University of Technology Sydney

Faculty of Engineering and Information Technology

Analysis of Residential Load Data and Its Applications for Smart Grids

A thesis submitted for the degree of

Doctor of Philosophy

Amin Rajabi

(2020)

Title of the thesis:

Analysis of Residential Load Data and Its Applications for Smart Grids

Ph.D. student:

Amin Rajabi

Email: [redacted]@student.uts.edu.au

Supervisor:

A/Prof. Li Li

E-mail: Li.Li@uts.edu.au

Co-supervisor:

Dr. Jiangfeng Zhang

E-mail: Jiangfeng.Zhang@uts.edu.au

Address:

School of Electrical and Data Engineering,

University of Technology Sydney, 15 Broadway, Ultimo, NSW 2007, Australia

Certificate of Original Authorship

I, Amin Rajabi, declare that this thesis is submitted in fulfilment of the requirements for the award of PhD degree, in the school of Electrical and Data Engineering, faculty of Engineering and Information Technology at the University of Technology Sydney. This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program Scholarship.

Signature:

Production Note:
Signature removed prior to publication.

Amin Rajabi

Date: 30/11/2019

Acknowledgments

I would like to take this opportunity to thank the following people and organisations for their assistance and support during my candidature.

First, I would like to express my deep gratitude to my principal supervisor Assoc. Prof. Li Li and my co-supervisor Dr Jiangfeng Zhang for their guidance, encouragement, and belief in my abilities. Their constructive comments contributed significantly to the quality of this research. I also would like to thank Prof. Jianguo Zhu for his help and advice.

I wish to thank the Irish Social Science Data Archive (ISSDA) for providing me with the datasets which are used in this thesis. I thank the UTS Advanced Research Computing Laboratory and its manager Dr Matthew Gaston for giving me substantial technical supports. I also appreciate the help and constructive advice of Prof. David McGloin.

I thank all my colleagues and friends, especially Sahand Ghavidel, Mojtaba Jabbari Ghadi, Ali Azizivahed, Mohsen Eskandari, and Mohammad Abuhilaleh with whom I have shared significant knowledge and experience during the course of my student life.

I should thank my friend Dr Ravindra Palavalli Nettimi for his special support, encouragement, and kind help during my studies.

Finally, none of this would be made possible without the endless love and unconditional supports of my family.

Publications

The following publications are part of the thesis.

Journal publications

[1] A. Rajabi, M. Eskandari, M. Jabbari Ghadi, S. Ghavidel, L. Li, J. Zhang, P. Siano, "A pattern recognition methodology for analyzing residential customers load data and targeting demand response applications," *Energy and Buildings*, vol. 203, p. 109455, 2019.

- **This paper is incorporated in Chapter 4.**

[2] A. Rajabi, M. Eskandari, M. Jabbari Ghadi, L. Li, J. Zhang, P. Siano, "A Comparative Study of Clustering Techniques for Electrical Load Pattern Segmentation," *Renewable and Sustainable Energy Reviews*, vol. 120, p.109628, 2020.

- **This paper is incorporated mainly in Chapter 3 and partially in Chapter 4.**

[3] A. Rajabi, M. Eskandari, L. Li, J. Zhang, "Trends in Applications of Load Data Clustering in Smart Grid Enviornment," (Under review, *IEEE Systems Journal*)

- **This paper is incorporated in Chapter 2.**

[4] A. Rajabi, L. Li, J. Zhang, K. Muttaqi, “A Feature-Based Data Mining Approach for Characterizing Residential Consumption Behavior in Smart Electricity Grids,” (Under review, *Energy Research and Social Sciences*)

- **This paper is incorporated in Chapter 5.**

[5] A. Rajabi, M. Jabbari Ghadi, S. Ghavidel, L. Li, J. Zhang, “A Clustering-Based Framework for Development of Time of Use Tariffs for Residential Electricity Customers,” (To be submitted)

- **This paper is incorporated in Chapter 6.**

Conference publications

[6] A. Rajabi, L. Li, J. Zhang, and J. Zhu, "Aggregation of small loads for demand response programs—Implementation and challenges: A review," in *Environment and Electrical Engineering and 2017 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe), 2017 IEEE International Conference on*, 2017, pp. 1-6

- **This paper is incorporated in Chapter 2.**

[7] A. Rajabi, L. Li, J. Zhang, J. Zhu, S. Ghavidel, and M. J. Ghadi, "A review on clustering of residential electricity customers and its applications," in *2017 20th International Conference on Electrical Machines and Systems (ICEMS)*, 2017, pp. 1-6.

- **This paper is incorporated in Chapter 2.**

Abstract

Smart grids are equipped with advanced smart metering infrastructures that enable the two-way communication between end-users and utilities and record the consumption data of electricity customers. The gathered smart meter data have opened up possibilities for analyzing the consumption behavior of customers and understanding the underlying factors affecting it. However, the dimensionality of recorded data necessitates the use of data analysis techniques to extract valuable information from the load profiles. In this regard, this thesis utilizes various methods to analyze the consumption and survey data of residential electricity customers.

In the first two chapters, the concepts of smart metering, residential customers' load patterns, and clustering of load data are extensively discussed, and a comprehensive discussion of the extant literature is presented.

Chapter 3 presents a comparative study of five main clustering approaches including K-means, fuzzy c-means, hierarchical, self-organizing map, and Gaussian mixture models for load pattern segmentation. Various parameters of each of these methods are explained in detail and their performances are compared using six cluster validity indexes. The obtained results are analyzed to find out the characteristic load shapes among the load curves of customers and to identify the main consumption patterns.

The problems of data deluge and residential DR establishment are addressed in Chapter 4 using a combination of symbolic aggregate approximation (SAX), as a suitable dimensionality reduction technique, a clustering algorithm, and the entropy concept. The use of SAX can assist in the clustering of residential load patterns, which usually display

high variability. Moreover, the results are utilized for ranking the customers based on their stability in usage patterns over time which is beneficial for different DR programs.

In Chapter 5, both the consumption data and survey data of residential electricity customers are used to find out the effects of the households' socio-demographic attributes and building characteristics on load patterns.

In Chapter 6, a combination of clustering algorithms and optimization models are used to design TOU tariffs for electricity customers. The problem is modeled as a mixed-integer linear programming problem with the objective of maximizing the profits of an electricity retailer that participates in different market settlements. The stochastic programming technique is used to address the uncertainties in future load and price.

Finally, in the last chapter, future directions in the analysis of smart meter data and the clustering of load patterns are briefly reported. Furthermore, the proposals for future work are elaborated in this chapter.

Keywords: Residential Load Data; Data Mining; Clustering; Demand Response; Data Size Reduction; Time of Use Tariffs

Contents

List of Tables

List of Figures

List of Abbreviations

AIC	Akaike's information criterion
AMI	Advanced Metering Infrastructure
AMR	Automatic Meter Reading
ARMA	Auto Regressive with Moving Average
BI	Business Intelligence
BIC	Bayesian Information Criterion
CER	Commission for Energy Regulation
CFL	Compact Fluorescent Lamp
CFSFDP	Fast Search and Find of Density Peaks
CVI	Clustering Validity Index
DBI	Davies-Bouldin Indicator
D^2R	Dynamic Demand Response
DFT	Discrete Fourier Transform
DLC	Direct Load Control
DR	Demand Response
DMS	Data Management System
DTW	Dynamic Time Warping
DWT	Discrete Wavelet Transform
EM	Expectation-Maximization
FCM	Fuzzy C-Means
GMM	Gaussian mixture model

HMM	Hidden Markov Model
HDF	Household and Dwelling Features
I/C	Interruptible/Curtailable load
ISODATA	Iterative Self-Organizing Data Analysis Technique Algorithm
KEPCO	Korea Electric Power Corporation
KNN	K-Nearest Neighbor
LR	Logistic Regression
MAPE	Mean Absolute Percentage Error
MIA	Mean Index Adequacy
MSE	Mean Square Error
NTL	Non-Technical Losses
OPF	Optimum-Path Forest
PAA	Piecewise Aggregate Approximation
PAM	Partitioning Around Medoids
PG&E	Pacific Gas and Energy Company
PCA	Principal Component Analysis
RLP	Representative Load Pattern
SAX	Symbolic Aggregate approXimation
SIL	Silhouette index
SOM	Self-Organizing Maps
SVM	Support Vector Machines
SVR	Support Vector Regression
TLP	Typical Load Profile
TOU	Time of Use
WCBCR	The ratio of within-cluster sum of squares to between-cluster variation

1 Introduction

1.1 Background and Research Question

The electricity industry has transformed dramatically in the last decade and embraced various novel technologies in different sectors. The significant increase in the installed renewable generation, the uptake of energy storage systems in large and small scales, the introduction of electric vehicles, and the vast deployment of smart meters are among the key technologies which directly affect electricity networks. There is an agreement that customers' load shapes, overall network demand, and the grid's control and operation will be greatly impacted by these advancements. However, these changes vary vastly based on different factors including but not limited to specific network structures, geographical locations, electricity tariffs, and government incentives. Specifically, residential sector, due to its underlying characteristics such as the sensitivity to costs and incentives and the comfort considerations, will exhibit a much more diverse behaviour by adopting such technologies. Therefore, the research studies usually consider various scenarios to predict the future domestic load shapes and the distribution network demand.

Among the mentioned technologies, smart meters will play an important role by providing electrical measurements of the network at the customer-side, thereby, increasing the visibility and controllability of the network. Enhancement of the power networks using advanced metering infrastructure (AMI), measuring equipment, and smart devices is expected to restructure the existing power grids into a cyber-physical system. Such a system is not only able to carry power flow but can also transmit data for advanced measurement

and control applications. The backbone of this cyber-physical system are smart meters and other sensory devices. Smart meters are specified with the sophisticated measurement, control and communication capabilities that they possess. Compared to a conventional energy meter, a smart meter includes measurement and calculation hardware, software, and communication capabilities that measures the energy consumption of a consumer and provides added information to the utility company [1], [2]. In the future distribution networks, smart meters will be integrated with different home, buildings, neighbourhood and wide area networks under different communication protocols [3]. They will be in constant communication with distribution (data) management system (DMS) for providing online information and receiving commands.

It is projected that the total number of installed smart meters will reach 780 million in 2020 [4]. As some critical studies pointed out [5], despite the ongoing rollouts, many utilities are still unclear about the optimal route to extracting value from these large investments. In North America, the main priorities are to use smart metering information as a means to support outage management and increase grid reliability. On the other hand, European utilities are more focused on consumer-related capabilities. Therefore, from the analytics point of view, the smart metering data is still mostly an underutilized area of value for existing deployments. However, it is becoming recognized as a strategic next step for many utilities. Another problem arises from the dimensionality of the data. The growing prevalence of installed SMs has resulted in the collection of consumption data at unprecedented scales [6], [7] as the energy consumptions of customers can be recorded in intervals of an hour or less. The rate and volume of these data not only outpace the capabilities of traditional systems of companies but also require redesigning the business

models of them [8]. As the deployment of smart meters is increasing, the main question is how to utilize such a wealth of hourly or half-hourly measured data to gain benefits for various stakeholders in power systems.

The overwhelming amounts of data gathered by smart meters call for a powerful and cost-effective information management system for data processing, analysis, and storage and necessitate the use of proper data analytics tools and techniques to analyze the load data [9]. The use of data mining techniques offers a variety of potentials within the power systems [10]. It provides unique opportunities for academia to study the consumption behavior of residential customers and understand its underlying affecting factors. On the other hand, it opens up possibilities for utilities to offer new services such as customized demand response (DR) programs or tariff structures to their electricity customers.

In spite of these potential advantages, in most companies, electrical engineers are not familiar with the data mining concepts. It is a serious obstacle to successfully reap the potential benefits of smart meter data. Therefore, one of the primary goals of this thesis is to utilize a cross-disciplinary approach spanning engineering and data science to present a systematic study for the use of smart meter data. The concepts and studies in this thesis are mostly oriented toward the practical applications and the results can guide the energy sector for implementations of the techniques in real-world scenarios.

In this thesis, we use different data mining techniques including the clustering and classification algorithms. Clustering is a well-known unsupervised data mining technique for segmentation of a data set by assigning its objects to a set of clusters [11]. It has numerous applications in different fields such as market segmentation analysis, biology, and social network studies. Classification is a supervised data mining technique that deals with

the categorical outputs or discrete class labels [11]. Classification is common in almost every aspect of everyday life, for example, to assign customers, stores, documents, emails, or any other type of instances into a set of known classes.

Along with the aforementioned data mining techniques, dimensionality reduction methods, statistical tests, as well as optimization methods are also used in different parts of this research. These algorithms are employed to cluster load data and analyze residential consumption patterns, propose improvements for residential DR programs, investigate the relationship between the household characteristics and consumption patterns, and design customized time of use (TOU) tariffs.

1.2 Research Objectives and Scope

The aims of this research can be summarized as:

- Evaluating the current status of residential DR programs and the challenges they face.
- Providing a detailed study of the benefits and applications of data mining methods, especially clustering of load data, for power systems.
- Presenting a comparative study of different clustering algorithms for load data clustering.
- Applying proper data size reduction techniques to deal with the dimensionality of smart meter data as well as difficulties of clustering due to variability of residential consumption patterns and to decide on the customers for various DR programs.

- Identifying the relationship between the consumption patterns of customers with the building features and household characteristics.
- Calculating customized TOU tariffs based on customer clustering.

1.3 Dataset

In this research, the initial aim was to use the data of the Australian project, "Smart Grid Smart City", which was completed in 2014. This dataset includes a large amount of information from the Australian electricity customers. Unfortunately, due to privacy issues, the website which was dedicated to this project was completely shut down. In this regard, another suitable dataset, which has been used in various academic publications, was selected for the case studies. This dataset contains the consumption as well as survey data of a large number of electricity customers and is collected as a part of the smart metering trial project that was carried out by Commission for Energy Regulation (CER) in Ireland [12]. In the following, the characteristics of this dataset are briefly described.

1.3.1 Consumption dataset

This data set contains the half-hourly consumption readings of 6445 customers consisted of 4225 residential consumers, 485 small-to-medium enterprises, and 1735 other customers for a period of one and half years, started from 14 July 2009 (day 194) and finished on 31 December 2010 (day 730). The original data is stored in 6 big text files in which each data instance is identified by three values: meter ID, a five digit code identifying the day and time, and the electricity consumption during 30 minute interval (in kWh) for the specified meter ID and time.

In this research, the focus has been on the residential customers and hence, only the data of these customers are used. It was observed that only 3639 customers have complete data (25730 recordings). For 539 users, a day of data was missing. For these customers, the missing day was identified and its consumption values were correspondingly calculated as the average of the consumption values of the day before and the day after it. In this way, the total number of customers with available data added up to 4178.

Furthermore, the correction of daylight saving time changes in spring and autumn was carried out. Two days (25 October 2009 and 31 October 2010) had two more recordings and one day (28 March 2010) had two fewer recordings and the data were modified accordingly.

The dataset is also divided into subsets based on the seasons. The seasons were chosen based on the meteorological seasons in the northern hemisphere. For example, based on this convention, spring starts on March 1 and ends on May 31.

Eventually, a year of data starting from 1 December 2009 and ending on 30 November 2010, which comprises of four seasons, was selected for the studies.

1.3.2 Survey dataset

The CER data set includes a survey data which provides significant information about the characteristics of those households that had participated in the trial. The responses to the questions are tabulated in a spreadsheet file. For the residential customers, the file has 4234 rows which specify the users with their unique ID and 143 columns which are dedicated to the questions. It should be noted that the responses for some of the customers are not recorded in the survey data. The total number of customers that are common in both datasets

account for 3148, whose consumption and survey data are accordingly used for relevant studies.

1.3.2.1 Survey structure

All the responses are saved as numbers. For example, the question about type of the home can take a value from 1 to 5 which defines the type as apartment, semi-detached house, detached house, terraced house, and bungalow respectively.

The proper code was written to extract the data. First of all, the response to some of questions should be inferred from different columns of the spreadsheet file. For instance, the question asking about the income of a household is like this:

Can you state which of the following broad categories best represents the yearly household income before tax?

Category	Recorded answer
Less than 15,000 Euros	1
15,000 to 30,000 Euros	2
30,000 to 50,000 Euros	3
50,000 to 75,000 Euros	4
75,000 or more Euros	5

Is that figure:

Category	Recorded answer
Per week	1
Per month	2
Per year	3

Can I just double check is that figure:

Category	Recorded answer
Before tax	1

Or after tax	2

So, the final response should be found out based on all the responses to these three sub-questions.

From this large number of questions the most relevant ones were selected which fall under one of these categories: the dwelling characteristics, household characteristics (socio-demographic data), appliances, and attitudes and knowledge toward energy saving.

1.4 Software

R software which is a well-known data analysis software and Matlab packages are used for the extraction of data, analyses, and depicting the results. In addition, General Algebraic Modeling System (GAMS) is used for solving the optimization problems.

Various packages in R including the following ones are used for the studies:

- *xlsx* package: for data extractions, data reading, and data writing
- *Cluster* package: especially for K-means clustering (Chapters 5 and 6)
- *biganalytics* package: for applying K-means for large matrixes (over 1,000,000 daily load shapes) (Chapter 4)
- *Stats* package: especially for hierarchical clustering (Chapter 3)
- *kohonen* package: for clustering with self-organizing map (SOM) and visualizing the results (Chapter 3)
- *NbClust* package: for calculating cluster validity indexes (Chapter 5)
- *nnet* and *broom* packages: for classification analysis (Chapter 5)
- *MASS* package: for Chi-squared test of independence (Chapter 5)

- *GoodmanKruskal* package: for Goodman and Kruskal's tau measure (Chapter 5)
- *Corrplot* package: for visualizing the results of correlation matrix (Chapter 5)
- *ggplot2* package: for plotting some of the figures.

