing, 2018, pp. 431–442. ISBN: 978-3-319-91473-2. DOI: 10.1007/978-3-319-91473-2_37.

[Bre01] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.

[Bun06] Bundesnetzagentur. *Richtlinie Datenaustausch und Mengenbilanzierung (DuM)*. 2006. URL: https://www.bundesnetzagentur.de/DE/Service-Funktionen/Beschlusskammern/1_GZ/BK6-GZ/2007/2007_0001bis0999/2007_001bis099/BK6-07-002/BK6-07-002_RichtlinieDatenaustauschundId8623doc_bf.html.

[Bun13] Bundesnetzagentur, Beschlusskammer 6. *Marktregeln für die Durchführung der Bilanzkreisabrechnung Strom (MaBiS)*. 2013. URL: https://www.bundesnetzagentur.de/DE/Service-Funktionen/Beschlusskammern/BK06/BK6_71_Bilanzabr/08%20achte%20Mitteilung/AchteMitteilung.html?nn=872140.

[Bun18] Bundesnetzagentur. *Festlegungsverfahren zur Änderung der Ausschreibungsbedingungen und Veröffentlichungspflichten von Minutenreserve (BK6-18-020)*. 2018. URL: https://www.bundesnetzagentur.de/DE/Service-Funktionen/Beschlusskammern/1_GZ/BK6-GZ/2018/2018_0001bis0999/BK6-18-020/BK6-18-020_Beschluss_vom_08_05_2018.html?nn=269594.

[Bun98] Bundesanzeiger. *Gesetz zur Neuregelung des Energiewirtschaftsrechts*. http://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&jumpTo=bgbl198s0730.pdf. Apr. 1998.

[BVK02] R. Babuka, P. J. van der Veen, and U. Kaymak. "Improved covariance estimation for Gustafson-Kessel clustering". In: *Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on*. Vol. 2. 2002, pp. 1081–1085. DOI: 10.1109/FUZZ.2002.1006654.

[BWS06] Mohamed Bouguessa, Shengrui Wang, and Haojun Sun. "An objective approach to cluster validation". In: *Pattern Recognition Letters* 27.13 (2006), pp. 1419–1430. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2006.01.015.

[Cao+] Feng Cao et al. "Density-Based Clustering over an Evolving Data Stream with Noise". In: *Proceedings of the 2006 SIAM International Conference on Data Mining*, pp. 328–339. DOI: 10.1137/1.9781611972764.29.

[CER] CER – The Commission for Energy Regulation. *Accessed via the Irish Social Science Data Archive*. http://www.ucd.ie/issda.

[Chi+01] Gianfranco Chicco et al. "Electric energy customer characterisation for developing dedicated market strategies". In: *2001 IEEE Porto Power Tech Proceedings (Cat. No.01EX502)*. Vol. 1. Sept. 2001. DOI: 10.1109/PTC.2001.964627.

[Chi+03] Gianfranco Chicco et al. "Customer characterization options for improving the tariff offer". In: *IEEE Transactions on Power Systems* 18.1 (Jan. 2003), pp. 381–387. ISSN: 0885-8950. DOI: 10.1109/TPWRS.2002.807085.

[Chi12] Gianfranco Chicco. "Overview and performance assessment of the clustering methods for electrical load pattern grouping". In: *Energy* 42.1 (2012). 8th World Energy System Conference, WESC 2010, pp. 68–80. ISSN: 0360-5442. DOI: 10.1016/j.energy.2011.12.031.

[Com12] European Commission. *2012/148/EU: Commission Recommendation of 9 March 2012 on preparations for the roll-out of smart metering systems*. http://data.europa.eu/eli/reco/2012/148/oj. Mar. 2012.

[Com14] European Commission. *2030 climate and energy policy framework*. https://ec.europa.eu/clima/policies/strategies/2030_en. Oct. 2014.

[Con07] 110th United States Congress. *Energy Independence and Security Act*. https://www.epa.gov/laws-regulations/summary-energy-independence-and-security-act. Dec. 2007.

[Dam64] Fred J. Damerau. "A Technique for Computer Detection and Correction of Spelling Errors". In: *Commun. ACM* 7.3 (Mar. 1964), pp. 171–176. ISSN: 0001-0782. DOI: 10.1145/363958.363994.

[DAY69] N. E. DAY. "Estimating the components of a mixture of normal distributions". In: *Biometrika* 56.3 (1969), pp. 463–474. DOI: 10.1093/biomet/56.3.463.

[Deg+13] Kathrin Degen et al. "Smart Metering, Beratung oder Sozialer Vergleich - Was beeinflusst den Elektrizitätsverbrauch?" In: (July 2013).

[DKK15] Panagiotis D. Diamantoulakis, Vasilios M. Kapinas, and George K. Karagiannidis. "Big Data Analytics for Dynamic Energy Management in Smart Grids". In: *CoRR* abs/1504.02424 (2015). URL: http://arxiv.org/abs/1504.02424.

[Dwo06] Cynthia Dwork. "Differential Privacy". In: *Automata, Languages and Programming*. Ed. by Michele Bugliesi et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–12. ISBN: 978-3-540-35908-1. DOI: 10.1007/11787006_1.

[EDI] EDI@Energy. URL: https://www.edi-energy.de.