Various M-file programs were written in Matlab environment, especially for developing the codes for:

- Clustering analysis including K-means, fuzzy c-means (FCM), hierarchical, and Gaussian mixture model (GMM) (Chapter 3)
- Cluster validity indexes (Chapter 3)
- Symbolic aggregate approximation (SAX) technique (Chapter 4)
- Data size reduction using principal component analysis (PCA) (Chapter 4)
- Entropy calculation (Chapter 4)
- Depicting the results

Furthermore, in Chapter 6, the large optimization problem is modeled in the GAMS environment.

The high-performance computing resources of Advanced Research Computing Laboratory (ARCLab) at University of Technology Sydney are used for studies in Chapter 4, especially for analyzing the large matrixes. Also, the server facilities and optimization solvers of Network-Enabled Optimization System (NEOS) are used for solving the optimization problem.

1.5 Contributions and Organization of the Thesis

The main contributions of this research and the structure of the rest of the thesis are briefly summarized in the following:

1. The underlying concepts of residential consumption patterns, residential DR, and residential load data clustering are thoroughly illustrated in Chapter 2. In this regard, the affecting factors on consumption patterns are summarized and the high variability of residential consumption, as a major problem when clustering the raw data, is explained. Accordingly, the use of representative load patterns (RLPs), as the most common practice in the literature for clustering the load patterns, is described. In addition, the various challenges of residential DR, which are neglected in most of the studies, are clarified. This chapter also explores in detail the integration of clustering techniques in business intelligence (BI) systems of utilities and reviews and demonstrates its benefits for power systems. In this respect, the applications of clustering for load forecasting, defining DR programs, design of tariff structures, classification of new electricity customers, detection of non-technical losses, and defining new class load profiles are explained in detail.
2. In Chapter 3, the major clustering approaches for load pattern segmentation are methodically explored and their applications are analyzed. Five major clustering algorithms including K-means, fuzzy c-means, hierarchical, self-organizing maps (SOM), and probabilistic and generative models and their relevant parameters are investigated. These clustering algorithms are firstly compared by using the daily data

of one customer and subsequently, the application of them for weekdays and weekend RLPs of more than 4000 customers are investigated.

3. In Chapter 4, a modified symbolic aggregate approximation (SAX) technique is utilized to characterize the variable time series consumption data of residential customers. The parameters of the method including the time axis partitioning and the number of alphabets are carefully investigated. By applying the SAX and a hierarchical clustering algorithm, the characteristic consumption patterns of customers are extracted. In addition, the entropy concept from information theory is utilized to compare the customers based on their stability over time which can help in establishing DR programs.

4. In Chapter 5, using a clustering and classification procedure, the relationship between the consumption data and household attributes are specified. Firstly, customers are clustered using their consumption data. The questions in survey data are transformed into appropriate variables and the Chi-squared test of independence and Goodman and Kruskal's tau measure are used to find the correlation and association among the variables. A multinomial logistic regression (MLR) technique is utilized to link the household characteristics with the cluster memberships.

5. A combination of clustering and optimization algorithms is employed in Chapter 6 to design customized tariffs for electricity customers. The proposed formulation uses the models of day-ahead markets, forward contracts, and storage units as well as the concepts of customer clustering, stochastic programming and risk to model the objective function of an electricity retailer. The problem is further linearized and is

represented through mixed-integer linear programming which is solved using off-the-shelf optimization solvers.

6. Finally, the possible future trends in load data clustering and the suggestions for future works are elaborated in Chapter 7. In this regard, the innovative methods such as distributed, online, and time series clustering of load data along with the applications of deep learning methods for load data analysis are briefly introduced to guide the interested readers for the future work and practical applications. The main directions for continuing the studies of the current research are also explained in this chapter.

2 Background and literature review

This chapter provides an introduction about the concepts of smart metering, residential load pattern characteristics, and clustering which is necessary for understanding the presented materials in the following chapters. Furthermore, it comprehensively discusses the potential advantages and applications of clustering techniques for power systems.

2.1 Smart Metering

Smart meters are considered as the third generation of meters after the electro-mechanical and electronic meters. They are specified with the sophisticated measurement, control and communication capabilities that they possess [1], [3]. Generally, the functionalities of a smart meter depend on the minimum functionality requirements which are set by local regulatory authorities. It can include a wide range of tasks including measurements, communication, monitoring, and execution of received commands from the control center. As shown in Fig. 2.1, in the future distribution networks, smart meters will be integrated with different home, buildings, neighborhood and wide area networks under different communication protocols and will communicate with data management systems for providing online information or receiving commands [13], [14].

The full-scale rollouts of smart meters will cost billions of dollars for countries which in some cases hinders the utilization of smart meter projects. Nevertheless, the potential benefits of the smart meters for the utilities, network, and customers have encouraged authorities to continue their integration into distribution grids [15].

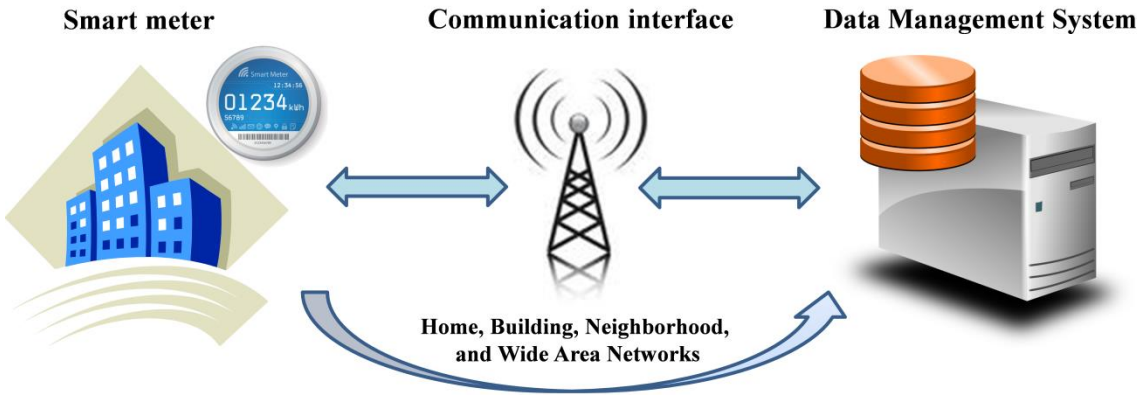


Fig. 2.1 Structure of smart metering system

In recent years, local authorities in different countries have issued various guidelines, directives, and standards about smart metering. They provide a perspective about the future expansion of smart meters in power networks and set policies for proper operation and management of them. It is projected that the total number of installed smart meters will reach 780 million in 2020 which is mostly driven by China as well as North America. Table 2-1 summarizes the trends of installations in some parts of the world [3], [4], [5, 16-18].

Most of these deployments are driven by government mandates as in the cases in China and the majority of European countries. The market-driven approach is accepted as the main force by other countries such as Germany and Japan. In some other places, a mixture of both approaches has been tried. For example, in Australia, the smart meter rollout was mandated in State of Victoria, which led to the replacement of 2.6 million of the traditional meters with the new meters [19]. The negative public view of this compulsory approach and technical shortcomings caused the other states to follow a market-driven approach [16], [5], [20]. Table 2-1 summarizes the trends of installations of smart meters in different parts of the world [3], [4], [5, 16-18].

Table 2-1 Smart meters' projected rollouts globally

Country	Trend
Australia	2.6 million SMs in 2014 and additional 5 million by 2020.
China	Plan for deployment of 300 million smart meters by 2015 and up to 380 million meters by 2020.
Japan	Near 60 million smart meters will be deployed by 2020.
EU	Enacting a mandate that requires utilities in all of its member states to provide smart meters to 80 per cent of their electricity consumers by 2020.
France	Plan for installation of a total of 35 million smart meters by 2020.
Italy	Almost all of 36 million customers had smart meters installed by 2011.
Spain	All meters for supplies of up to 15 kW have to be replaced by new meters by the end of 2018.
UK	Plan for replacement of all 30 million traditional meters in homes and small businesses with the SMs between 2015 and 2020.
USA	Near 52 million in 2014 and an estimated amount of 130 million installed smart meters by 2020.

While the number of smart meters is increasing rapidly in power networks, various regulatory concerns still exist regarding minimum functionalities of smart meters, communication protocols [3], [21], and privacy and security issues [12], [22]. The other main challenge arises from the volume of AMI data. The data sampling resolution of smart meters is usually set to 15 minutes, 30 minutes, or 1 hour. Sampling, reporting, and processing energy consumption at such rates, demands more complicated hardware and software capabilities. This necessitates the use of data mining tools such as clustering to

analyze the datasets and extract value from the large amount of recorded data. This aspect will be discussed in more detail in this thesis.

2.2 Residential Load Shape Characteristics

The main characteristic of residential customers' load patterns is that they usually show high variability which makes them different from the industrial and commercial load shapes. Consequently, two major problems should be tackled when studying residential load curves. The first challenge is that electricity consumption patterns of two residents could be significantly different. On the other hand, daily load patterns can be different even for the same customer [22].

Different factors are responsible for the energy use of a dwelling. In general, they can be divided into three main categories [23-25]: 1) physical determinants which are based on the building design and environmental factors, 2) time of day, week or year, and 3) behavioral determinants which are habit driven. The first factor shows the effect of building design like dwelling type, insulation and heating devices, and local climate factors such as temperature, humidity or solar radiation on the daily energy use of a household. The second factor demonstrates the relationship between the diurnal, intra-day and seasonal variations of electricity demand and the time of the week, season and year [26]. Typically a customer shows different energy usage patterns over the weekdays, weekend and holidays, and on different seasons. For example, it has been noted that there are some unusual peaks occurring on specific days such as Christmas and New Year's Eve/day and there are increases in demand during the Easter holidays [27]. The third determinant, which is related to individual consumer's behavior and their everyday practices, accounts for a substantial

proportion of household energy consumption [28]. Consequently, the load curves reflect the energy usage habits of customers, for example, the pattern of using different devices [29].

Fig. 2.2 displays the typical energy curve of a household during a week. Fig. 2.3 shows the variation of consumption pattern of this dwelling for all the days in different seasons.

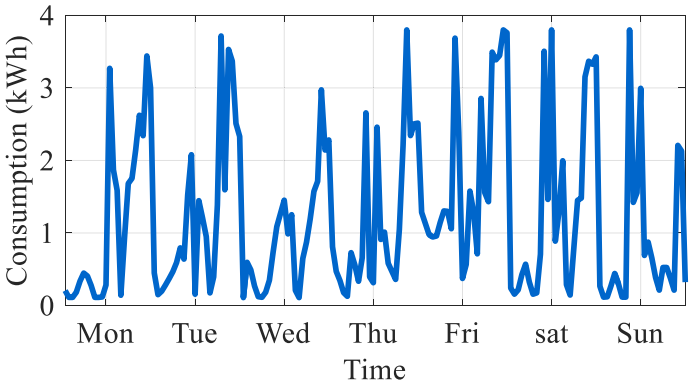


Fig. 2.2 Variability of the consumption of a household over a week

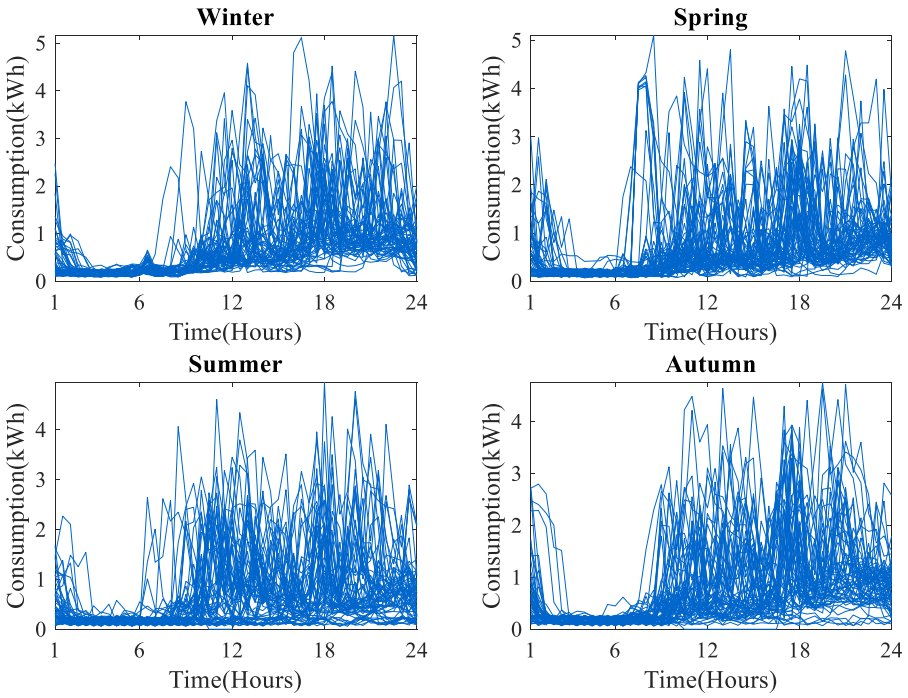


Fig. 2.3. Effect of seasonality on the consumption of a household

As these figures demonstrate, the consumption patterns of a residential customer can vary on a daily basis and can be largely affected by seasonal changes. Such behavior makes the analysis of energy data of residential customers more difficult. Especially, it imposes challenges to the clustering of customers into a certain number of classes since the daily load curves of a customer might assign to many different clusters.

Instead of dealing with many daily curves which may considerably differ from each other, a representative load pattern (RLP) can be defined and assigned to each customer. The common practice in most of the studies in the power system literature is to first define an RLP for each customer based on the whole recorded data and then apply clustering techniques on these representative curves. However, as explained in Chapter 4, there are alternative approaches that are able to reduce the dimensionality of data while preserving the temporal characteristics of load data.

Stages of constructing an RLP are visualized in Fig. 9. Initially, a set of different loading conditions are defined based on the user preferences, climate conditions, and other affecting parameters. For example, a good choice is to divide the year into 4 or 5 seasons [30], [31] and then define for each season two or three types of days like weekdays and weekends. By considering 4 seasons and two types of days, 8 different data sets will be available that can be assessed separately. This partitioning among different time periods improves the clustering results since consumption patterns of a customer in different loading conditions are usually different from each other. The daily load data of each customer in a specific loading condition can be organized to represent the customer's consumption by means of just one load pattern [32]. To this end, the daily load patterns are combined instant-by-instant based on a statistical criterion like mean or median to create a

representative load diagram. Finally, the normalized representative load pattern of each customer can be made by normalizing the original representative load diagram with respect to a reference power which is usually assumed as the maximum value of the diagram [33]. All values of this normalized curve lie in the [0, l] range. This allows the clustering of the customers with similar load shapes into a class, regardless of the actual quantities of consumptions.

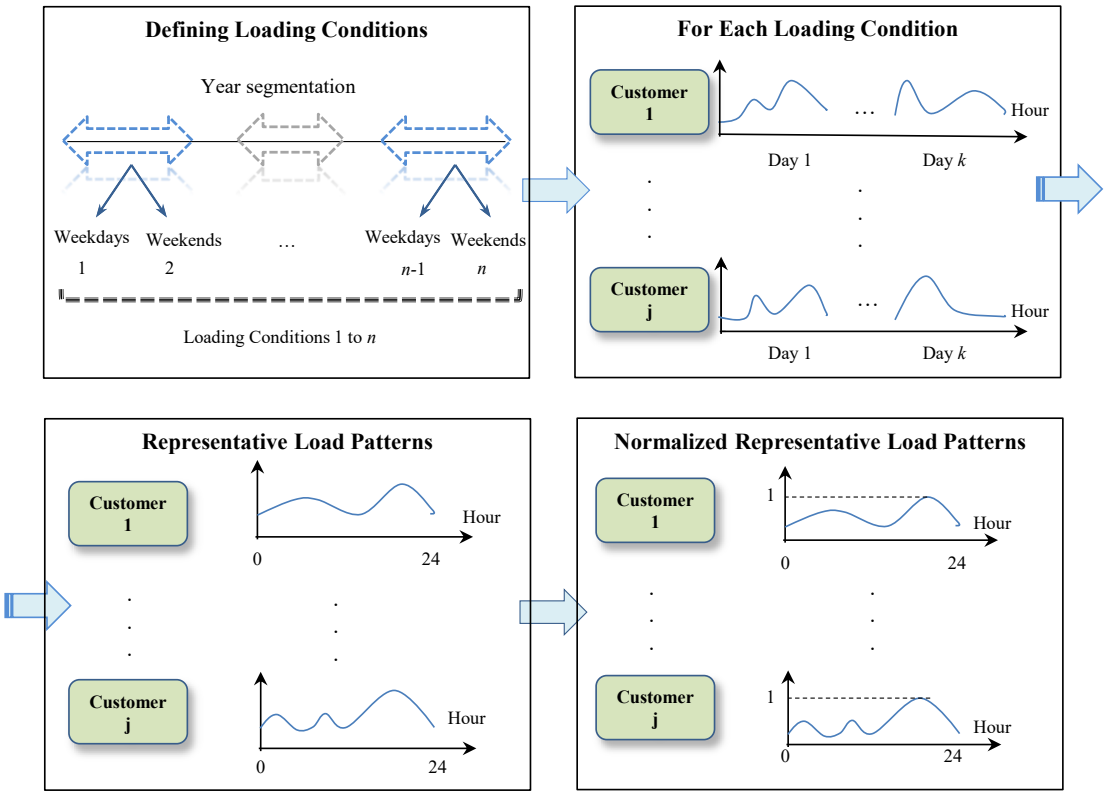


Fig. 2.4 Construction of representative load patterns for electricity customers

2.3 Effects of New Technologies on Load Patterns

2.3.1 Renewable energy status

The share of renewable energy in power grids has been constantly increasing in the last two decades. Just in 2019, 200 GW of renewable generation is added to the electricity grids, increasing the total installed capacity to 2600 GW [34]. From 2012, the annual additions of renewable resources have been higher than the combined additions of conventional fuel-based and nuclear power plants. Table 2-2 and Fig. 2.5 report the latest data on the total power capacity of each renewable technology worldwide [34] [35].

Table 2-2 Renewable energy capacity

Technology	Total Installed Capacity (GW)	
	2018	2019
Hydropower	1135	1150
Wind Power	591	651
Solar PV	512	627
Bio-power	131	139
Geothermal power	13.2	13.9
Concentrating Solar thermal power	5.6	6.2
Ocean power	0.5	0.5

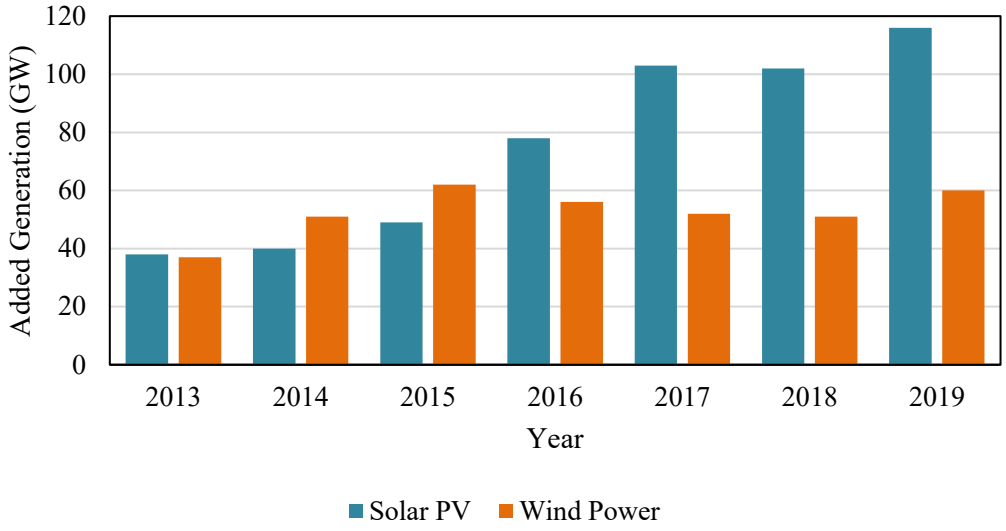


Fig. 2.5 Annual additions of solar PV and wind power

As can be seen, most of this renewable generation is produced by utility-scale plants, although the share of rooftop solar photovoltaic (PV) panels has been growing gradually. Solar PV is the main renewable generation available for use by residential dwellings. At the end of 2019, the total renewable production and PV generation accounted for almost 27 and 3 per cent of global electricity generation, respectively. In the 2018-2019 period, the solar PV capacity had the highest increase among all renewables that can be attributed to different influencing factors such as the price drop as well as governments' incentives. Nevertheless, different issues need to be addressed in order for solar PV to become a main source in the electricity networks including defining proper regulatory frameworks, introducing support schemes, and resolving technical barriers. Government policies such as feed-in premiums are still the main driver of the global PV market [36].

2.3.2 Storage systems

Electricity storage systems (ESSs) can benefit the electricity grids in various ways. They include a diverse range of technologies with different characteristics and are generally classified into: i) mechanical such as pumped hydro and compressed air, ii) electrochemical including various kinds of batteries, iii) chemical such as fuel cells, iv) electrical including supercapacitors and superconducting magnetic coils, and v) thermal such as heat storage systems [37].

ESSs have been deployed in large scales for grid applications such as backup power and frequency regulation for many decades. In recent years, there has been an increasing interest in ESS utilisation in smaller scales for example, for microgrids and commercial and residential buildings. This interest stems from the advantages for customers including the short-term power supply in case of system failure, storage of renewable generation surplus, and arbitrage of electricity based on time of use (TOU) tariffs.

Batteries are seen as the most common ESS for future residential applications. Different manufacturers now offer batteries for dwellings. Lithium-ion batteries are the dominant battery technology for households with desirable features such as relatively high energy and power densities and high efficiency. The price of batteries has decreased rapidly as shown in Fig. 2.6 [38] [39] due to technology advancements and large-scale manufacturing.

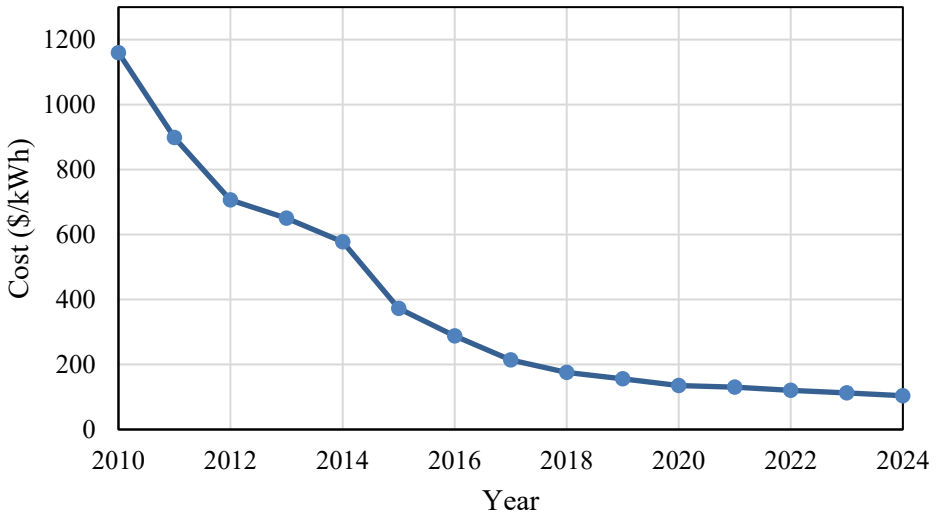


Fig. 2.6 Cost of battery in different years

Unlike the solar PV, the battery storage installation around the world has shown mixed results in the past years as it slowed down in some markets and increased significantly in some others [34]. In Australia, 233 MWh of new home batteries were added in 2019, expanding the total capacity of home batteries to an estimated amount of 1 GWh [40]. On the other hand, countries such as Korea and China experienced a decrease in battery installation growth compared with the previous years.

2.3.3 Electric vehicles

Electric vehicles (EVs) are seen as a promising technology for future green cities. The EV market is expanding vastly in different parts of the world owing to the sharp decrease in prices and the availability of a variety of models to the customers. There were around 7.2 million EVs on the world's roads in 2019, showing an increase of 2 million units from 2018 [34].

While the share of EVs in the global stock is still low (around 0.6%), the fast adoption of EVs in upcoming years can cause serious challenges for electricity networks. Lithium-ion batteries represent the most used technology in EVs. Currently, the average battery capacity in EVs is around 45 kWh with some models offering a capacity as high as 100 kWh (for instance, Tesla X) or even higher. This large capacity can put pressure on distribution networks and cause congestions in some parts. At the same time, with the proper management and aggregation, the flexibility of EVs in operating in different modes including the grid-to-vehicle (G2V) and vehicle-to-grid (V2G) can be harnessed to support the electricity networks.

2.3.4 Impact of newer technologies on load profiles

The introduction of newer technologies will affect different parts of electricity grids. These impacts have been studied vastly by academia and industry. On the grid-scale, these effects will be evident on the annual energy consumption, daily and seasonal peaks of the network, and aggregated load shape of the grid. They may also disrupt the normal operation of distribution and transmission networks for example, by interfering the protection coordination and causing congestion, voltage rise, transformer overloading, and so on. On the household level, the impacts will be mainly on the load shapes of the customers and their energy costs. Here, we limit the discussion to the effects on residential users.

Normally, there is a need to define various scenarios in order to project into the future and predict the probable changes in domestic energy consumption. Typical scenarios should involve the level of PV, storage, and EV uptake, the inclusion of an individual technology or a combination of technologies, and the customers' preferences in using these devices, for

instance, the EV driving and charging pattern and the attitude towards the cost saving. Besides, other external factors such as the geographical location influence the conclusions. For example, the PV generation level is directly affected by the location and climate. Therefore, not surprisingly, different studies can reach different outcomes based on the specific inputs and scenarios.

A) Effects of solar PV without the storage system

Using PV systems without the storage can decrease the daytime load of the dwelling but does not affect the evening peak. As will be explained more in the next chapters, different households exhibit different load patterns and can be classified into certain clusters based on their consumption habits. Therefore, households with a peak happening at daytime can benefit most from the PV system. Also, as indicated in different studies [41] [42], the PV impact is directly related to seasonal changes. For example, some studies showed that even with high penetration levels, the PV may not make any contribution to the network daytime peak during the winter [41].

Besides the technical aspects of solar PV integration, another important question is the customers' willingness for its deployment. The cost/benefit analysis is needed when predicting the implementation of PV systems by customers. It is shown that the cost-effectiveness of PV systems for users is related to different factors such as the customer's electricity usage and energy consumption behavior, PV capacity, dwelling's location, and electricity tariffs [43]. For instance, a study performed for New Zealand network has modelled different combinations of these factors and has shown that (in the defined circumstances) only a small fraction of these cases can be cost-effective [44]. This is due to the high initial costs and the low tariffs for exporting the excess PV energy to the grid. By

changing these two factors in the next years, the authors have demonstrated that PV uptake can increase to 40%. As mentioned, these studies are completely dependent on the specific scenario and differ among different countries. Overall, it can be concluded that in many parts of the world, the government incentives are still necessary since the upfront costs and relatively long payback periods might hinder the PV implementation by households [34] [35].

B) Effects of electric vehicle

EVs have the potential to cause great changes in demand. Modelling the impact of EVs on load profiles is difficult since it is largely dependent on the customers' behavior. Different approaches are proposed for EV modelling. Three key factors are important when considering the effect of EV on load shapes including the charging location, the charging need, and the charging moment [45]. The charging location can be the residential buildings, workplaces, shopping centers, or specific charging stations. The charging need refers to the amount of energy that EV needs when it is connected to the grid which depends on the used energy during the driving. Modeling this aspect has been done in different ways, for example, using probabilistic distributions. The charging moment can be modelled as either the start time of the charging or the charging duration i.e. the period that EV is connected to the grid. Again, various methods are proposed for modeling this EV aspect. For instance, the EV can be charged immediately after the journey which might coincide with the customer/system peak or be charged at off-peak periods (lower electricity prices).

The EVs can also act as unidirectional or bidirectional based on their ability to feed the energy back into the grid. Additionally, many proposals are suggested for the aggregation of EVs for grid-support applications such as the frequency regulation or demand response

programs. Therefore, customers may participate in such programs which in turn will affect their charging patterns.

Overall, the effect of EVs on the network and an individual household depends on the mentioned factors. For the grid, there is a possibility of large demand increase, power losses, overload, etc. if the EVs charge/discharge patterns remain uncontrolled [43].

C) Effects of storage systems combined with solar PV

Home ESSs are mostly used in combination with solar PVs. They have this potential to significantly alter the household's load curve by storing the energy provided by solar PV and the grid and releasing this energy when it is needed. Selecting the best sizes of Solar PV and ESS for a household depends on the household's socio-demographic features (such as young adults, older family, retired, big energy user, etc.) as well as their energy usage habits. For example, an optimal choice for a family with 20 kWh average usage per day can be a solar PV and a battery with the sizes of 5 kW and 3.5 kWh, respectively [46].

The charge/discharge strategies of batteries can be done in different ways based on the defined goals, although, the control analysis and predictions are less complicated compared with EVs. Three main strategies can be defined with these goals: i) to smooth the household's load, ii) to achieve the highest profit, and iii) a combination of the first two strategies. In the first case, the battery acts with the purpose of flattening the load curve by discharging in household' high consumption periods (above a certain limit) and charging in low load times (below the certain limit). While this strategy partially benefits the customer, it is advantageous for the network as well by reduction of peak load. The second approach maximizes the customers' profit by considering electricity prices. The hybrid strategy is a

combination of these two approaches in which the ESS discharges when the dwelling load is above the limit and charges during low-price periods.

In addition to control strategies applied within a home, there have been initiatives for the utilization of household batteries in an aggregated way. For instance, a prototype of a virtual power plant consisting of 1000 residential and business premises was established by AGL company in South Australia [47]. In this project, the solar battery ESSs are managed on a cloud-based system and form a 5 MW power plant. In such schemes, the charge/discharge of batteries can be done based on the system needs and/or for the maximization of the overall profits of participants.

The use of home ESSs even with solar PV is not still considered a cost-effective option in many parts of the world due to high prices of batteries. For example, in Australia, the total costs of the battery system, inverter, supporting hardware, and installation is generally around 1000$ per kWh. Ref. [46] presents a cost/benefit analysis for the utilization of batteries in the Australian framework. It categorizes households into three different groups based on their consumption levels and analyses the cost-effectiveness of the combined solar/battery system for different flat and TOU tariff structures currently available in different states in Australia (48 scenarios in total). As per 2020, this analysis suggests that, except for one special case in Western Australia, none of the other scenarios makes the financial sense since the payback period is almost the same as the lifetime of the battery.

It should be noted that the other storage systems such as the heat storage have been also implemented in the residential premises. Some of these technologies offer a proper solution for storing excess PV energy. However, they are less controllable compared with the batteries.

The methods that are suggested in the next chapters could be applied for both the households equipped with the discussed technologies and the dwellings without them. In the bigger picture, these technologies will be used along with the smart meter data and AMI facilities to achieve smarter grids. Such structure will use renewable generations at the household level, residential storage systems, EVs flexibility, demand response programs, and the two-way communication using AMI to flatten the load curve of the system and achieve greener networks with lower emissions and higher efficiencies.

2.4 Load Pattern Clustering

2.4.1 Clustering concepts

Clustering is an unsupervised data mining technique that enables the determination of intrinsic patterns in data sets. The main aim of clustering is to partition data instances (objects) of a data set into a number of groups (called clusters) which are as similar as possible. Objects belonging to a cluster are more similar to each other than to those in other clusters. Therefore, the ultimate goal is to achieve high intra-cluster similarity and low inter-cluster similarity.

Expressing mathematically, a data set S which has n records (observations) can be partitioned into a set of K clusters $C_1, C_2, \dots, C_K$ that do not intersect (however, this assumption is sometimes violated when the soft clustering is applied) and the union of them is equal to the full data set as shown in (2.1):

$$S = \cup_{i=1}^{K} C_i \quad and \quad C_i \cap C_j = \emptyset \ for \ i \neq j \tag{2.1}$$

2.4.2 History of electricity customer clustering

In the power system domain, utilities and system operators are interested to classify the electricity users into distinct groups as it offers advantages in decision making and control of power network. However, this trend was naturally limited in power system studies as the system operators and researchers did not have access to the fine grained consumptions of the customers. Before the widespread availability of AMI data, little information about each household's consumption or energy use habits was available. The monthly usage of each household and some fixed information such as voltage level and nominal demand were the main sources of information for categorizing the households. On the other hand, utilities and researchers also conducted in-home surveys in order to evaluate the effect of various variables on consumption patterns. These variables could be categorized under one of these elements: dwelling characteristics, demographics and socio-economic factors, habits of energy use such as consumption timing, attitudes toward energy use like level of concern toward energy conservation, knowledge about electricity consumption, and energy efficiency goals. Table 2-3 summarizes these elements which are usually addressed in the questionaries and the most important variables for each of them [48-52].

Based on those available data, it was a common practice among distribution companies to define a set of classes and assign each customer to a specific class. For example, Pacific Gas and Energy Company (PG&E) segmented its residential customers into eleven clusters using a broad range of attitudinal and demographic variables [53]. In another attempt, 46 different customer class load profiles were defined by Finnish Electricity Association for categorizing customers in residential, agricultural, industrial, and services sectors [23], [25]. 18 of these load curves are for the residential users and each customer is attached to a

specific load curve which is used as the base for billing and electricity distribution planning [23], [25].

Such approaches were inherently inaccurate as they did not have access to the real consumption data of customers. Installation of smart meters has fundamentally changed this situation. Nowadays, the fine-grained measurements are available in a large scale for tens of thousands to millions of users and moreover, they are accessible for successive years. As a result, the customer categorization can be achieved by implementing appropriate clustering techniques which are applied on the load data of the customers.