[EK13] Helmut Edelmann and Thomas Kästner. *Kosten-Nutzen-Analyse für einen flächendeckenden Einsatz intelligenter Zähler*. Graf-Adolf-Platz 15, 40213 Düsseldorf, Deutschland, July 2013. URL: https://www.bmwi.de/Redaktion/DE/Publikationen/Studien/kosten-nutzen-analyse-fuer-flaechendeckenden-einsatz-intelligenterzaehler.html.

[Ene16] Bundesverband der Energie- und Wasserwirtschaft (BDEW). *Positionspapier: Umsetzung Gesetz zur Digitalisierung der Energiewende*. https://www.bdew.de/service/stellungnahmen/positionspapier-umsetzung-gesetz-digitalisierung-energiewende/. May 2016.

[ES00] Martin Ester and Jörg Sander. *Knowledge Discovery in Databases: Techniken und Anwendungen*. Springer Berlin Heidelberg, 2000. ISBN: 9783540673286. DOI: 10.1007/978-3-642-58331-5.

[Est+96] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *Kdd*. Vol. 96. 34. 1996, pp. 226–231.

[Eur50] Council of Europe. *Convention for the Protection of Human Rights and Fundamental Freedoms*. https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/005. Nov. 1950.

[Fer+15] Marta C. Ferreira et al. "Fuzzy modeling based on Mixed Fuzzy Clustering for health care applications". In: *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. Aug. 2015, pp. 1–5. DOI: 10.1109/FUZZ-IEEE.2015.7338028.

[FPS+96] Usama M Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, et al. "Knowledge Discovery and Data Mining: Towards a Unifying Framework." In: *KDD*. Vol. 96. 1996, pp. 82–88.

[FT00] Christian Fünfgeld and Remo Tiedemann. *Anwendung der Repräsentativen VDEW-Lastprofile step - by - step*. Tech. rep. VDEW-Materialien M-05/2000, Frankfurt. Brandenburgische Technische Universität Cottbus, Lehrstuhl für Energiewirtschaft, 2000.

[FWV07] Damien François, Vincent Wertz, and Michel Verleysen. "The Concentration of Fractional Distances". In: *IEEE Transactions on Knowledge & Data Engineering* 19 (2007), pp. 873–886. ISSN: 1041-4347. DOI: 10.1109/TKDE.2007.1037.

[GG89] I. Gath and A.B. Geva. "Unsupervised Optimal Fuzzy Clustering". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11 (1989), pp. 773–780. ISSN: 0162-8828.

[GK78] D. E. Gustafson and W. C. Kessel. "Fuzzy clustering with a fuzzy covariance matrix". In: *1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*. Jan. 1978, pp. 761–766. DOI: 10.1109/CDC.1978.268028.

[GP11] Arne Grein and Martin Pehnt. "Load management for refrigeration systems: Potentials and barriers". In: *Energy Policy* 39.9 (2011), pp. 5598–5608. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2011.04.040.

[GRS98] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. "CURE: An Efficient Clustering Algorithm for Large Databases". In: *SIGMOD Rec.* 27.2 (June 1998), pp. 73–84. ISSN: 0163-5808. DOI: 10.1145/276305.276312.

[Hal99] Mark Andrew Hall. "Correlation-based Feature Selection for Machine Learning". PhD thesis. 1999.

[Hav+12] Timothy Havens et al. "Fuzzy c-Means Algorithms for Very Large Data". In: 20 (Dec. 2012), pp. 1130–1146.

[Hay+14] Marian Hayn et al. "Electricity load profiles in Europe: The importance of household segmentation". In: *Energy Research & Social Science* 3 (2014), pp. 30–45. ISSN: 2214-6296. DOI: 10.1016/j.erss.2014.07.002.

[HB01] R. J. Hathaway and J. C. Bezdek. "Fuzzy c-means clustering of incomplete data". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 31.5 (Oct. 2001), pp. 735–744. ISSN: 1083-4419. DOI: 10.1109/3477.956035.

[HC10a] Ludmila Himmelspach and Stefan Conrad. "Clustering approaches for data with missing values: Comparison and evaluation". In: *2010 5th International Conference on Digital Information Management, ICDIM 2010*. Aug. 2010, pp. 19–28. DOI: 10.1109/ICDIM.2010.5664691.

[HC10b] Ludmila Himmelspach and Stefan Conrad. "Fuzzy Clustering of Incomplete Data Based on Cluster Dispersion". In: *Computational Intelligence for Knowledge-Based Systems Design*. Ed. by Eyke Hüllermeier, Rudolf Kruse, and Frank Hoffmann. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 59–68. ISBN: 978-3-642-14049-5. DOI: 10.1007/978-3-642-14049-5_7.

[HCC12] Ludmila Himmelspach, João Paulo Carvalho, and Stefan Conrad. "On Cluster Validity for Fuzzy Clustering of Incomplete Data". In: *Scalable Uncertainty Management*. Ed. by Eyke Hüllermeier et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 612–618. ISBN: 978-3-642-33362-0. DOI: 10.1007/978-3-642-33362-0_50.

[HHC11] Ludmila Himmelspach, Daniel Hommers, and Stefan Conrad. "Cluster Tendency Assessment for Fuzzy Clustering of Incomplete Data". In: *Proceedings of the 7th conference of the European Society for Fuzzy Logic and Technology*. Atlantis Press, Aug. 2011, pp. 290–297. ISBN: 978-90-78677-00-0. DOI: 10.2991/eusflat.2011.136.