Table 2-3 The most important variables of in-home surveys

Elements	**Variables**
Demographics	Number of occupants / Age of dwellers / Presence of children or elderly persons / Income / Education / Employment situation of residents / Information and communication technology
Building design (Dwelling Characteristics)	Dwelling type / Size of home (number of rooms) / Age of home / Insulation / Heating devices / Home appliances / Presence and number of compact fluorescent lamps (CFLs) installed
Attitude	Thinking about ways to save energy / Thinking about ways to control energy cost / Level of concern about energy or conservation / Strength of link between efficiency and environment / Effect of high energy use on global warming / Thinking about local energy issues / Paying more for environmentally friendly products / Desired comfort levels

Habits	Consumption timing (day, night, etc.) / Reducing temperature in unused rooms / Washing machine: full loads or using cold water / Full dishwasher or air dry dishes / Turning off or using minimum lights / Turning off water heater if away / Dressing warmer at cold weather or keeping thermostat at lower degrees / Leaving windows open for ventilation in winter / Regularly reviewing energy use
Knowledge	Aware of ways to save energy / Aware of energy efficiency and DR programs / Know where to get renewables information
Energy efficiency goals	Replace major appliances with Energy Star or energy efficient ones / Maintenance of devices for improving efficiency / Set thermostats for efficiency / Replace bulb or fixtures with energy efficient ones / Install insulation or windows

2.4.3 Stages of load pattern clustering

Stages of electricity customers' clustering is summarized and depicted in Fig. 2.7. These stages are as follows:

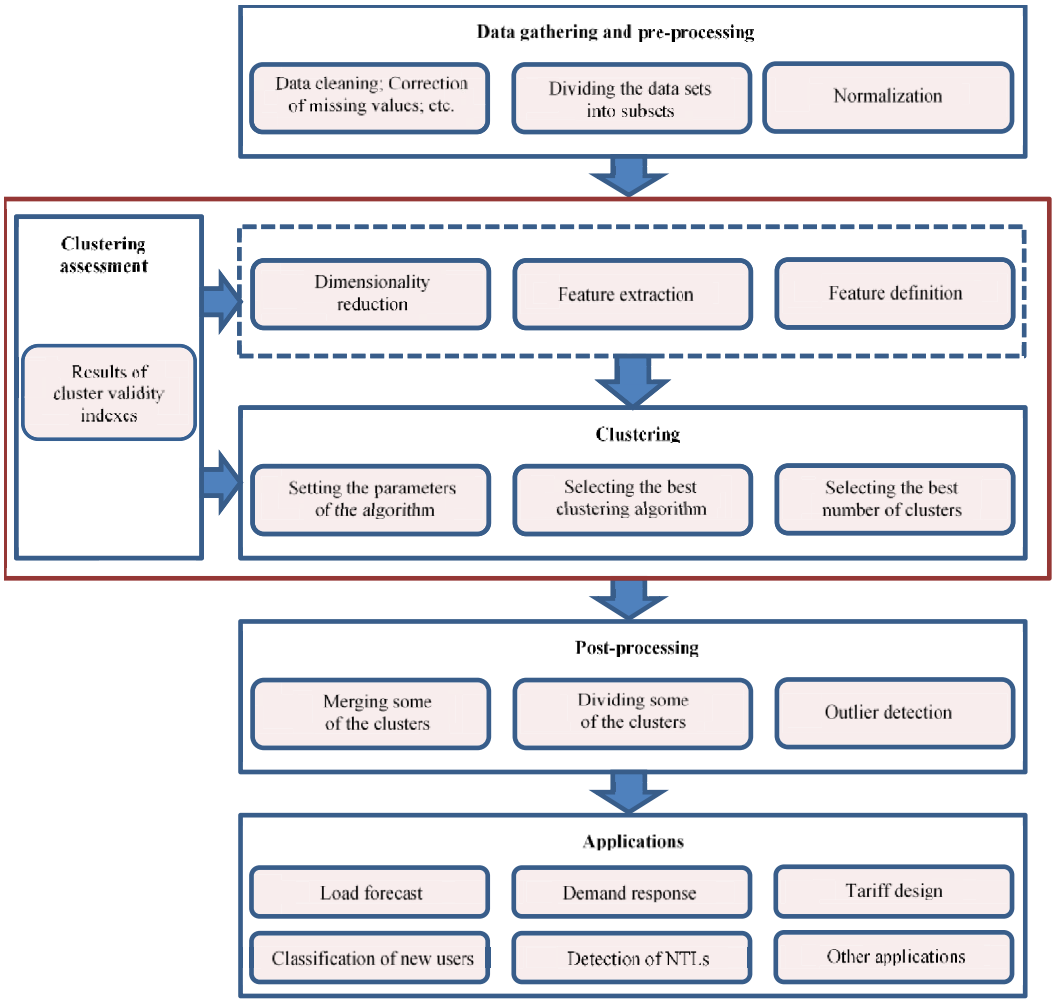


Fig. 2.7 Stages of load pattern clustering

Electricity consumption data gathering: The first step includes the collection of the consumption data of electricity customers. Like any other practical world data gathering, a pre-processing is needed to discover the missing (incomplete) or bad data in the data set. The missing data can be repaired at this stage by using different techniques like regression methods or can be handled by special ways as a part of clustering stage [11]. A customer's data may also be corrupted by the noise or by the occurrence of uncommon situations like

anomalous days or unexpected failures. Consequently, elimination or replacement of bad data is another essential pre-processing step.

Data size reduction/feature definition/ feature extraction: In some cases, before the main clustering stage, the collected smart meter data are processed in some ways to reduce the scale of input data or to define more meaningful features for categorizing the customers. This preliminary stage can be categorized by feature definition (expert knowledge-based feature extraction), feature extraction, and dimensionality reduction techniques. These methods are described in Chapter 4.

Clustering stage: Use of proper clustering techniques and accurate selection of parameters is vital in this stage; although, it depends on various factors such as the size of available data, the final goal of clustering, on-line or off-line clustering, the analytics and computational facilities, and user preferences. Sometimes more than one clustering method may be applied to the load patterns, and final results will later be compared to attain the best results. Furthermore, a combination of different clustering techniques is also possible to speed up the process or to obtain better outcomes [54]. We will discuss the main clustering techniques and their relevant parameters in Chapter 3 and study their applications for load data segmentation.

Clustering performance assessment: Since the clustering of a data set is an unsupervised process, it is not very clear how to assess the quality of the resulted clusters in an objective way [11]. Intuitively, a good clustering method must ensure that each cluster is compact and different clusters are widely separated from each other [55]. Usually, various clustering validity indexes (CVIs) are used to evaluate the clustering results.

Formation of customer classes: This stage represents the post-processing of the formed clusters, mostly based on the real-world scenarios. For example, the final number of clusters cannot be more than a certain number if the final goal of clustering is to define cluster-specific tariffs or to apply DR programs. So, the number of customer segments should be specified by the ultimate user like the retailer or DR aggregator. In this case, some clusters that have similar patterns may be consolidated [53].

2.5 Literature on Clustering of Load Data

Many clustering algorithms are proposed in the data mining community, and for each method, different variations are developed. In the power system literature, some of these techniques have been applied to load patterns of customers.

In the following, a review of the most important works from the literature is presented. Table 2-4 reports these clustering methods and the corresponding references.

K-center family methods are by far the most common approaches used in the literature. Ref. [56] utilizes the electricity consumption usage of 103 residential dwellings with the time resolution of 1 minute. The data are firstly averaged over each hour to build up the hourly load profiles and a RLP is created for every home for each season of the year. K-means is applied to partition the dwellings into two clusters for each season.

Table 2-4 Clustering methods for load data clustering

Method	References
K-means	[6] [53] [32] [57] [58] [59] [33] [60] [61] [62] [63] [64] [56] [65] [66]
FCM	[6] [67] [32] [57] [59] [33] [60] [64] [68] [69]
Hierarchical	[70] [32] [59] [54] [33] [30] [60] [64]

SOM	[6] [71] [72] [58] [33] [60] [61] [73] [74] [75] [76]
Model-based approaches	[48] [70] [27] [77]
Other methods	
K-medoids	[29] [58]
Adaptive K-means	[78] [22] [54]
K-Shapes	[65]
Follow the leader	[32] [79] [57]
DBSCAN	[80] [81]
ISODATA	[82]
Optimum-path forest	[83]

A similar procedure is followed in [62], where the data of just working days are used. An improved FCM is used in [67] to cluster the electricity consumption data of one month of 938 households in China. FCM is also applied along with the K-means and hierarchical algorithms to the consumption data of a group of South Korean high voltage customers [59].

Ref. [84] uses a fuzzy Gustafson-Kessel clustering for identification of non-technical losses. This clustering method can be seen as an extension of regular FCM in which Euclidean distance is replaced by a dissimilarity measure that results in hyperellipsoidal clusters. This method can provide greater flexibility for the shape of clusters.

Clustering of a set of LV substations in the United Kingdom is performed in [30] using a hierarchical algorithm. 15 different loading conditions are considered by dividing the year into 5 seasons and 3 types of days.

Clustering with SOM has been done in several studies. In [71] an SOM-based clustering of Finnish electricity consumers is presented. The aim is to introduce a visual

data mining driven application to exemplify the potentials of real-time business intelligence for electricity companies. In [74], besides the annual electricity usage, various physical characteristics and property features are used for the clustering. On the other hand, an SOM-based methodology used in [73] to segregate customers based on three different sets of indices: information on the clients' climate areas, quantitative information extracted from daily load patterns, and quantitative and qualitative information obtained from questionnaires.

GMM models are recently used in some studies to cluster electricity customers. Labeeuw et al. [48] analyze the electricity demand of 58 households. They favour a GMM approach to K-means, FCM and hierarchical algorithms because of the smoothing effect of GMM and the need for data upsclaing. Moreover, GMM is used for segregating 3600 residential customers [27] and its performance is compared with K-means [77] and hierarchical and K-means methods [70]. A few other algorithms such as adaptive K-means [78] [22] [54], follow the leader [32] [79], k-shapes [65], and density-based spatial clustering of applications with noise (DBSCAN) [80] [81] have been also used in the literature for clustering of load data.

K-means method needs to determine the number of clusters before running the algorithm. Instead of trying out several candidate values for K, an adaptive K-means algorithm can be utilized to determine the final number of clusters during the cluster formation process [85]. This algorithm starts with an initial best guess $k = k_0$, but permits changing it on the go whenever it appears too large or too small for a given dataset [11]. Kwac *et al.* [54] proposed a clustering methodology which combines adaptive K-means and

hierarchical clustering. Firstly, adaptive K-means is applied to segregate customers to a large number of clusters. In the next stage, a hierarchical clustering merges those clusters that are highly correlated.

Fahiman *et al.* [65] compare the performance of K-means with a newly introduced clustering method called K-shapes algorithm to cluster several thousands of dwellings. K-shapes considers the shape of time series during clustering rather than treating the observations as independent attributes. It consists of three main components [86]: 1) a shape-based distance measure which is based on a cross-correlation measure, 2) time series shape extraction which defines a centroid based on an optimization problem, 3) shape-based time series clustering which clusters time series data based on the last two steps. The authors claim that K-shapes significantly outperforms the K-means with respect to clustering accuracy.

DBSCAN technique clusters those observations which are closely packed together and specifies the data points in low-density regions as outliers. This technique is employed in [80] for clustering customers' load patterns and designing customized tariffs for each household based on its dominant load pattern. Ref. [81] uses an adaptive DBSCAN to find a typical consumption pattern in each season for each individual customer. K-means is then applied to group these typical load curves into several clusters.

Biclustering techniques are used in [87] to analyze the building consumption data. The biclustering allows simultaneous clustering of both the observations (buildings) and features (days). The proposed method obtains subgroups of buildings that exhibit a similar consumption pattern during a specific time period.

Furthermore, Markov model is used in [22] to capture the dynamics of the load data and transfer the large data set of load curves to some state transition matrices which are used for clustering. Ref. [52] suggests that when weather effects are accounted for, household consumption is solely based on the occupancy. Here, occupancy refers to socio-demographic factors and the lifestyle. A hidden Markov model (HMM) framework is utilized to infer the occupancy states from consumption data. Spectral clustering is used to segment the collection of HMMs.

2.6 Clustering Applications in Smart Grid Environment

The availability of consumption data offers unique opportunities to electricity companies to recapture the investment costs of AMI systems and to gain benefits through new services. However, to make use of such opportunities, data analytics tools and techniques need to be implemented in the business intelligence (BI) systems of electricity companies [88], [71]. BI includes applications, technologies, and processes for gathering, storing, accessing, and analyzing data to help end users make better decisions [89]. The introduction of smart meters has opened up possibilities for companies to design various BI-enabled business applications such as billing, tariff designs, and DR programs. Moreover, these systems can integrate measured energy consumptions with the other data such as demographics of customers as well as external data like market prices or weather information to improve network operation, define new services, and benefit power companies [71]. Fig. 2.8 illustrates such a structure including the BI system. As this figure shows the data mining tools, specially clustering techniques, can be included in the BI systems for defining innovative business applications.

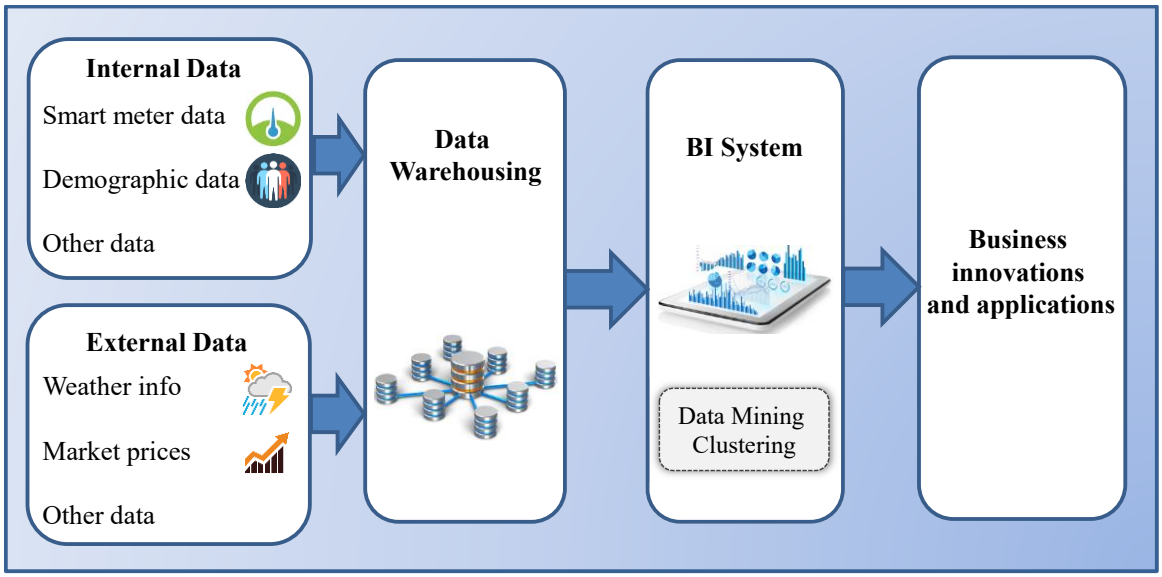


Fig. 2.8 Architecture of BI systems of companies

Despite the considerable advancements in the clustering of load patterns and the recent publications which have explored the different applications of clustering of load data, there is not any comprehensive study which summarizes and analyses these applications in power systems. The previous publications in this area either reviewed merely the clustering of load patterns [32, 55, 90] or briefly discussed some of the applications that can be facilitated by clustering techniques [88], [2], [91]. In this regard, this section tries to fill this gap by highlighting the major applications that can be established by including the clustering techniques in the BI systems of companies. Each of these applications is described in detail, with appropriate studies from the literature being reviewed. In addition, in each section, characteristics of many projects/trials/studies, which have been conducted in various countries and which utilized clustering techniques for mining load data, are summarized and reported.

2.6.1 Load forecasting

Load forecasting has always been an important issue for power system operation and planning. Energy consumption prediction can be categorized by very short-term [78], short-term [92], medium-term [93], and long-term load forecasting for different time horizons. Short-term forecasting is critical in unit commitment, task scheduling of both power generation and distribution facilities, and DR applications [94], [95]. Therefore, accurate estimate of the electric load is essential for electricity providers and system operators. Any error in this forecast might result in electricity shortage and reliability issues. On the other hand, accurate forecasts allow utilities to operate at lower costs and save considerable amounts of money each year. Also, decision makers need to forecast the load in the long-term for planning and expansion of the network.

The prediction of the load at a specific time t in the future depends on a weather independent component containing trend, seasonality and calendar effects, and a weather dependent component that shows the effect of variables such as temperature and cloud cover [96]. A wide variety of forecasting techniques have been proposed in the last two decades including regression techniques [97], time series models such as auto regressive with moving average (ARMA) models [98], support vector machines (SVM) [99] and neural networks [100], [101].

The aforementioned techniques have been typically used at large scales, for example, for forecasting the load of an entire region or even a country. So, the predictions are made based on the aggregated data which comprise a large number of households. The global rollout of smart meters has created new opportunities for further improvement of electric

load forecast accuracy [94], [65]. In this regard, cluster-based load forecasting can be used to divide customers into classes with similar consumption behaviors and perform forecasting for each cluster of customers separately. This method has obvious advantages in comparison with two extreme approaches that have been previously considered in the literature [65], [101]: (1) completely aggregated method which aggregates the energy consumption of all households into one time series (the aggregate consumption) and uses this accumulated data for forecasting, and (2) completely disaggregated method that forecasts the energy consumption of each individual household separately, and then adds the individual predictions to obtain the prediction at the aggregated level. The latter approach is complicated as the consumption of a single household is highly variable and fluctuates widely. The former method, on the other hand, can be improved if more information of individual customers' load changes will be considered in the forecasting process. The clustering-based approach allows cancelling out the individual variations in consumption by aggregating the load in each cluster, and enables achieving higher accuracies in the prediction by applying forecasting methods on each class separately. Furthermore, it is possible to apply different forecasting techniques for different clusters. The structures of these methods are shown in Fig. 2.9.