[HHG07a] P. Hore, L. O. Hall, and D. B. Goldgof. "A fuzzy c means variant for clustering evolving data streams". In: *2007 IEEE International Conference on Systems, Man and Cybernetics*. Oct. 2007, pp. 360–365. DOI: 10.1109/ICSMC.2007.4413710.

[HHG07b] P. Hore, L. O. Hall, and D. B. Goldgof. "Single Pass Fuzzy C Means". In: *2007 IEEE International Fuzzy Systems Conference*. July 2007, pp. 1–7. DOI: 10.1109/FUZZY.2007.4295372.

[Him16] Ludmila Himmelspach. "Fuzzy Clustering of Incomplete Data". PhD thesis. Heinrich-Heine-University, Institute of Computer Science, 2016.

[Hof+12] Patrick Hoffman et al. *Praxistest "Moderne Energiesparsysteme im Haushalt"*. Altenkesseler Straße 17, D-66115 Saarbrücken, Germany, 2012.

[Hop+16] Konstantin Hopf et al. "Feature extraction and filtering for household classification based on smart electricity meter data". In: *Computer Science - Research and Development* 31.3 (Aug. 2016), pp. 141–148. ISSN: 1865-2042. DOI: 10.1007/s00450-014-0294-4.

[Hop+17] Konstantin Hopf et al. "A Decision Support System for Photovoltaic Potential Estimation". In: *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*. IML '17. Liverpool, United Kingdom: ACM, 2017, 3:1–3:10. ISBN: 978-1-4503-5243-7. DOI: 10.1145/3109761.3109764.

[Hor+08] P. Hore et al. "Online fuzzy c means". In: *NAFIPS 2008 - 2008 Annual Meeting of the North American Fuzzy Information Processing Society*. May 2008, pp. 1–5. DOI: 10.1109/NAFIPS.2008.4531233.

[HPK11] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011. ISBN: 9780123814791.

[HRR14] Michael Hinterstocker, Serafin von Roon, and Marina Rau. "Bewertung der aktuellen Standardlastprofile Österreichs und Analyse zukünftiger Anpassungsmöglichkeiten im Strommarkt". In: *Proc. 2014 13. Symposium Energieinnovation*. 2014.

[HSK16] Konstantin Hopf, Mariya Sodenkamp, and Ilya Kozlovskiy. "Energy Data Analytics For Improved Residential Service Quality And Energy Efficiency". In: *ENERGY* (2016). URL: http://aisel.aisnet.org/ecis2016_rip/73.

[HSS18] Konstantin Hopf, Mariya Sodenkamp, and Thorsten Staake. "Enhancing energy efficiency in the residential sector with smart meter data analytics". In: *Electronic Markets* (Mar. 2018). ISSN: 1422-8890. DOI: 10.1007/s12525-018-0290-9.

[IA09] Ali Ipakchi and Farrokh Albuyeh. "Grid of the future". In: *IEEE Power and Energy Magazine* 7.2 (Mar. 2009), pp. 52–62. ISSN: 1540-7977. DOI: 10.1109/MPE.2008.931384.

[IDK18] Omar A. Ibrahim, Yizhuo Du, and James Keller. "Robust On-Line Streaming Clustering". In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*. Ed. by Jesús Medina et al. Cham: Springer International Publishing, 2018, pp. 467–478. ISBN: 978-3-319-91473-2. DOI: 10.1007/978-3-319-91473-2_40.

[Ita75] F. Itakura. "Minimum prediction residual principle applied to speech recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 23.1 (Feb. 1975), pp. 67–72. ISSN: 0096-3518. DOI: 10.1109/TASSP.1975.1162641.

[JJR11] Marek Jawurek, Martin Johns, and Konrad Rieck. "Smart Metering Depseudonymization". In: *Proceedings of the 27th Annual Computer Security Applications Conference*. ACSAC '11. Orlando, Florida, USA: ACM, 2011, pp. 227–236. ISBN: 978-1-4503-0672-0. DOI: 10.1145/2076732.2076764.

[Jus05] Bundesministerium der Justiz und für Verbraucherschutz. *Stromnetzzugangsverordnung: Verordnung über den Zugang zu Elektrizitätsversorgungsnetzen*. https://www.gesetze-im-internet.de/stromnzv/. July 2005.

[Jus16] Bundesministerium der Justiz und für Verbraucherschutz. *Gesetz über den Messstellenbetrieb und die Datenkommunikation in intelligenten Energienetzen*. https://www.gesetze-im-internet.de/messbg/. Aug. 2016.

[KKK04] Karin Kailing, Hans-Peter Kriegel, and Peer Kröger. "Density-Connected Subspace Clustering for High-Dimensional Data". In: *Proceedings of the 2004 SIAM International Conference on Data Mining*. 2004, pp. 246–256. DOI: 10.1137/1.9781611972740.23.

[KL83] Joseph B. Kruskal and Mark Liberman. "The symmetric time-warping problem: From continuous to discrete". In: (Jan. 1983).

[Koh89] Teuvo Kohonen. *Self-Organization and Associative Memory*. 3rd ed. Springer Series in Information Sciences 8. Springer-Verlag Berlin Heidelberg, 1989. ISBN: 978-3-540-51387-2,978-3-642-88163-3.

[KP01] Eamonn J. Keogh and Michael J. Pazzani. "Derivative Dynamic Time Warping". In: *Proceedings of the 2001 SIAM International Conference on Data Mining*. 2001, pp. 1–11. DOI: 10.1137/1.9781611972719.1.