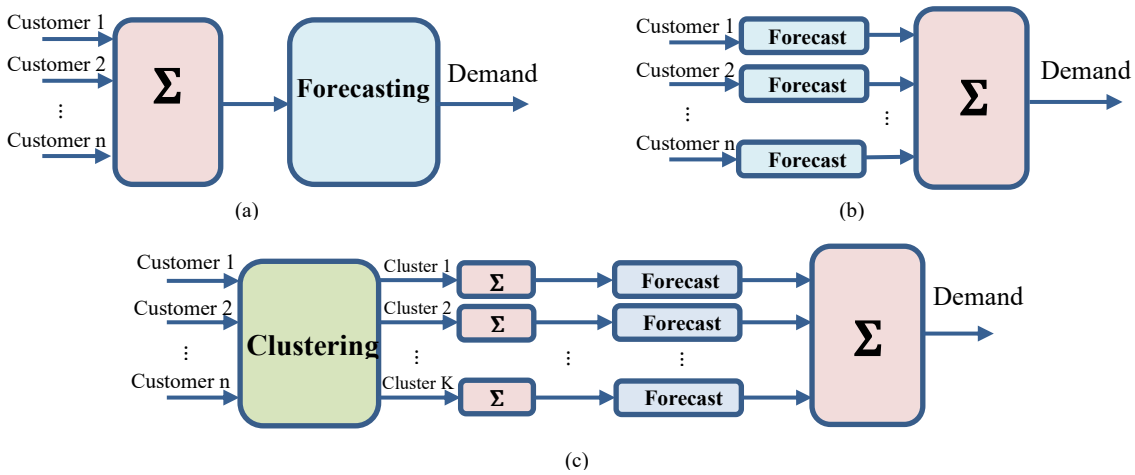


Fig. 2.9 Different methods for load forecasting: (a) Completely aggregated. (b) Completely disaggregated. (c) Clustering-based forecasting

The recent studies on cluster-based load forecasting can be categorized according to their feature set selection methods, clustering techniques, and the applied prediction methods. Table 2-5 summarizes the accomplished studies and their specifications.

The analysis in [101], which uses various forecasting algorithms including support vector regression (SVR), linear regression, and multi-layer perceptron for one hour- and 24 hour-ahead forecast, confirms that cluster-based approach displays better results if a suitable number of clusters are selected and the size of the customer base is large enough.

The load forecasting of the distribution system of a district in China is presented in [102] using a combination of spatial load forecasting and clustering analysis. The distribution network under study is divided into small regions called cells and the load is predicted for each cell using different forecasting algorithms and partitioning method.

Table 2-5 Studies that used clustering techniques for the improvement of load forecasts

Ref.	Sample Size	Dataset	Time resolution (minutes)	Clustering method	Load forecasting method	Descriptions
[65]	3176 residential consumers	CER (Ireland)	30	K-means & K-shapes	Traditional neural networks & deep neural networks	Use of a weighted summation method based on the number of members in each cluster
[101]	782 customers	CER (Ireland)	30	Max-AC (maximize the autocorrelation of the energy consumption of the clusters) Min-Stdev (minimize the fluctuation in the clusters' energy consumption) Max-Sim (maximize the similarity among customers within a cluster.)	Seasonal ARIMA Linear Regression Multilayer Perceptron SVR	
[78]	50 smart meters	N/A	15	Online clustering algorithm	ARIMA	
[96]	2309 industrial customers	France	60	Hierarchical	General forecasting method employed by French Electrical Company	Pre-processing of individual customer load data using WT
[103]	~ 6000 customers	CER (Ireland)	30	Kernel spectral clustering	A periodic autoregressive model with exogenous variables (PARX)	Comparing different distance measures
[94]	3176 residential customers	CER (Ireland)	30	K-means	Neural Network	Examines three representations of data for clustering
[8]	N/A	1st dataset: Consolidated Edison Company of New York, Inc. 2nd dataset: CER (Ireland)	1st dataset: 15 2nd dataset: 30	K-means	Neural networks	

Table 2-5 Studies that used clustering techniques for the improvement of load forecasts (Cont'd)

[104]	1st dataset: 5066 residential consumers 2nd dataset: 3639 residential consumers 3rd dataset: 3630 consumers	1st dataset: Low Carbon London project 2nd dataset: CER (Ireland) 3rd dataset: Slovak electricity consumption data	1st dataset: 30 2nd dataset: 30 3rd dataset: 15	Partitioning around medoids (PAM)	Triple exponential smoothing Recursive partitioning regression tree Conditional inference trees Random Forests	Evaluating various representations of data prior to clustering
[105]	1st dataset: 3639 residential consumers 2nd dataset: 3607 consumers 3rd dataset: 300 residential customers	1st dataset: CER (Ireland) 2nd dataset: Slovak electricity consumption data 3rd dataset: Ausgrid dataset (Australia)	1st dataset: 30 2nd dataset: 15 3rd dataset: 30	K-means & PAM	Seasonal naïve method Multiple Linear Regression Random Forests Conditional inference trees	Examining different model-based representations of data prior to clustering stage
[106]	11281 customers	Slovak electricity consumption data	15	K-shape	Deep neural networks with a Sequence to Sequence (S2S) architecture	
[107]	Two case study buildings (a mall and a hotel)	China	60	Fuzzy c-means	SVR	Use of a combination of SVR and wavelet decomposition
[108]	N/A	N/A	N/A	K-means	SVR/ Seasonal decomposition of time series based on loses regression / Random Forest/ Gradient Boosting Machine / Regression Trees	Studying four representations of data prior to clustering stage
[109]	N/A	China	N/A	Ant colony fuzzy clustering	SVM	
[110]	N/A	China	N/A	Fuzzy clustering	Target theory	Forecasting the maximum load of a grid

Ref. [78] investigates the applications such as dynamic demand response (D^2R) that require very short-term load forecasting. Therefore, ARIMA model is used as a time series forecasting method for the prediction of the load. The aim is to forecast the household's energy consumption. Hence, the goal of clustering is to minimize the ARIMA prediction error for the cluster as a whole to minimize the cumulative consumption prediction error. For the experimental stage, the data of 50 smart meters for a period of 3 months are used. The first 2 months are used as training data and the last one month as evaluating set.

Both [111] and [95] use ant colony clustering for short-term load forecasting. In [111], an SVM forecasting model based on ant colony fuzzy clustering algorithm is presented. The main objective is to overcome the problems associated with SVM algorithm such as slow convergence speed and low forecasting accuracy under the condition of much redundancy and noise in the training data. So, historical data are preprocessed by clustering method. Comparison of the results of the proposed method with the SVM results shows the improvement in the processing speed and forecasting accuracy.

Authors in [110] use fuzzy clustering to address the correlation among different load categories. By use of clustering, the influencing groups are determined and the groups which have a strong correlation are classified as a category and the load forecasting is performed for them.

Misiti *et al.* [96] investigate creating customer clusters to improve the accuracy of load forecast of the French Electrical Company. The aim is to create customer clusters such that the sum of disaggregated forecasts will perform significantly better than the aggregated forecast. The method comprises three main steps: a pre-processing of individual customer load data using WT, primary customer clustering which produces a large number of clusters,

and finally an iterative optimization that reduces the number of clusters. Clustering is performed using the hierarchical algorithm and the load is predicted based on the general forecasting method employed by French Electrical Company. The forecasting performance measured by long-term Mean Absolute Percentage Error (MAPE) shows a significant improvement for disaggregated case (2.49%) compared with the aggregated approach (4.06%).

The improved forecasting is achieved for CER data set [12] in different studies using kernel spectral clustering [103], a combination of deep learning and K-shapes [65], and application of K-means on three representations of the data set [94]. Ref. [65] compares the performance and forecasting accuracy of four different approaches. Two clustering algorithms, K-means and K-shapes, and two forecasting methods based on the traditional neural networks and deep neural networks are utilised for this purpose. Another contribution is the use of a weighted summation method for accumulating the forecasting results of each cluster according to the size of their membership. Furthermore, three different representations of CER data set are examined in [94] for clustering including the full load pattern, average daily load pattern, and regression coefficients while the number of clusters is fixed. The coefficients obtained from the regression method are different for each customer and are used as the features for clustering. The artificial neural network is used for forecasting purposes. The forecast error decreases dramatically using regression coefficients in comparison with the case without clustering.

Neural networks are also deployed in [8] for forecasting the load of each cluster. K-means is adopted for clustering load profiles and the number of clusters is decided based on forecasting performance which is measured by MAPE.

2.6.2 Demand response

DR is considered as one of the fundamental parts of demand side management along with the energy efficiency and behavioral change programs [112], [113]. As the smart grid concept continues to evolve, various methods have been developed to enhance the efficiency of power system. DR is considered as one of the most cost-effective and reliable solutions for the smoothing of the load curve and helping system when it is under the stress [112]. According to the Federal Energy Regulatory Commission, DR is defined as: "Changes in electric use by end-use customers from their normal consumption patterns in response to changes in the price of electricity over time, or to incentive payments designed to induce lower electricity use at times of high wholesale market prices or when system reliability is jeopardized" [114]. Another recent definition by the European University Institute designates it as: "changes in electric usage implemented directly or indirectly by end-use customers/prosumers from their current/normal consumption/injection patterns in response to certain signals" [48, 115]. This customer-enabled power consumption management enables the adaption of power demands to time pricing or incentives which consequently can improve the efficiency and reliability of power network [112, 116].

The research on DR provision from residential customers has significantly increased in recent years as DR can contribute to ancillary service provision [117], facilitating renewable energy integration [118-121] and making virtual power plants [122], [123].

DR programs were traditionally focused on larger electricity customers like industries or commercial users. Residential sector was mostly neglected as in most cases there was no access to loads of individual customers. The insignificant level of individual residential

loads in comparison with the large loads was another determining factor in disregarding consumption of the dwellings for DR applications. Introduction of AMI and smart meters, however, allows the procurement of DR from the households. The new agents in electricity system structures called DR aggregators are responsible for this [124] [125]. However, there are still many unsolved questions regarding the role of these aggregators, the interactions of them with customers and system (market) operators, and the challenges that they in practice face.

Generally, the fundamental functions of DR can be explained by these three actions as shown in Fig. 2.10: peak clipping which refers to reducing the peak energy consumption; valley filling which refers to endorsing the off-peak energy consumption; and load shifting which is a combination of the two above mentioned methods and aims at shifting the energy consumption over the time horizon for example from peak to off-peak periods.

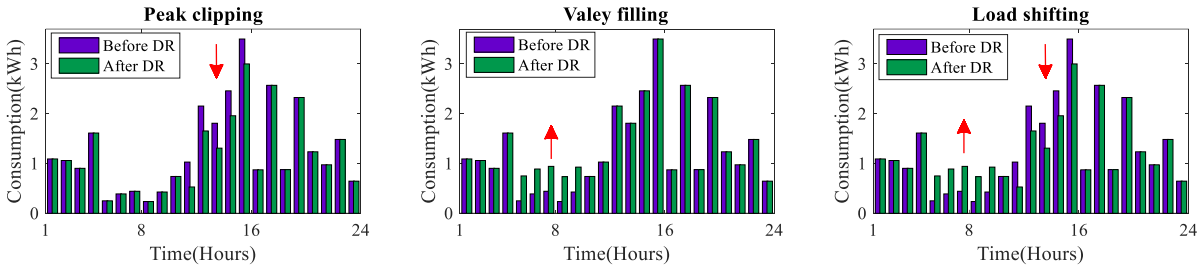


Fig. 2.10 Fundamental functions of DR programs

DR programs are traditionally divided into two main classes: Price (time)-based DR and incentive (event)-based DR [112, 126] [127]. Price-based program provides users with different electricity prices at different times. They can also offer customers time varying prices that are defined based on the cost of electricity in different time periods [128]. Incentive-based programs pay participating users for demand reduction during periods of

system stress like peak load or system contingencies. The main price-based DR programs include time of use pricing and critical peak pricing which respectively charge users with different prices for different time of day (on-peak, mid-peak, off-peak) and for special peak periods when system is under the stress. Major incentive-based programs are direct load control (DLC) and interruptible/curtailable (I/C) load. DLC loads can be remotely cycled or turned off by the utility, and can normally be deployed within a relatively short notice [129]. In I/C program customers receive a discounted rate for agreeing to reduce load on request when the grid reliability is jeopardized [130]. They are also various other options for DR which mostly have been used by large customers in industrial and commercial sectors.

Consumption of a household can be distinguished by three important features [54]: load shape that displays households time of day energy consumption, actual volume of energy use, and daily usage pattern stability over time. Different DR methods can target each of these three aspects. Customer behavior, usage pattern, and willingness to participate in DR are other important issues that significantly affect DR policies. Clustering of customers to different classes that show similar patterns is a promising way for DR program targeting and customer engagement. For example, if households whose peak demand corresponds to the total system peak are identified, they may be good prospects for recruiting for residential DR programs. However, the willingness and ability of households to shift energy usage off peak must be determined by household surveys [53]. Clustering can help in this stage too to divide the customers based on their attitude toward DR [48]. The studies on the application of clustering for DR are reported in Table 2-6.

Table 2-6 Studies that used clustering techniques for DR analysis

Ref.	Sample Size	Dataset	Time resolution (minutes)	Clustering method	Descriptions
[53]	8337 residential households	PG&E (US)	15	K-means	Finding the best candidates for reduction of summer peak
[54]	218,000 (66 million load shapes)	PG&E (US)	60	Adaptive K-means + Hierarchical	Different aspects are analyzed like entropy, time and quantity of use, effect of climate zone, and seasonality./ Household targeting for energy efficiency programs and recommendations for time of use shifts
[48]	Load data of 1693 households/Surveys for 500 households/ Electricity measurement at appliance level for 58 households	Belgium	15	Expectation-maximization (EM) model for clustering based on the load data K-means for segmentation based on household attitudes	In-home surveys are used for social segmentation. Finding DR potentials of wet appliances
[29]	13827 load curves	Opower Corporation (US)	N/A	K-medoids (DTW similarity metric) & K-means	
[72]	1800 users/ 700 network DR plan, 700 retailer DR plan, 400 control group (55800 load profiles totally)	Australia	30	SOM	Use of PCA for data size reduction Finding the effects of DR plans on consumption
[27]	3622 customers	CER (Ireland)	30	Gaussian Mixture Model (GMM)	To understand the peak demand and major sources of variability in customers' behavior
[52]	1100 users/ socio-economic data of 950 households using surveys/ weather measurements (5- to 15- min resolution)	US based Google employees	10	Spectral clustering and K-medoids	Use of surveys for relating the consumption data to certain household characteristics Understanding inter-temporal consumption dynamics of users for DR management
[22]	6,445 customers (4,511 residents, 391 industries, and 1533 unknown)	CER (Ireland)	30	Distributed clustering	SAX for data size reduction Investigating electricity consumption dynamics/ Addressing the challenges of high-dimensional consumption data

Authors in [54] evaluate a huge amount of consumption data in different ways to understand the main aspects of usage: time of peak, variability, and quantity as well as the effect of seasonality and climate zone. For this purpose, they firstly decompose daily usage patterns into daily total usage, which is later described through a log normal distribution, and a normalized daily load shape which is used for clustering. Clustering is performed through an adaptive K-means and a hierarchical algorithm. Based on this study, they propose a multidimensional analysis which divides households to heavy, moderate and light based on the consumption magnitudes and stable, moderate and variable based on the variability of usage. All of these analyses can help to develop proper DR programs for each residential household.

The potential of wet appliances (washing machines, tumble dryers, and dishwashers) for DR is investigated in [48]. In this research, three sets of data are used: consumption measurement, in-home surveys which ask customers about their attitudes towards DR, and measurement at appliance level for a limited number of households. Firstly, a model-based clustering algorithm is used to divide the households into ten classes based on their load curves. Secondly, by using the K-means method, households are categorized into four groups of advocates, supporters, sceptics, and refusers based on their attitude toward demand reduction. Finally, the individual appliance measurements are scaled up among clusters to characterize the potential of load reduction by wet appliances.

The authors in [29] explain three issues that needs to be tackled for providing DR from customers: firstly, categorizing those users that have the most important impact on energy reduction; secondly, predicting a given customer's energy consumption; and thirdly, estimating the devices and the time of use of them for that customer. The exact time of use

or sequence in which devices are used can change greatly from one day to another. Thus, the authors suggest that using a similarity metric such as dynamic time warping (DTW) which tries to characterize underlying time shifted consumption behavior is most appropriate. Since DTW is invariant to contraction and expansion and small shift, prediction error from DTW is less.

Ref [131] addresses the problem of segmentation of electricity users for the utilities by using consumption, demographics, and previous program enrolment data. The final goal is to extract those users that are most probable to enrol in different energy efficiency or DR programs and to target each group with efficient appropriate messages. The study is performed on a large population of about 1 million users.

The effect of acceptance of two different DR programs on consumption patterns of Australian households is studied in [72]. The authors try to give an insight into the change of users' actual electricity load profiles after participating in these programs. To this end, two groups of customers, control group customers and customers under DR, are compared. After applying a two-stage clustering by PCA and SOM algorithms, consumptions behaviour are detected and customers are categorized by their consumption levels and time of use. Analysis results show the effectiveness of DR programs as customers change their time of use after subscribing to these programs.

2.6.3 Tariff design

One of the most important applications of clustering of electricity users is to design suitable tariffs for different customers based on the classes that they belong to. A successful tariff structure should be capable of shifting the load demand over the time and encourage

customer acceptance by providing economic benefits. In the last two decades, TOU tariffs have been employed by the utilities as an alternative to flat tariffs to encourage customers to decrease their energy level during peak hours and save on their own bills. The findings of clustering process could guide the electricity companies to formulate new pricing differentiation strategies [71]. Therefore, as different customers show different load patterns, clustering can help to design cluster-specific tariff structures which can result in the reduction of peak load. For example, the customized electricity retail prices offered to different clusters are shown in Fig. 2.11.

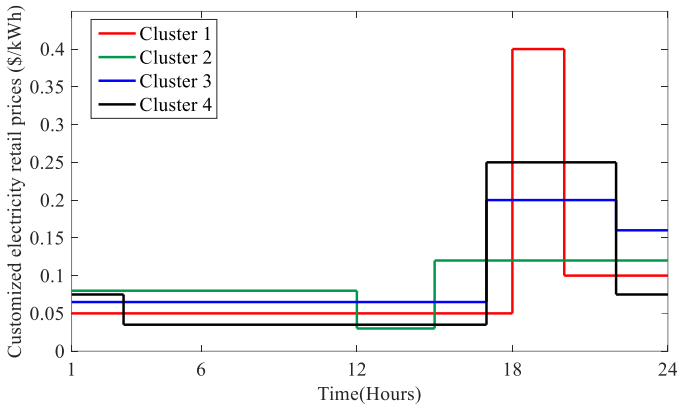


Fig. 2.11. Customized tariff structures for different clusters

In practice, however, it is not possible to define various tariff patterns as it is not practical for retailers or utilities to manage many kinds of tariff programs. A suitable mechanism is to include the clusters with the most number of users or the clusters which contain those customers that are categorized as heavy and medium energy users. The appropriate tariff is identified for each class and the other customers can be assigned to the most similar groups. The other major design concerns [70] include considering the yearly load demand variations and seasonality effects, as well as customer acceptance of the proposed tariff structure. Finally, the related utility or retailer should periodically monitor

the daily load patterns of its customers, updates customer classes by automatic clustering, and adjust the tariffs applied to each customer class [79].

Table 2-7 highlights the major works which utilized clustering techniques for defining customized tariff structures.

Table 2-7 Studies that used clustering techniques for designing tariff structures

Ref.	Sample Size	Dataset	Time resolution (minutes)	Clustering method	Descriptions
[70]	N/A	Ameren Illinois database	N/A	K-means, Hierarchical, GMM	
[132]	210 load profiles of medium voltage consumers	Iran	N/A	Fuzzy K-means	Determining the optimal selling price for an electricity retailer
[79]	471 non-residential customers	Romania	N/A	Follow the leader	Defining a set of indices for customer classification/ Studying the customized tariffs
[10]	N/A	Germany	15	K-means	Suggestions for segment-specific tariffs
[80]	31 customers (2771 daily load profiles)	Smart Grid, Smart City project (Australia)	30	DBSCAN	Use of a mixed integer nonlinear programming model for customizing retail prices
[133]	2000 customers	N/A	15	Use of an artificial neural network and a locality sensitive hashing algorithm	Use of meta-data (load condition) along with the load data for real-time pricing
[134]	N/A	UK	30	GMM	Defining price-oriented and load-oriented tariffs and exploring the effects of these tariffs on domestic DR

The work in [70] proposes a clustering-based methodology for determining optimal TOU structures. The authors outline the underlying parameters of any TOU structure, and compare three clustering algorithm (K-means, hierarchical, and Gaussian mixture model) to characterize the monthly variation of TOU arrangement. To test the efficacy of clustering algorithms, the total monthly bill and the degree of granularity of clusters are used as the main metrics.

In [132] an electricity retailer groups its customers into different classes in order to maximize the annual profits. A weighted fuzzy K-means method is used for this purpose. Then, the optimal selling price of each cluster is determined based on a profit function.

A multiple rate tariff structure is proposed in [79] to replace the single rate structure. It defines four features and clusters customers by using a follow the leader method. To study the new tariff structure, it is assumed that the total revenues under the new tariffs do not exceed the total revenues with the previous one. Moreover, it suggests a procedure for identifying the customer's class and assigning an appropriate tariff to the customer.

Clustering of a group of customers in Germany is studied in [10] in which the authors use K-means for clustering. A number of time zones based on the peaks and valleys of customer are defined and a time-variable rate is assigned to them. In the next step, the authors propose a segment-specific rate design which sets a different rate for each segment of individual customers.

In a recent publication, the problem of customized retail prices are studied by Yang *et al.* [80]. The customers are initially clustered using the DBSCAN algorithm. In the next step, a mixed integer nonlinear programming (MINLP) model is formulated to optimize the structure of TOU retail price and the price level for each cluster.

The developed method in [133], is suitable for real-time applications such as the real-time pricing. The association between the load data and the corresponding meta-data allows judging whether a certain load diagram represents a cost-efficient or expensive behavior at a moment of time. In other words, the focus here is on a behavior classification of customers rather than a standard customer classification.

2.6.4 Classification of new electricity customers

Classification is considered as one of the major data mining tasks. The goal of classification is to classify instances (observations) into a set of predefined classes or categories [11]. In other words, classification is the problem of identifying to which of a set of classes a new observation belongs, on the basis of a training set of data containing instances whose class membership is known. Classification is common in almost every aspect of everyday life, for example, to assign customers, stores, documents, emails, or any other type of instances into a set of known classes. The algorithm that implements classification and maps the instances to the classes is known as a classifier.

In power systems, classification can be used to assign new customers or the customers without smart meters to the classes that are previously formed by the clustering process. A set of features (explanatory variables) can be defined based on the survey data (or the fixed data) and a limited amount of consumption data. Surveys contain the information regarding the physical aspects of dwellings and household characteristics. Therefore, various features can be extracted from the survey responses and a limited amount of energy data of the customer. Based on these features, the classifier returns a variable indicative of the customer group in which the customer best fits [62]. Such structure is illustrated by Fig. 2.12.

Table 2-8 shows the major works which have used the clustering process prior to the classification.

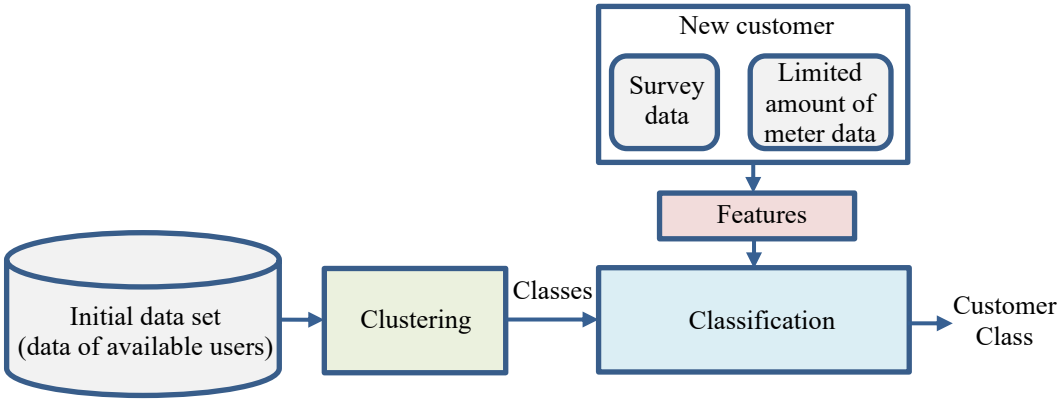


Fig. 2.12. Classification of electricity customers

Table 2-8 Studies that used clustering techniques as a preliminary stage prior to the classification

Ref.	Sample Size	Dataset	Time resolution (minutes)	Clustering method	Classification method	Descriptions
[62]	3440	CER (Ireland)	30	K-means	Logistic regression (LR) Classification and Regression Tree (CART)	Use of feature selection techniques Classification of new customers
[59]	3183 high voltage customers	Korea Electric Power Cooperation (KEPCO)	15	K-means, Hierarchical, FCM	N/A	Generating typical load profiles of non-AMR customers
[63]	1022 medium voltage customers	Portugal	15	K-means, other variations of K-means	C5.0 algorithm	Classification of new customers
[61]	165	Portugal	15	SOM, K-means	C5.0 algorithm	Classification using the load shape indexes
[30] [31]	824 substations, 3500 feeders	UK	10	Hierarchical	Multinomial LR	Classifying unmonitored substations/ Estimating the load peak of substations by use of regression methods
[73]	625 (216376 daily load profiles)	Spain	60	SOM & K-means	SOM classifier	Classification of new users by a SOM classifier and estimating the characteristics of energy consumption habits of them

Ref. [62] presents a detailed procedure for classification of new electricity customers based on the survey responses and some information of the electricity consumptions. A

limited amount of metering data is used in order to emulate the analysis of new electricity customers for which only a small amount of data is available. In the first stage, clustering is performed separately for each season and a total number of 4 clusters are decided for each of them. In the next step, a large set of features are considered for classification including 47 features extracted from the survey information and a set of features which are defined based on the available load data. To further improve the classification accuracy, application of three different feature selection methods and two different classifiers are examined.

The attained results show that the use of survey data significantly increases the accuracy of classification task (up to 20%). With the growth of available smart metering data, the simulations show an increase in accuracy achieving up to 60%, 70%, and 80% accuracy, respectively, with 1, 4 and 8 weeks of data.

The study in [59] seeks to generate the load profile of users without automatic meter reading (AMR) facilities based on the usage data of customers equipped with AMR. It firstly clusters AMR users and calculates a typical load profile (TLP) for each cluster as the mean of load patterns belonging to that cluster. Then the daily load profile of each non-AMR user is created by comparing different attributes of that customer with TLPs. So, the most similar TLP is selected and the classification of non-AMR users is performed. Ref. [63] also classifies the non-AMR medium voltage customers. The initial data set is clustered by K-means algorithm and other variations of K-means and classification is achieved using a rule set obtained by C5.0 classification algorithm.

Ref. [61] presents a framework for electricity consumer characterization. This framework includes two parts: the load profiling module and the classification module. In the first stage, the customers are grouped into clusters by using a combination of SOM and

K-means algorithms and a load profile for each cluster is constructed. In the classification stage, three load shape indexes (load factor, night impact, and lunch impact) and commercial indexes are used to classify customers.

[30] and [31] report the findings of a project for estimating the peaks and shapes of unmonitored low voltage substations in the UK. As the cost of installation of proper measurement equipment for thousands of substations will be very high, a reasonable approach is to assign the unmonitored ones to the formed clusters of monitored substations. So, in the first step monitored substations are clustered using hierarchical and K-means algorithms and a template is defined as the average normalized load pattern of load shapes belonging to that cluster. In the second step, unmonitored substations are assigned to these templates just based on their own fixed data like the capacity of low voltage substation or information of outgoing feeders. For this purpose, a multinomial LR model is utilized.

The authors in [73] use an SOM classifier which allows a correspondence between new users' load profiles, and one of the obtained patterns from the clustering. Furthermore, the characteristics of new customers based on the classification of their load curves are predicted.

2.6.5 Non-technical loss detection

Losses in power grids include both technical losses and non-technical losses (NTLs). NTLs are closely related to frauds through illegal electricity connection, hardware tampering, and unpaid bills [135] [136]. NTLs have major negative consequences for both developed and developing economies since they pose significant costs to utilities, states, or

legitimate consumers [135]. The total amount of NTL costs is estimated to be around $25 billion every year [137]. Consequently, the reduction of NTLs has been receiving increased attention both in academia and industry.

Different methods are proposed for the detection of NTLs which work based on the analysis of consumption data, assuming that NTLs cause a deviation from normal consumption behavior. The classification-based techniques including artificial neural networks, SVM, decision trees, and naive Bayes are widely used for inferring the probability of the presence of NTLs [138-140]. State estimation is another popular method which works based on the estimation of the power flow to the customer node. They can find a possible NTL when there is a significant deviation between the estimated power flow and the billed energy consumption [141].

Clustering techniques are sometimes used directly to detect NTLs by identifying the customers with irregular consumption behavior. However, in most cases, they are used as a preliminary step to divide customers or load patterns into the groups that display similar characteristics which are then used by the classification methods to discover NTLs. Clustering is also utilized to calculate prototypes or power profiles. A significant difference between a new sample from these prototypes suggests a possible fraud [142]. The literature on NTL which employed the clustering techniques is reported in Table 2-9.

Ref. [143] proposes an approach which involves an SOM clustering to identify possible periods of frauds for high voltage electricity customers. A detailed study of NTL identification using OPF clustering algorithm is presented in [83] for two different datasets consisting of commercial and industrial customers and the results are compared with other traditional clustering algorithms.

Table 2-9 Studies on NTL which used clustering techniques

Ref.	Sample Size	Dataset	Time resolution (minutes)	Clustering method	Classification method	Descriptions
[83]	1st dataset: 3178 industrial users/ 2nd data set: 4948 commercial customers	Brazil	N/A	Optimum-path forest (OPF), K-means, GMM	OPF classifier	Use of 8 features
[136]	A subset of 20,126 consumers	Brazil	N/A	FCM	Fuzzy classification	Defining and calculating 5 attributes for each customer
[144]	Training data: 4701 load profiles/ test data: 1273 load profiles	KEPCO	N/A	K-means	LR, SVM, K-nearest neighbour (KNN)	Use of PCA before the clustering
[84]	2515 customers	CER (Ireland)	30	Fuzzy Gustafson-Kessel clustering	×	The method is compared with K-means, GMM, DBSCAN, and SVM
[145]	2982 consumers	CER (Ireland)	30	DBSCAN	×	Use of PCA before the clustering
[146]	200	Malaysia	30	K-means, EM	×	

An anomaly detection method for identifying cyber intrusion attempts is introduced in [145]. The method combines PCA technique and DBSCAN clustering to verify the integrity of measurements. The case studies are conducted for nearly 3000 customers where the 60-week data of each customer is represented by only 2 principle components. DBSCAN algorithm is then applied to distinguish between the points corresponding to regular weeks and the points corresponding to anomalous weeks.

The NTL detection framework in [146] uses a clustering module, which compares three clustering algorithms, in conjunction with a classification module. Angelos et al. [136] build a two-step model encompassing a FCM clustering and a fuzzy classification for identification of abnormalities in consumption patterns. Based on the six-month values of consumption levels and inspection remarks, they define five features which are

consequently used for the clustering. The procedure results in an index score with the highest scores referring to the potential fraudsters.

Ref. [144] incorporates an additional stage in abnormality detection based on the conditional probability. They assert that a load profile can be considered normal or abnormal based on the different conditions and hence, the condition associated with a load profile should also be considered in the abnormality detection. Therefore, the proposed approach consists of three stages including building the representative prototypes using a combined PCA and K-means clustering, generating a two-dimensional space using conditional probability and similarity, and applying the classification techniques.

2.6.6 Other applications

i) Finding the relationship between household characteristics and consumption patterns

The surveys and questionnaires include various data which can help in understanding the energy consumption behavior of customers. The availability of survey data along with the smart meter data provides this opportunity to assess the importance of each variable on the electricity usage.

For an individual dwelling, it is usually hard to evaluate the correlation between these attributes and its consumption behavior, as no comparisons can be made. On the other hand, clustering can reveal those possible correlations between household features and electricity usage, as customers with specific attributes usually belong to the same cluster. In this respect, several studies paid attention to this possible application of clustering results.

Table 2-10 displays those projects and their specifications which employed clustering techniques to assess the relationship between the survey information and the consumption data.

Table 2-10 Use of clustering techniques for studying the characteristics of households

Ref.	Sample Size	Dataset	Time resolution (minutes)	Clustering method	Descriptions
[56]	103	Austin/Texas	1	K-means	Exploring the relation between electricity price and load shapes/ Finding the correlation between household characteristics and load patterns
[58]	3941	CER (Ireland)	30	K-means, K-medoids, SOM	Profile classes are determined for each customer using the clustering and are linked to household characteristics by applying a multi-nominal LR.
[75]	3941	CER (Ireland)	30	SOM	Investigating the effect of household characteristics on the cluster they belong to.
[74]	~8000 customers	Finland	Annual consumption	SOM+K-means	Comparing each customer with similar customers and creating customer specific electricity saving guidance.

In [56], customers are firstly clustered using K-means. In the next step, the correlation between the explanatory variables in the survey data and the clustering results are found using a regression model. Explanatory variables include a wide range of variables like the number of males, females and kids, vehicles and computers, income, and other physical and behavioral features. The results show that these variables are significant determiners for the inclusion of a household in a special cluster.

A linear multivariate regression model is applied in [58] to figure out the association of each explanatory variable (customer characteristics) with the profile classes which are obtained using the clustering. They use the detailed information of dwellings, occupants and appliance specifications for this purpose. On the other hand, McLoughlin et al. [75] segregate four thousand electricity customers to nine groups and then investigate the

dwelling and occupant characteristics including number and age bracket of dwellers and social class of them to determine whether there is any significance to these characteristics within each group.

The study in [74] creates comparison groups based on the characteristics of customers' buildings. In this way, electricity usage of each customer can be compared with its corresponding comparison group to encourage households to reduce their energy consumption and think about the methods of energy conservation. Moreover, the presented method provides a tool to target and create customer specific electricity saving tips.

ii) Defining new class load curves

Previous practices of customer categorization were involved building the customer class load profiles through sampling consumption data in some pilot projects. In distribution system calculations, such customer class load profiles were used extensively to model the load [82], estimate the state of the network [147] and predict future loads in the distribution system planning. Usually, distribution operators acquired different customer data including: 1) the customer connection information such as customer location, supply voltage, and number of phases; 2) customer class as residential, industrial, agricultural, public and commercial; 3) energy consumption which included some information about monthly or yearly usage; and 4) additional information such as the type of heating systems or physical characteristics of the home. These data were used to assign a predefined load profile to an individual customer.