[KR90] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 1990. ISBN: 9780471878766. DOI: 10.1002/9780470316801.

[Lam12] Oscar Perpiñan Lamigueiro. "solaR: Solar Radiation and Photovoltaic Systems with R". In: *Journal of Statistical Software* 50.9 (Aug. 2012), pp. 1–32. URL: http://oa.upm.es/20821/.

[Lev66] Vladimir Iosifovich Levenshtein. "Binary codes capable of correcting deletions, insertions and reversals." In: *Soviet Physics Doklady* 10.8 (Feb. 1966). Doklady Akademii Nauk SSSR, V163 No4 845-848 1965, pp. 707–710.

[Lia05] T. Warren Liao. "Clustering of time series data — a survey". In: *Pattern Recognition* 38.11 (2005), pp. 1857–1874. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2005.01.025.

[Lit88] Roderick J. A. Little. "A Test of Missing Completely at Random for Multivariate Data with Missing Values". In: *Journal of the American Statistical Association* 83.404 (1988), pp. 1198–1202. DOI: 10.1080/01621459.1988.10478722.

[Löd+10] Martin Lödl et al. "Abschätzung des Photovoltaik-Potentials auf Dachflächen in Deutschland". In: *11. Symposium Energieinnovation „Alte Ziele–Neue Wege "*. 2010.

[Lop11] Raul H. C. Lopes. "Kolmogorov-Smirnov Test". In: *International Encyclopedia of Statistical Science*. Ed. by Miodrag Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 718–720. ISBN: 978-3-642-04898-2. DOI: 10.1007/978-3-642-04898-2_326.

[LR02] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data, Second Edition*. New York: John Wiley & Sons, 2002. ISBN: 978-0-471-18386-0. DOI: 10.1002/9781119013563.

[Mac+67] James MacQueen et al. "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. Oakland, CA, USA. 1967, pp. 281–297.

[Mei+99] Hermann Meier et al. *Repräsentative VDEW-Lastprofile*. Tech. rep. VDEW-Materialien M-32/99, Frankfurt. Brandenburgische Technische Universität Cottbus, Lehrstuhl für Energiewirtschaft, 1999.

[MLB] Masud Moshtaghi, Christopher Leckie, and James C. Bezdek. "Online Clustering of Multivariate Time-series". In: *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 360–368. DOI: 10.1137/1.9781611974348.41.

[Moh+10] A. H. Mohsenian-Rad et al. "Autonomous Demand-Side Management Based on Game-Theoretic Energy Consumption Scheduling for the Future Smart Grid". In: *IEEE Transactions on Smart Grid* 1.3 (Dec. 2010), pp. 320–331. ISSN: 1949-3053. DOI: 10.1109/TSG.2010.2089069.

[MRR80] C. Myers, L. Rabiner, and A. Rosenberg. "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.6 (Dec. 1980), pp. 623–635. ISSN: 0096-3518. DOI: 10.1109/TASSP.1980.1163491.

[MRT12] Eoghan McKenna, Ian Richardson, and Murray Thomson. "Smart meter data: Balancing consumer privacy concerns with legitimate applications". In: *Energy Policy* 41 (2012). Modeling Transport (Energy) Demand and Policies, pp. 807–814. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2011.11.049.

[Mül+16] Oliver Müller et al. "Utilizing big data analytics for information systems research: challenges, promises and guidelines". In: *European Journal of Information Systems* 25.4 (2016), pp. 289–302. DOI: 10.1057/ejis.2016.2.

[OCa+02] Liadan O'Callaghan et al. "Streaming-data algorithms for high-quality clustering". In: *Proceedings 18th International Conference on Data Engineering*. Feb. 2002, pp. 685–694. DOI: 10.1109/ICDE.2002.994785.

[Ord+10] J. Ordóñez et al. "Analysis of the photovoltaic solar energy capacity of residential rooftops in Andalusia (Spain)". In: *Renewable and Sustainable Energy Reviews* 14.7 (2010), pp. 2122–2130. ISSN: 1364-0321. DOI: 10.1016/j.rser.2010.01.001.

[Par06] European Parliament. *Directive 2006/32/EC of the European Parliament and of the Council of 5 April 2006 on energy end-use efficiency and energy services and repealing and repealing Council Directive 93/76/EEC*. http://data.europa.eu/eli/dir/2006/32/oj. Apr. 2006.

[Par09] European Parliament. *Directive 2009/72/EC of the European Parliament and of the Council of 13 July 2009 concerning common rules for the internal market in electricity and repealing Directive 2003/54/EC*. http://data.europa.eu/eli/dir/2009/72/oj. Aug. 2009.

[Par15] European Parliament. *Directive (EU) 2015/1535 of the European Parliament and of the Council of 9 September 2015 laying down a procedure for the provision of information in the field of technical regulations and of rules on Information Society services (codification)*. http://data.europa.eu/eli/dir/2015/1535/oj. Sept. 2015.

[Par96] European Parliament. *Directive 96/92/EC of the European Parliament and of the Council of 19 December 1996 concerning common rules for the internal market in electricity*. http://data.europa.eu/eli/dir/1996/92/oj. Dec. 1996.

[PSM09] Michael Angelo A. Pedrasa, Ted D. Spooner, and Iain F. MacGill. "Scheduling of Demand Side Resources Using Binary Particle Swarm Optimization". In: *IEEE Transactions on Power Systems* 24.3 (Aug. 2009), pp. 1173–1181. ISSN: 0885-8950. DOI: 10.1109/TPWRS.2009.2021219.

[Ren08] Rena Tangens. *BigBrotherAward 2008 in der Kategorie "Technik"*. Tech. rep. BigBrotherAwards, 2008. URL: https://bigbrotherawards.de/2008/technik-yello-strom.

[Rod+03] Fátima Rodrigues et al. "Machine Learning and Data Mining in Pattern Recognition: Third International Conference, MLDM 2003 Leipzig, Germany, July 5–7, 2003 Proceedings". In: ed. by Petra Perner and Azriel Rosenfeld. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003. Chap. A Comparative Analysis of Clustering Algorithms Applied to Load Profiling, pp. 73–85. ISBN: 978-3-540-45065-8. DOI: 10.1007/3-540-45065-3_7.

[Roo+14] Serafin von Roon et al. *Statusbericht zum Standardlastprofilverfahren Gas*. https://www.ffegmbh.de/kompetenzen/system-markt-analysen/508-statusbericht-standardlastprofile-gas. Am Blütenanger 71, 80995 München, Deutschland, Nov. 2014.

[Rou78] Marc Roubens. "Pattern classification problems and fuzzy sets". In: *Fuzzy Sets and Systems* 1.4 (1978), pp. 239–253. ISSN: 0165-0114. DOI: 10.1016/0165-0114(78)90016-7.

[Rou87] Peter 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. ISSN: 0377-0427. DOI: 10.1016/0377-0427(87)90125-7.

[RV08] B. Ramanathan and V. Vittal. "A Framework for Evaluation of Advanced Direct Load Control With Minimum Disruption". In: *IEEE Transactions on Power Systems* 23.4 (Nov. 2008), pp. 1681–1688. ISSN: 0885-8950. DOI: 10.1109/TPWRS.2008.2004732.

[RW+12] C. Rudin, D. Waltz, et al. "Machine Learning for the New York City Power Grid". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 34.2 (Feb. 2012), pp. 328–345. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2011.108.

[SC07] Stan Salvador and Philip Chan. "Toward Accurate Dynamic Time Warping in Linear Time and Space". In: *Intell. Data Anal.* 11.5 (Oct. 2007), pp. 561–580. ISSN: 1088-467X. URL: http://dl.acm.org/citation.cfm?id=1367985.1367993.

[SC78] H. Sakoe and S. Chiba. "Dynamic programming algorithm optimization for spoken word recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26.1 (Feb. 1978), pp. 43–49. ISSN: 0096-3518. DOI: 10.1109/TASSP.1978.1163055.

[Sch+15] Hanna Schäfer et al. "Analysing the segmentation of energy consumers using mixed fuzzy clustering". In: *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. Aug. 2015, pp. 1–7. DOI: 10.1109/FUZZ-IEEE.2015.7338120.

[Sch12] Tim Schlüter. "Knowledge Discovery from Time Series". PhD thesis. Heinrich-Heine-University, Institute of Computer Science, 2012.

[She00] Colin Shearer. "The CRISP-DM model: the new blueprint for data mining". In: *Journal of Data Warehouse* 5.4 (2000), pp. 13–22.

[Sic] Bundesamt für Sicherheit in der Informationstechnik. *BSI TR-03109 Technische Vorgaben für intelligente Messsysteme und deren sicherer Betrieb*. https://www.bsi.bund.de/DE/Publikationen/TechnischeRichtlinien/tr03109/index_htm.html.

[Sil+05] Vera Silva et al. "An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques". In: *Power Systems, IEEE Transactions on* 20 (June 2005), pp. 596–602. DOI: 10.1109/TPWRS.2005.846234.

[SL01] Manish Sarkar and Tze-Yun Leong. "Fuzzy K-Means Clustering with Missing Values". In: *Proceedings of American Medical Informatics Association Annual Fall Symposium (AMIA)*. 2001, pp. 588–592.

[SL07] Kathleen Spees and Lester B. Lave. "Demand Response and Electricity Market Efficiency". In: *The Electricity Journal* 20.3 (2007), pp. 69–85. ISSN: 1040-6190. DOI: 10.1016/j.tej.2007.01.006.

[SM17] D. Scholz and F. Müsgens. "How to improve standard load profiles: Updating, regionalization and smart meter data". In: *2017 14th International Conference on the European Energy Market (EEM)*. June 2017, pp. 1–6. DOI: 10.1109/EEM.2017.7981939.

[Sod+17] Mariya A. Sodenkamp et al. "Smart Meter Data Analytics for Enhanced Energy Efficiency in the Residential Sector". In: *Towards Thought Leadership in Digital Transformation: 13. Internationale Tagung Wirtschaftsinformatik, WI 2017, St.Gallen, Switzerland, February 12-15, 2017*. Ed. by Jan Marco Leimeister and Walter Brenner. 2017. URL: http://aisel.aisnet.org/wi2017/track12/paper/10.

[Sta13] International Organization for Standardization (ISO). *ISO/IEC 27001:2013*. BIBC II, Chemin de Blandonnet 8, CP 401, 1214 Vernier, Geneva, Switzerland, 2013.

[Sto74] M. Stone. "Cross-Validatory Choice and Assessment of Statistical Predictions". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 36.2 (1974), pp. 111–147. ISSN: 00359246. URL: http://www.jstor.org/stable/2984809.