However, this approach involves several sources of error and presents important uncertainties in dealing with distribution state estimation and planning. First of all, these

customer class load curves are constructed using some sample measurements which comprise an insufficient number of customers. Also, the data is out of date, which cannot reflect the current consumption patterns of electricity customers. Secondly, the information that is used for assigning a load profile to a customer is hardly updated. Therefore, the possible changes in the customer habits, activities and appliances are not taken into account and hence, it is very probable to have improper load profiles for the customers. In addition, some customers may have irregular load patterns or patterns that cannot be fitted to any of the existing customer class load profiles.

Now that thanks to smart meters the consumption data of customers are available, defining customer class load profiles can be performed based on the clustering of customers. To this end, [82] investigates the customer classification for Finish electricity grid. This paper utilizes smart meter data to classify 660 customers which primarily belong to 6 general groups: residential, agricultural, industrial, public administration, commercial, and the remaining customers. Iterative self-Organizing data analysis technique algorithm (ISODATA) is used for the clustering and based on that, customer class load profiles are calculated.

Ref. [23] addresses the problem of classification in Finland too in which 1035 customers are classified based on their hourly energy uses. Each of these customers is assigned to a specific load category by distribution companies, so, correspondence between these original load curves and energy use of customers is investigated. Then new load curves based on the clustering results are created. The new load curves are constructed by calculating the mean of electricity use of each cluster.

Specifications of the two above mentioned projects are further shown in Table 2-11.

Table 2-11 Studies that used clustering techniques for defining new class load curves

Ref.	Sample Size	Dataset	Time resolution (minutes)	Clustering method	Descriptions
[82]	660 customers	Western Finland	60	ISODATA	The method also includes temperature dependency correction and outlier filtering.
[23]	1035 customers	Eastern Finland	60	K-means	Use of 7 features for clustering

2.7 Summary

In this chapter, firstly, the main concepts, trend, and policies of smart metering were explained. Secondly, the main characteristics of residential load pattern were briefly illustrated. In the next step, the concepts, history, and stages of load pattern clustering were presented in detail and a comprehensive review of current literature was provided.

Finally, we highlighted some of the most important applications of clustering techniques for power systems. In each section, firstly, the basic concepts are introduced and secondly, the major recent works from the literature were summarized and reviewed. In Table 2-5 to Table 2-11, we summarized the major characteristics of various projects/studies/trials which are carried out worldwide to characterize customers' energy usage. These projects not only differ in terms of the geographical location and scale but also vary in terms of the time resolution, applied techniques, and final goals. Some of these projects had limited domain by nature. On the other hand, some of them gathered a huge number of load profiles [12], [148] and several studies analyzed millions of daily load shapes [54], [131]. They are useful in providing an insight into the projects and research studies that are carried out globally. The presented materials can greatly help the researchers and the engineers form the energy sector to become familiar with the underlying concepts as well as the applications of clustering techniques for power grids.

3 A Comparative Study of Clustering Techniques for Electrical Load Pattern Segmentation

3.1 Motivation and Objectives

In Chapter 2, the basic concepts of load data clustering were introduced. In this chapter, we will examine various clustering methods and their corresponding parameters which affect the load data segmentation. This study is opportune, because, despite the considerable changes which happened in this area, there is no comprehensive study on the application of clustering techniques for power systems. It also lays the groundwork for the successive chapters in which understanding the load data clustering process and interpreting its outcome are necessary.

There are several features that distinguish the current work in this chapter from the previous publications [2], [90], [32], mainly:

- Five major clustering techniques are introduced and the effects of their different parameters for load pattern clustering are analyzed. Besides the well-studied clustering methods such as K-means, fuzzy c-means and hierarchical algorithms, clustering with the probabilistic and generative models and self-organizing maps (SOM) are also discussed.
- These clustering techniques and their applications in customer segmentation are compared.
- The results are analyzed to identify the main consumption patterns of customers.

The following sections are organized as follows. Section 3.2 introduces the major clustering techniques and their theoretical concepts. Section 3.3 discusses the clustering parameters and their effects on clustering outcomes. Analysis of the methods and their comparisons are carried out in Section 3.4 for two different datasets, one containing the yearly load data of only one customer and the other one comprising of the RLPs of all the customers for two different loading conditions.

3.2 Clustering Algorithms

In the following, the major clustering techniques that will be studied in this thesis are briefly illustrated.

3.2.1 Distance-based methods

Distance-based methods are the most popular clustering algorithms since they are generally fast and easy to implement. These algorithms use "similarity (dissimilarity) measures" to construct the clusters. As the main purpose of clustering is to group similar instances, defining proper measures that can express numerically the degree to which two objects are similar to or dissimilar from each other is required. The main types of similarity measures used in the literature can be categorized as [11], [149]: 1) difference (distance)-based measures such as Minkowski distance (L_P-norm distance), Canberra distance, and Gower's coefficient, and 2) Correlation-based measures (similarity functions) such as cosine measure and Pearson's correlation measure.

Here, we confine the discussion to Minkowski measures as they are the most common measure used in the power system literature. These similarity measures try to calculate a

distance value based on the differences between the features (attributes) of the two compared objects. If two load curves x_i and x_j are represented by h recordings, the Minkowski distance of order p between them can be calculated as follows:

$$d_{Mink,p} = \left(\left|x_{i,1} - x_{j,1}\right|^p + \left|x_{i,2} - x_{j,2}\right|^p + \cdots + \left|x_{i,h} - x_{j,h}\right|^p\right)^{1/p} \quad (3.1)$$

For $p = 1$ and $p = 2$ the L_P-norm distance is usually called the Manhattan distance (or city block distance) and Euclidean distance, respectively. Euclidean distance is by far the most widely used dissimilarity measure.

Two well-known and frequently used distance-based clustering algorithms are partitioning methods and hierarchical clustering methods which are well presented in the data mining literature.

A) K-centers family

K-centers family including K-means, K-medians, and K-medoids are the most widely used partitioning clustering techniques. They do not create a tree structure to describe the groupings of data, but rather create a single level of clusters. They share the same basic operation principle which is outlined in Algorithm 1 [150], [151].

Algorithm 1. K-centers clustering

Require: Number of clusters and cluster centers as follows:

- The number of clusters is predetermined (***k*** clusters)
- ***k*** points are selected as the initial cluster centers.

Repeat:

1- Assign each instance to the closest center until ***k*** clusters are formed.

2- Recompute the center of each cluster based on all instances that belong to it.

Until: The convergence criterion is met.

K-means is a commonly used algorithm, which minimizes the square-error function, defined as:

$$E = \sum_{k=1}^{K} \sum_{x \in C_k} |x - c_k|^2 \tag{3.2}$$

where K is the number of clusters and c_k is the center of kth cluster donated by C_k.

Fuzzy c-means (FCM) is another popular method of K-centers family [152]. It is similar to K-means clustering, but each instance has a grade of membership to each cluster [33]. FCM minimizes the following objective function:

$$J_m = \sum_{l=1}^{N} \sum_{k=1}^{K} {\mu_{lk}}^m \|x_l - c_k\|^2 \tag{3.3}$$

where N is the number of load curves (observations), μ_{lk} is the degree of membership of lth load curve in kth cluster, and m is the parameter that controls the amount of fuzziness.

In fuzzy clustering, each load curve does not belong to only one cluster. Instead, the degree of membership determines the amount of membership of each load curve to each cluster, where:

$$\sum_{k=1}^{K} \mu_{lk} = 1 \tag{3.4}$$

An observation is assigned to the cluster to which it has the maximum value of membership degree [69]. The membership degrees are updated in each step as:

$$\mu_{lk} = \left[\sum_{j=1}^{K} \left[\frac{\|x_l - c_k\|}{\|x_l - c_j\|} \right]^{\frac{2}{m-1}} \right]^{-1} \tag{3.5}$$

Fuzzy overlap refers to how fuzzy the boundaries between clusters are and can take a value above 1. The higher values of this parameter will result in fuzzier clusters.

B) Hierarchical clustering

Hierarchical clustering is a more flexible and deterministic algorithm than K-centers method. The hierarchical algorithm produces a tree or dendrogram by either agglomerative (bottom-up) or divisive (top-down) methods. In the agglomerative method, initially each instance is classified as a cluster and then clusters are merged iteratively to build a bottom-up hierarchy of the clusters until a single root cluster is reached. The divisive approach, on the other hand, starts with a single root cluster and splits it into subclusters continuously, generating a top-down hierarchy of clusters. Fig. 3.1 displays the hierarchical tree or dendrogram of an agglomerative clustering method.

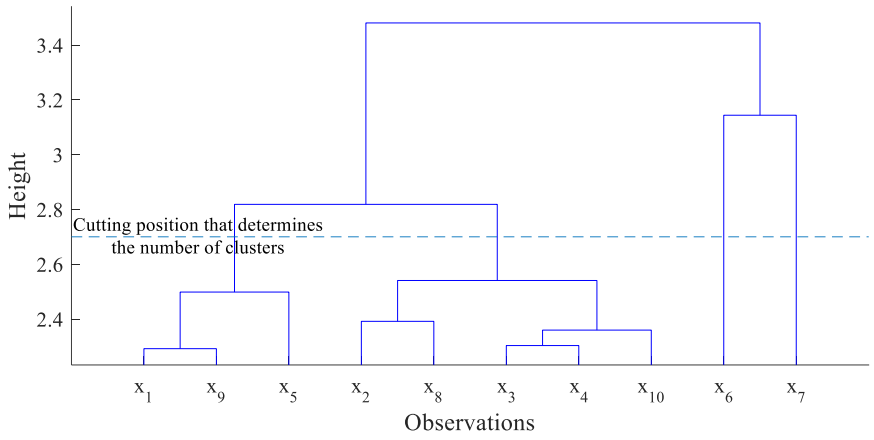


Fig. 3.1 Dendrogram formed by a hierarchical clustering method

This formed hierarchy can be cut at any given level which allows obtaining the corresponding clusters. This is the main advantage of hierarchical algorithms that makes them considerably different from partitioning methods which require the number of clusters before starting the algorithm. Also, hierarchical clustering has fewer assumptions about the

distribution of data. However, it should be noted that hierarchical clustering is generally more computationally expensive than K-means (time complexity of $O(n^3)$ where n is the number of observations compared to the linear complexity of K-means).

3.2.2 Self-Organizing Map

SOM is an unsupervised artificial neural network that projects the original input space to a reduced output space [33]. It produces a graphical representation of the data which allows an easy evaluation of the results and grouping them into clusters by visual inspection [60] [153]. The SOM consists of a grid containing $W_1 \times W_2$ map units (neurons). The original h-dimensional data vector is transformed to a (typically) bi-dimensional space where similar observations in the input space are mapped into nearby units. Each unit i is represented by a prototype vector $w_i = [w_{i1}, w_{i2}, \dots, w_{ih}]$, which has the same dimension of input vectors (h). The number of units can vary from a few dozen up to several thousand [153]. Each unit is connected to adjacent units by a neighbourhood relation, which determines the topology or structure of the map.

The SOM is trained iteratively. In each training step, a sample vector x from the data set is picked out randomly. The distance of this vector and all prototype vectors are calculated and the unit whose prototype vector is closest to x is selected as the best-matching unit (BMU) or winning unit:

$$\|x - w_b\| = \min_i \|x - w_i\| \tag{3.6}$$

The learning algorithm updates the weight of the winning unit and also the weights of its adjacent units. The prototype vector of unit i is updated using the following equation:

$$w_i(t+1) = w_i(t) + \alpha(t)h_{bi}(t)[x(t) - w_i(t)] \quad (3.7)$$

where t represents time, $\alpha(t)$ is the learning rate or adaptation coefficient at time t, and $h_{bi}(t)$is the neighborhood kernel (neighborhood symmetrical function) around the winner unit b. The units that are topologically close to the winning unit b are activated using $h_{bi}(t)$:

$$h_{bi}(t) = \exp\left(-\frac{\|r_b - r_i\|^2}{2{\sigma_t}^2(t)}\right) \quad (3.8)$$

where r_i represents the coordinates of unit i in the SOM grid and $\sigma(t)$ is the neighborhood radius function. Both $\alpha(t)$ and $\sigma(t)$ decrease monotonically with time. Therefore, the neighborhood size of each unit reduces in each training step and, finally, it ends with a single unit.

3.2.3 Probabilistic and generative models

In the model-based clustering, it is assumed that instances arise from a distribution that is a mixture of several components. The problem is to estimate the parameters of each component (i.e., cluster) and identify which component produced each observation [150]. This process leads to the clustering of the data. In practice, the attention is mostly paid to parametric mixture models, where all the components are from the same family of distributions. Gaussian (normal) distributions are by far the most commonly used representation in the model-based clustering. In this case, the mixture model is the Gaussian mixture model (GMM), where components are Gaussian distributions with different means and variances. The mathematical formulations of GMM are illustrated in the following [154], [155], [156].

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ be a set of n observations. The variable $\mathbf{x}_i$ is assumed to be distributed according to a mixture of K components. The probability density function (or mixture distribution) of $\mathbf{x}_i$ can be written as:

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{m=1}^{K} \alpha_m \, p(\mathbf{x}_i|\boldsymbol{\theta_m}) \tag{3.9}$$

Where $\alpha_1, \dots, \alpha_K$ are the *mixing probabilities* and each θ_m is the set of parameters which define the mth component. $\boldsymbol{\theta} = \{\boldsymbol{\theta_1}, \dots, \boldsymbol{\theta_K}, \alpha_1, \dots, \alpha_K\}$ is the complete set of parameters. The α_m must satisfy:

$$\alpha_m \geq 0, \;\; m = 1, \dots, K, \qquad and \qquad \sum_{m=1}^{K} \alpha_m = 1 \tag{3.10}$$

In the Gaussian mixture model, each component is specified by the parameters of a multivariate Gaussian distribution:

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = p(\mathbf{x}_i|\alpha, \mu, \Sigma) = \sum_{m=1}^{K} \alpha_m \, \mathcal{N}(\mathbf{x}_i|\mu_m, \Sigma_m) \tag{3.11}$$

where,

$$\mathcal{N}(\mathbf{x}_i|\mu_m, \Sigma_m) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x_i} - \mu_m)^T \Sigma_m^{-1} (\boldsymbol{x_i} - \mu_m)\right\}}{(2\pi)^{\frac{D}{2}} \, |\Sigma_m|^{\frac{1}{2}}} \tag{3.12}$$

For a D-Dimensional vector $\mathbf{x}$, μ and Σ are the D-dimensional mean vector and the D × D covariance matrix respectively. In the case of a single variable $\mathbf{x}$, (3.12) is reduced to:

$$\mathcal{N}(\mathbf{x}_i|\mu_m, \sigma_m^2) = \frac{\exp\left\{-\frac{1}{2\sigma_m^2}(\mathbf{x_i} - \mu_m)^2\right\}}{(2\pi\sigma_m^2)^{\frac{1}{2}}} \tag{3.13}$$

where, μ and σ^2 are the mean and variance respectively.

Usually, Expectation-Maximization (EM) algorithm is utilized for parameter estimation of the model. Bayesian information criterion (BIC) and Akaike's information criterion (AIC) are the main criteria for choosing the best number of components (clusters) [27].

3.3 Discussion on the Algorithms

Each of the presented methods for the clustering has its advantages and disadvantages. In addition, different considerations need to be taken into account before applying the clustering algorithms. In this section the affecting parameters of each of the clustering methods are discussed.

3.3.1 K-center family

For K-centers methods various parameters including the number of clusters, initial centres, and the dissimilarity measure must be initially determined. Each of these parameters can affect the final outcomes of the clustering. Initial centers can be selected by a random fashion among the instances of the data set [33]. The random selection of cluster centers may affect the final cluster formations. CVI measures can be used to find the best choice for multiple runs of the clustering algorithm with different random initial centers [11]. In addition, numerous initialization methods are also proposed for the selection of the centers. Ref. [157] provides a thorough study of various initialization methods and compares their performance for real and synthetic datasets. Table 3-1 compares the main characteristics of the three main K-centers methods.

Table 3-1 Characteristics of main methods of K-centers family

Method	Calculation of center	Best dissimilarity measure	Disadvantages	Advantages
K-means	Center is calculated as the mean of members of the cluster	Euclidean	Not applicable to discrete attributes; Handling the data containing outliers; Handling asymmetrically distributed data. Cluster centers might not be similar to any instance.	Easy to implement and efficient.
K-medians	Center is selected as the median of members of the cluster	Manhattan	More costly to calculate; Cluster centers might not be similar to any instance.	More robust to asymmetric distributions and outliers; Not skewed so much by extremely large or small values.
K-medoids	Center is the cluster member that is the least dissimilar to other cluster members, on the average	Different measures can be used	More expensive computationally than K-means and K-medians.	Robust with respect to noise and outliers; Guarantees convergence.

For FCM, the degree of fuzzy overlap needs to be decided. The selection of parameter m has been the subject of many studies in the data science literature [158] [159] [160]. These studies follow different approaches for the selection of optimal m and suggest different values and ranges for that. For example, [160] and [161] propose the selection of m from the range of [1.5, 4] and [1.5, 2.5] respectively. The most frequently used and accepted value in various applications is m=2 [161] [162] which is also the suggested value in MATLAB software. In this thesis, the value of m is selected based on the results of CVIs. FCM computation time is longer compared to K-means since the degrees of membership need to be updated at each step.