[Str08] Goran Strbac. "Demand side management: Benefits and challenges". In: *Energy Policy* 36.12 (2008). Foresight Sustainable Energy Management and the Built Environment Project, pp. 4419–4426. ISSN: 0301-4215. DOI: 10.1016/j.enpol.2008.09.030.

[VVS15] Joaquim P. L. Viegas, Susana M. Vieira, and João M. C. Sousa. "Fuzzy clustering and prediction of electricity demand based on household characteristics". In: *2015 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT-15), Gijón, Spain., June 30, 2015.* Ed. by José M. Alonso, Humberto Bustince, and Marek Reformat. Atlantis Press, 2015. ISBN: 978-94-62520-77-6.

[VVS16] Joaquim L. Viegas, Susana M. Vieira, and João M. C. Sousa. "Mining Consumer Characteristics from Smart Metering Data through Fuzzy Modelling". In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems: 16th International Conference, IPMU 2016, Eindhoven, The Netherlands, June 20-24, 2016, Proceedings, Part I.* Ed. by Paulo Joao Carvalho et al. Cham: Springer International Publishing, 2016, pp. 562–573. ISBN: 978-3-319-40596-4. DOI: 10.1007/978-3-319-40596-4_47.

[WB05] Christel Weiß and Peter Bucsky. *Basiswissen Medizinische Statistik.* Springer-Verlag Berlin Heidelberg, 2005. ISBN: 978-3-540-28549-6. DOI: 10.1007/3-540-28549-0.

[Wik] Wikipedia. *Stromversorgung.* https://de.wikipedia.org/wiki/Datei:Stromversorgung.svg.

[Wir16] Bundesministerium für Wirtschaft und Energie. *Gesetz zur Digitalisierung der Energiewende.* Aug. 2016. URL: https://www.bmwi.de/Redaktion/DE/Downloads/Gesetz/gesetz-zur-digitalisierung-der-energiewende.html.

[WL11] Zhimin Wang and Furong Li. "Critical peak pricing tariff design for mass consumers in Great Britain". In: *2011 IEEE Power and Energy Society General Meeting.* July 2011, pp. 1–6. DOI: 10.1109/PES.2011.6039603.

[WNP10] L.K. Wiginton, H.T. Nguyen, and J.M. Pearce. "Quantifying rooftop solar photovoltaic potential for regional renewable energy policy". In: *Computers, Environment and Urban Systems* 34.4 (2010). Geospatial Cyberinfrastructure, pp. 345–357. ISSN: 0198-9715. DOI: 10.1016/j.compenvurbsys.2010.01.001.

[WS18] James Watson and Michael Schmela. *Global Market Outlook 2018-2022.* Tech. rep. Solar Power Europe, 2018. URL: http://www.solarpowereurope.org/global-market-outlook-2018-2022/.

[Wüs17a] Christian Wüst. "Blackout im Parkhaus". In: *DER SPIEGEL* 43 (Oct. 2017), pp. 116–117.

[Wüs17b] Christian Wüst. "Strom-Illusionen". In: *DER SPIEGEL* 34 (Aug. 2017), pp. 118–120.

[XB91] Xuanli Lisa Xie and Gerardo Beni. "A Validity Measure for Fuzzy Clustering". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 13.8 (Aug. 1991), pp. 841–847. ISSN: 0162-8828.

[Zha+14] Jing Zhao et al. "Achieving differential privacy of data disclosure in the smart grid". In: *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*. Apr. 2014, pp. 504–512. DOI: 10.1109/INFOCOM.2014.6847974.

List of Figures

1.1 Overview of the transformation of the electricity grid due to digitization and growing deployment of *Intelligent Metering Systems*. Adapted from [EK13] with some assets taken from [Wik]. 2
1.2 Overview of the processes associated with *Knowledge Discovery in Databases*, adapted from [FPS+96]. 3
1.3 Schematic example for two time series whose distance is computed using (a) the euclidean distance and (b) *Dynamic Time Warping (DTW)* [BK59]. The *Warping Path* is visualized by black lines connecting the measurements of the two time series which have been matched up. . . . 10
1.4 Illustration of extensions to *Dynamic Time Warping (DTW)* [BK59]. The figures show two time series (red and cyan graph) aligned to the $n \times m$-matrix used to construct the *Warping Path*. The depicted approaches are (a) the *Sakoe-Chiba Band* [SC78] and (b) the *Itakura Parallelogram* [Ita75]. The area colored in dark green correspond to the region the *Warping Path* of DTW is supposed to not trespass. 11

2.1 Overview of the general layout of electricity grids in europe. *Extra-High Voltage* is notated by dark red lines, *High Voltage* by bright red lines, *Medium Voltage* by yellow lines and *Low Voltage* by black lines. Interleaving circles depict transformer stations. Adapted from [Wik]. . 14
2.2 Overview of the Standard Load Profiles (a) *H0* (household profile) and (b) *G0* (general purpose industrial profile) from the *BDEW*, valid only during summer. The time series shown are normalized and describe the expected consumption behavior for the day-types *Working day*, *Saturday* and *Sunday and Holiday*. 19
2.3 Depiction of how load profiles are used to forecast the electric meter reading of a given customer. The normalized time series of a load profile are concatenated based on the day-type segmentation and integrated over the course of one year to form a time series describing the normalized meter reading of the customer, which is then scaled based on the *YCF* to yield a forecast for the actual meter reading of the customer. . 20
2.4 Schematic representation of how the forecast of the total energy consumption is derived based on the customer base of an energy provider, their individual *YCF* as well as the load profile assignments. 21

2.5 Comparison of the ratios of the absolute difference between the actual consumption and the predicted load yielded when using the *BDEW Standard Load Profiles* to the actual consumption in percent for 10 randomly selected energy providers in Germany. The monthly ratios shown correspond to the year 2017. The energy providers are sorted ascendingly according to their yearly average deviation ratio. 24

3.1 Comparison of capabilities of traditional electricity meters against *Modern Metering Devices* and *Intelligent Metering Systems*. 28

3.2 Visualization of how a consumption time series (left side) is transformed into a binary feature-vector $\phi(f_{i,l})$ depicted as a grid (right side) as part of the *Linking by behavior anomaly* attack vector [JJR11]. Black colored bins correspond to a value of 1, white colored bins to a value of 0. . . . 34

3.3 Overview of Demand-Side Management strategies focusing on (a) individual interactions between the energy provider and each customer and (b) the Smart Grid with enabled communication between the customers and the energy provider. Adapted from [Moh+10] with some assets taken from [Wik]. 44

4.1 Overview of the (a) *HHU-Dataset*, (b) *CER-Dataset* and (c) *IZES-Dataset*. The blue colored graphs show the total energy consumption (primary axis); the purple colored graphs show the number of non-missing values per time slot (secondary axis). 58

4.2 Visualization of how the Dataset D is constructed from the time series X by splitting the time series into tuples at the start of a new calendar day. 67

4.3 Ratio of the deviations and the actual consumption in percent yielded by the load profiles generated using different values for the number of day-types and consumption patterns using the experimental setup described in section 4.4.1.1. The graphs visualize the results for (a)(b) the *HHU-Dataset*, (c)(d) the *CER-Dataset* and (e)(f) the *IZES-Dataset* using (a)(c)(e) *K-Means++* and (b)(d)(f) *Random Coordinates* to generate the starting configuration of the clustering process. 74

4.4 Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on 2 day-types and 2 consumption patterns per day-type and using *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.1.1 for (a) the *HHU-Dataset*, (b) the *CER-Dataset* and (c) the *IZES-Dataset*. The green graph shows the absolute deviation of the forecast from the actual consumption. . . 75

4.5 Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on 10 day-types and 25 consumption patterns per day-type and using *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.1.1 for (a) the *HHU-Dataset*, (b) the *CER-Dataset* and (c) the *IZES-Dataset*. The green graph shows the absolute deviation of the forecast from the actual consumption. . . 76

4.6 Overview of the day-type segmentations for the (a) (b) *HHU-Dataset* and (c) (d) *CER-Dataset* yielded by using (a) (c) 4 day-types and (b) (d) 8 day-types. The graphs have been colorized depending on which cluster the total energy consumption time series has been assigned to on a given calendar day. 77

4.7 Visualization of the dynamization functions discovered by applying the *Fourier transformation* on the *HHU-Dataset* (green graph), the *CER-Dataset* (blue graph) and the *IZES-Dataset* (orange graph). For comparison, the dynamization function for the *BDEW Standard Load Profiles* is depicted in this diagram as the red graph. 79

4.8 Overview of the (a) *HHU-Dataset*, (b) *CER-Dataset* and (c) *IZES-Dataset*, where each measurement has been divided by the corresponding value of $DynFactor_{HHU}$, $DynFactor_{CER}$ and $DynFactor_{IZES}$, respectively. 80

4.9 Ratio of the deviations and the actual consumption in percent yielded by the load profiles generated using different values for the number of day-types and consumption patterns using the experimental setup described in section 4.4.2.1. The graphs visualize the results for (a)(b) the *HHU-Dataset*, (c)(d) the *CER-Dataset* and (e)(f) the *IZES-Dataset* using (a)(c)(e) *K-Means++* and (b)(d)(f) *Random Coordinates* to generate the starting configuration of the clustering process. 81

4.10 Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on 2 day-types and 2 consumption patterns per day-type and using *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.2.1 for (a) the *HHU-Dataset*, (b) the *CER-Dataset* and (c) the *IZES-Dataset*. The green graph shows the absolute deviation of the forecast from the actual consumption. . . 82

4.11 Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on 10 day-types and 25 consumption patterns per day-type and using *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.2.1 for (a) the *HHU-Dataset*, (b) the *CER-Dataset* and (c) the *IZES-Dataset*. The green graph shows the absolute deviation of the forecast from the actual consumption. . . 83

4.12 Ratio of the deviations and the actual consumption in percent yielded by the load profiles generated using different values for the number of day-types and consumption patterns using the experimental setup described in section 4.4.3.1. The graphs visualize the results for (a)(b) the *HHU-Dataset*, (c)(d) the *CER-Dataset* and (e)(f) the *IZES-Dataset* using (a)(c)(e) *K-Means++* and (b)(d)(f) *Random Coordinates* to generate the starting configuration of the clustering process. 85

4.13 Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on 2 day-types and 2 consumption patterns per day-type and using *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.3.1 for (a) the *HHU-Dataset*, (b) the *CER-Dataset* and (c) the *IZES-Dataset*. The green graph shows the absolute deviation of the forecast from the actual consumption. . . 86

4.14 Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on 10 day-types and 25 consumption patterns per day-type and using *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.3.1 for (a) the *HHU-Dataset*, (b) the *CER-Dataset* and (c) the *IZES-Dataset*. The green graph shows the absolute deviation of the forecast from the actual consumption. . . 87

4.15 Visualization of two customers and the clustering prototype that the clustering process would yield if the position of the clustering prototype was based solely on the distance in the feature space (black dot) and weighted according to the product of *Year Consumption Forecast* of the data tuple and the distance in the feature space (gray dot). 87

4.16 Ratio of the deviations and the actual consumption in percent yielded by the load profiles generated using different values for the number of day-types and consumption patterns using the experimental setup described in section 4.4.4.1. The graphs visualize the results for (a)(b) the *HHU-Dataset*, (c)(d) the *CER-Dataset* and (e)(f) the *IZES-Dataset* using (a)(c)(e) *K-Means++* and (b)(d)(f) *Random Coordinates* to generate the starting configuration of the clustering process. 88

4.17 Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on 2 day-types and 2 consumption patterns per day-type and using *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.4.1 for (a) the *HHU-Dataset*, (b) the *CER-Dataset* and (c) the *IZES-Dataset*. The green graph shows the absolute deviation of the forecast from the actual consumption. . . 89

4.18 Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on 10 day-types and 25 consumption patterns per day-type and using *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.4.1 for (a) the *HHU-Dataset*, (b) the *CER-Dataset* and (c) the *IZES-Dataset*. The green graph shows the absolute deviation of the forecast from the actual consumption. . . 90

4.19 Comparison of the actual total energy consumption (red graph) and the consumption predicted using the load profiles based on *K-Means++* for generating the starting cluster configuration (orange graph) according to the experimental setup described in section 4.4.4.1 for (a) the *HHU-Dataset* when using 4 day-types and 25 consumption patterns per day-type and (b) the *CER-Dataset* when using 2 day-types and 25 consumption patterns per day-type. The green graph shows the absolute deviation of the forecast from the actual consumption. 91

5.1 Overview of a subset of the *HHU-Dataset* used for the experimental evaluation of the approach described in section 5.3 (a) *as-is* and (b) normalized using the dynamization function described in equation 4.28. The blue colored graphs show the total energy consumption (primary axis); the purple colored graphs show the number of non-missing values per time slot (secondary axis). 99

5.2 Ratio of the deviations and the actual consumption in percent yielded by the load profiles generated using different values for the number of day-types and consumption patterns using the experimental setup described in section 5.3. The graphs visualize the results for the dataset depicted in figure 5.1 using *K-Means++* to generate the starting configuration of the clustering process. The results shown correspond to (a) Experiment No. 1, (b) Experiment No. 2, (c) Experiment No. 3, (d) Experiment No. 4, (e) Experiment No. 5 and (f) Experiment No. 6 as outlined in table 5.1. 100

5.3 Ratio of the deviations and the actual consumption in percent yielded by the load profiles generated using different values for the number of day-types and consumption patterns using the experimental setup described in section 5.3. The graphs visualize the results for the dataset depicted in figure 5.1 using *Random Coordinates* to generate the starting configuration of the clustering process. The results shown correspond to (a) Experiment No. 1, (b) Experiment No. 2, (c) Experiment No. 3, (d) Experiment No. 4, (e) Experiment No. 5 and (f) Experiment No. 6 as outlined in table 5.1. 101

List of Tables

2.1 Overview of the *BDEW Standard Load Profiles* used by most German energy providers and their common use-cases as defined in [Mei+99; FT00]. 23

3.1 Overview of the Smart Metering pilot projects in different european countries. Adapted from [EK13]. 31

4.1 Overview of the results of the χ^2 test for the MCAR failure mechanism according to [Lit88]. 60

5.1 Overview of the segmentations of the dataset shown in figure 5.1 into training data and test data for the experimental evaluation discussed in section 5.4. 99

List of Own Publications

[Boc16] Christian Bock. "Clustering-Ansatz zur Erstellung von Lastprofilen zur Vorhersage des Stromverbrauchs". In: *Proceedings of the 28th GI-Workshop Grundlagen von Datenbanken (GvDB 2016)*. Ed. by Lena Wiese, Hendrik Bitzmann, and Tim Waage. Nörten Hardenberg, Germany: CEUR-Workshop Proceedings, Vol. 1594, May 2016, pp. 21–26. URL: http://ceur-ws.org/Vol-1594/.

[Boc17] Christian Bock. "Generating Load Profiles Using Smart Metering Time Series". In: *Advances in Fuzzy Logic and Technology 2017: Proceedings of: EUSFLAT-2017 – The 10th Conference of the European Society for Fuzzy Logic and Technology, September 11–15, 2017, Warsaw, Poland IWIFSGN'2017 – The Sixteenth International Workshop on Intuitionistic Fuzzy Sets and Generalized Nets, September 13–15, 2017, Warsaw, Poland, Volume 1*. Ed. by Janusz Kacprzyk et al. Cham: Springer International Publishing, 2017, pp. 211–223. ISBN: 978-3-319-66830-7. DOI: 10.1007/978-3-319-66830-7_20.

[Boc18] Christian Bock. "Forecasting Energy Demand by Clustering Smart Metering Time Series". In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*. Ed. by Jesús Medina et al. Cham: Springer International Publishing, 2018, pp. 431–442. ISBN: 978-3-319-91473-2. DOI: 10.1007/978-3-319-91473-2_37.