3.3.2 Hierarchical

In almost all the studies in the power system domain, the agglomerative approach is used as the preferred hierarchical method. For agglomerative methods, the formation of clusters is based on the similarity measures. Firstly, using a distance criterion, a similarity

matrix D is built in which $d_{ij} \in D$ shows the distance between the observation i and the observation j. In the next step, based on this similarity matrix, instances are grouped into clusters using a linkage criterion. The linkage is an evaluation function which indicates the best candidates for merging. Therefore, at each level the closest sets of clusters are merged until the final cluster (which contains all the observations) is obtained.

Some of the most important linkage criteria and their features are reported in Table 3-2 [11], [150], [156], [163].

Table 3-2 Linkage criteria for hierarchical clustering

$\boldsymbol{C_i}$: ith cluster; c_i: center of cluster $\boldsymbol{C_i}$; n_i : number of data points belonging to cluster i; $d(x,y) =$ distance between the objects x and y		
Linkage criterion	**Description**	**Features**
Single	$\min_{\substack{x\in C_i \\ y\in C_j}} d(x,y)$	Neglects the overall cluster structure Sensitive to noise and outliers Capable of clustering non-elliptical shaped groups of data points Not affected by the monotone transformations (like the logarithmic transformation) of the original data
Complete	$\max_{\substack{x\in C_i \\ y\in C_j}} d(x,y)$	Obtains more compact shaped clusters Sensitive to outliers
Average	$\frac{1}{n_i . n_j} \sum_{\substack{x\in C_i \\ y\in C_j}} d(x,y)$	A compromise between single and complete linkages Computationally expensive, especially for large datasets Noise resistant
Centroid	$d(c_i, c_j)$	Does not have monotonic property i.e. a merged cluster might become closer to other clusters than its descendants which is usually undesirable.
Ward	$\sqrt{\frac{n_i . n_j}{n_i + n_j}} d(c_i, c_j)$	Not directly based on similarities between data points of the two clusters, instead works based on an objective function

In single linkage, the similarity of two clusters is determined based on the similarity between their most similar members. On the other hand, in complete linkage the similarity

of two clusters is measured as the similarity of their most dissimilar members. Average linkage method (sometimes called UPGMA which stands for "unweighted pair group method using arithmetic averages") diminishes the problems associated with single and complete linkage methods by considering the similarity between all pairs of instances present in both of the clusters. So, the average dissimilarity between instances from two clusters serves as the dissimilarity between the clusters. Another method called centroid linkage clustering computes the dissimilarity between the center for cluster i and the center for cluster j. In addition, linkages can be defined based on a specific quality criterion or objective function. The most famous one among these linkage methods is Ward criterion with the objective to minimize the total sum of squared dissimilarities between cluster members and cluster centers for all the clusters. In other words, for every two clusters C_i and C_j, Wards' criterion measures the increase in the value of sum of squared errors for the clustering obtained by merging them into $C_i \cup C_j$.

Likewise the dissimilarity measure, the choice of linkage can also have a significant impact on the final clustering outcomes.

3.3.3 SOM

The parameters of neural network such as the learning rate and the radius of the neighbourhood might slightly affect the partitioning of the data set by SOM. In addition, the SOM results depend on the population of neurons as well as the topology or structure of the map. For N data points, the number of neurons is recommended to be between $5 \times \sqrt{\mathrm{N}}$ to $20 \times \sqrt{\mathrm{N}}$ [60]. In addition, different topologies can be selected for the neural lattice.

Traditionally, hexagonal or rectangular arrangements of neurons are chosen, in them internal neurons are bounded by six and four adjacent neurons respectively.

A visual inspection of the generated SOM map can give an initial idea of the number of clusters. This is particularly performed using a the unified distance matrix (called U-matrix) that shows the distances between prototype vectors of adjacent units and can visualize the cluster structure of the SOM. However, this process does not guarantee the best results. Sometimes a two-level approach is used in which the prototypes are formed using the SOM and then, a clustering algorithm is applied on the prototypes to obtain the final clusters [61] [56]. This is especially beneficial when the data set contains a large number of data points.

3.3.4 GMM

GMM is able to model both continuous and categorical data which is an advantage of this method over many clustering techniques such as K-means [27]. Interested readers are referred to [164] for technical explanations and examples of applying mixture models on mixed continuous and categorical variables. GMM requires the number of clusters to be specified before fitting the model. In addition, for applying GMM, the parameters of covariance matrix of each component need to be specified. The structure of the covariance defines the shape of a confidence ellipsoid over a cluster. The detailed technical discussion of the various covariance structures is beyond the scope of this paper. Interested readers are referred to [165] and [166] for practical implementations in Matlab and R programs, respectively. In this thesis, different configurations of covariance matrices are examined and their effects on clustering results are studied. Firstly, two different structures for covariance matrices, which specify the cases with correlated and uncorrelated predictors, are

considered. In the case studies, the former and latter cases are denoted as full and diagonal, respectively. Secondly, the effects of shared or unshared covariance matrices among all components are investigated. Each combination of these parameters defines the orientation and shape of ellipsoids. Since the appropriate covariance structure and number of components (clusters) are not known, the information criteria like AIC or BIC are used to compare different models. Lower values of AIC and BIC indicates better models with the most suitable parameters or the best number of components.

Furthermore, EM algorithm that fits the GMM is sensitive to initial conditions and might converge to a local optimum. To ensure global convergence is achieved, the algorithm can be run repeatedly with different initial conditions [27]. The initial component parameters can be decided in various ways, for example, in a random fashion or by applying a k-means clustering to choose a number of observations [167].

3.4 Application of Clustering Algorithms to the Load Curves of Customers

In this section, the impacts of discussed parameters of presented algorithms on clustering of daily loads curves of electricity customers are discussed.

In the following, most of the case studies are carried out for clustering of 356 daily load curves of one residential customer (sections 3.4.2 to 3.4.6). In addition, to investigate the clustering of a large number of users, clustering techniques are also applied on a data set comprising more than 4000 customers (section 3.4.7).

Since the aim is to cluster the daily load curves based on their shapes, each daily load curve is normalized based on the maximum consumption of that day. Without normalization

of daily curves, the resulted clusters will only reflect load magnitudes. The majority of clustering studies mostly focus on the shape of load curves. The effect of customers' consumption values can be examined with other methods. For instance, [54] clusters the customers based on their load shapes. It also fits a mixture of log-normal distributions to the daily consumption values of customers and based on that, divides them into heavy, moderate and light energy users.

In the following, firstly, the effect of different parameters of each algorithm is investigated and then, the performances of clustering methods are compared.

3.4.1 Cluster validity indexes

As explained in the preceding sections, parameters of each clustering method and initial conditions affect the final results and hence, clustering outcomes should be evaluated considering a range of parameters and conditions. CVIs can be used to study various aspects of clustering results and to compare the methods. In the electricity customer categorization, the CVIs may be used for different purposes, mainly:

- To determine the suitable number of customer clusters [23], [61]: Many clustering methods require that the number of clusters be specified by the user. Therefore, there is a need for the criteria to determine the best number of clusters. In this regard, clustering process can be repeated for different pre-set number of clusters, and based on the values of CVIs for each case, the best case can be selected.
- To compare the performance of different clustering techniques [32], [57]: Here, the same data set is grouped into classes by using different clustering algorithms and the

role of CVIs is to select the best clustering results. Also in some studies, they are used to assess the performance of dimensionality reduction methods [168].

- To investigate the effect of method parameters on clustering results [169], [170]: The variants and parameters of each clustering technique can have a significant effect on the final clustering outcomes.
- To evaluate the performance of clustering when some attributes (features) are added or removed [79], [171]: this is important in the process of feature selection since it is desirable to have a set of features that can include all the necessary information of the consumption pattern and provide a sound basis for comparisons.

Here, the comparison of parameters and methods are conducted based on 6 different CVIs: mean square error (MSE), Silhouette index (SIL), Davies-Bouldin indicator (DBI), mean index adequacy (MIA), the ratio of within-cluster sum of squares to between-cluster variation (WCBCR), and Dunn index. The definitions of these CVIs are given in Table 3-3. The rule in this table refers to the interpretation of the CVIs for choosing among the results. For example, the minimum value of DBI indicates the best result. The DBI, SIL, Dunn, and WCBCR indexes are explained in more detail in the following. Besides these CVIs, AIC is utilized for evaluating the results of GMM method.

Table 3-3 List of CVIs

N: Number of observations (load curves); K: number of clusters; c_i: center of cluster i; $d(x,y)$ = distance between the objects x and y; n_i: number of data points belonging to cluster i		
Cluster validity index	**Descriptions**	**Rule**
$MSE = \frac{1}{N}\left(\sum_{k=1}^{K}\sum_{x_i \in C_k} d^2\,(x_i, c_k)\right)$		min
$SIL = \frac{1}{K}\sum_{i=1}^{K} \mathcal{s}_i$, where, $\mathcal{s}_i = \frac{1}{n_i}\sum_{x_j \in C_i} s(j)$	$s(i) = \frac{\mathcal{B}(i)-\mathcal{A}(i)}{\max(\mathcal{B}(i),\mathcal{A}(i))}$, where, $\mathcal{A}(i) = \text{within} - \text{cluster mean distance} = \frac{1}{n_k - 1}\sum_{\substack{j \in C_k \\ j \neq i}} d(x_i, x_j)$, $\mathcal{B}(i)$ = the smallest of mean distances to other clusters $= \min_{k' \neq k}\left(\frac{1}{n_{k'}}\sum_{j \in C_{k'}} d(x_i, x_j)\right)$	max
$DBI = \frac{1}{K}\sum_{i=1}^{K} max_{j \neq i}\left\{\frac{\left[\frac{1}{n_i}\sum_{x \in C_i} d(x, c_i) + \frac{1}{n_j}\sum_{x \in C_j} d(x, c_j)\right]}{d(c_i, c_j)}\right\}$		min
$MIA = \sqrt{\frac{1}{K}\sum_{k=1}^{K} {d_{C_k}}^2}$	d_{C_k} = the distance between cluster center c_i and the member of the cluster i= $\sqrt{\frac{1}{n_k}\sum_{x_i \in C_k} d^2(x_i, c_k)}$	min
$WCBCR = \frac{\sum_{k=1}^{K}\sum_{x_i \in C_k} d^2(x_i, c_k)}{\sum_{1 \leq i < k}^{K} d^2(c_i, c_k)}$		min
$Dunn = \frac{\min\limits_{i \neq j} d_{ij}}{\max\limits_{i} D_i}$	d_{ij}: the distance between the closest instances of two clusters (separation) D_i: the largest distance between two instances that belong to the cluster i (diameter)	max

Let us consider the results of a clustering process that leads to the formation of K clusters $C_1, C_2, \dots, C_K$ with the cluster centers and the number of instances in each cluster denoted by $c_1, c_2, \dots, c_K$ and $n_1, n_2, \dots, n_K$, respectively. Now, a set of distances can be defined accordingly:

- Distance between two instances: $d(x_i, x_j)$
- Distance between an instance and a cluster center: $d(x_i, c_i)$
- The mean distance of the instances belonging to the cluster i to their corresponding center:

$$\delta_i = \frac{1}{n_i} \sum_{x \in C_i} d(x, c_i)$$

- Distance between two cluster centers: $\Delta_{i,j} = d(c_i, c_j)$
- The distance between the closest instances of two clusters (separation): d_{ij}
- The largest distance between two instances that belong to the cluster i (diameter): D_i

Based on these definitions, the most important CVI measures are illustrated in the following.

Davies-Bouldin indicator (DBI):

$$DBI = \frac{1}{K} \sum_{i=1}^{K} max_{j \neq i} \left(\frac{\delta_i + \delta_j}{\Delta_{i,j}} \right)$$

$$= \frac{1}{K} \sum_{i=1}^{K} max_{j \neq i} \left\{ \left[\frac{1}{n_i} \sum_{x \in C_i} d(x, c_i) + \frac{1}{n_j} \sum_{x \in C_j} d(x, c_j) \right] / d(c_i, c_j) \right\} \quad (3.14)$$

To calculate this CVI, firstly, the values of $\left((\delta_i + \delta_j)/\Delta_{i,j}\right)$ between a cluster and all other clusters are calculated and the maximum amount of these values is selected for each cluster. Then, the average of all these maximum values is selected as DBI index. The lower DBI values indicate the better clustering results.

Dunn Index: Let us denote by d_{min} the minimal distance between points of different clusters and d_{max} the largest within-cluster distance. So, d_{min} and d_{max} can be written as:

$$d_{min} = \min_{i \neq j} d_{ij} \tag{3.15}$$

$$d_{max} = \max_{i} D_i \tag{3.16}$$

Now, Dunn index can be defined as:

$$Dunn = \frac{d_{min}}{d_{max}} \tag{3.17}$$

Higher Dunn index values correspond to the more preferred results as they correspond to inter-cluster dissimilarities that are large in comparison to intra-cluster dissimilarities [11].

SIL (Silhouette Index): Generally, three Silhouette measures can be defined: Silhouette width for an instance, cluster mean Silhouette, and global mean Silhouette over all the clusters [172].

For calculation of Silhouette width, we first need to explain two other distances: within-cluster mean distance, $\mathcal{A}(i)$, and the smallest of mean distances to other clusters, $\mathcal{B}(i)$, which are defined for an instance $x_i \in C_k$ as the following:

$$\mathcal{A}(i) = \frac{1}{n_k - 1} \sum_{\substack{x_j \in C_k \\ j \neq i}} d(x_i, x_j) \tag{3.18}$$

$$\mathcal{B}(i) = \min_{k' \neq k}\left(\frac{1}{n_{k'}} \sum_{x_j \in C_{k'}} d(x_i, x_j)\right) \tag{3.19}$$

The Silhouette width of instance x_i can then be defined as:

$$s(i) = \frac{\mathcal{B}(i) - \mathcal{A}(i)}{\max\left(\mathcal{B}(i), \mathcal{A}(i)\right)} \tag{3.20}$$

The cluster mean Silhouette, which is defined for a cluster C_i, is calculated as:

$$\mathcal{s}_i = \frac{1}{n_i} \sum_{x_j \in C_i} s(j) \tag{3.21}$$

Finally, global Silhouette index can be defined as the mean of the mean Silhouettes through all the clusters:

$$SI = \frac{1}{K} \sum_{i=1}^{K} \mathcal{s}_i \tag{3.22}$$

The Silhouette width of an instance is a number between -1 and 1, with those values approaching 1 indicating that the instance is very well placed in the right cluster while negative values near -1 displaying that the instance should be assigned to another cluster.

WCBCR (the ratio of within-cluster sum of squares to between-cluster variation): Intuitively, clusters formed by clustering process should have these two features: they should be compact and they should be as far from each other as possible. These notions can be mathematically expressed by these definitions: within-cluster variation (*WC*) and between cluster variation (*BC*) [173] as follows:

$$WC = \sum_{k=1}^{K} \sum_{x_i \in C_k} d^2(x_i, c_k) \tag{3.23}$$

$$BC = \sum_{1 \le i < k}^{K} d^2(c_i, c_k) \tag{3.24}$$

Finally, WCBCR index is constructed by the combination of these two measures:

$$WCBCR = \frac{WC}{BC} = \sum_{k=1}^{K} \sum_{x_i \in C_k} d^2(x_i, c_k) \Big/ \sum_{1 \le i < k}^{K} d^2(c_i, c_k) \tag{3.25}$$

3.4.2 Fuzzy c-means

For FCM, the main parameter is the fuzziness degree, characterized by parameter m in Eq. 3.3. Fig. 3.2 shows the effect of this parameter on clustering results where the number of clusters is fixed at 10. The value of m is changed from 1.05 to 4 at the steps of 0.05. Since the initial centers are selected randomly, the clustering results slightly change in each execution of the method. Thus, for each value of fuzziness degree, the clustering is carried out ten different times and the outcomes are averaged for each CVI. As this figure shows, the CVIs indicate that the best results happen at around 1.9 to 2.

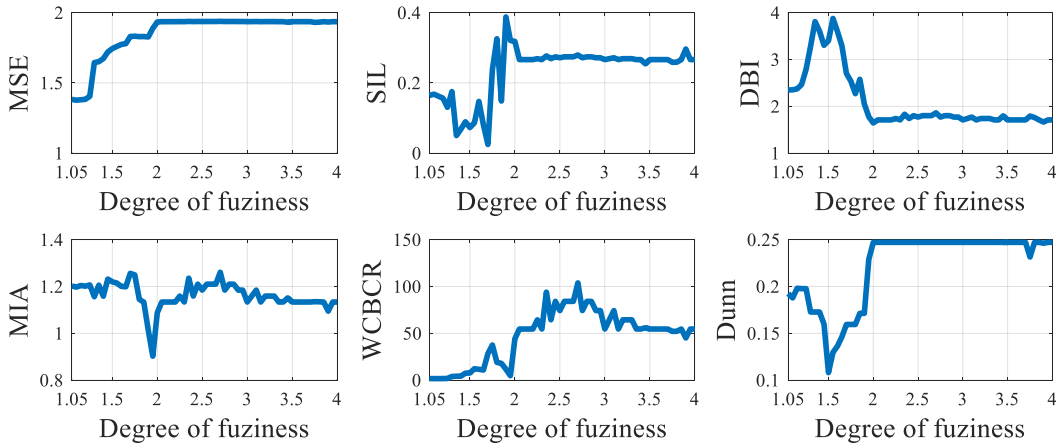


Fig. 3.2 Effect of fuzziness degree on the clustering results for FCM method

3.4.3 Hierarchical clustering

Here, 5 different hierarchical methods with different linkage criteria are compared for varying number of clusters and the results are displayed in Fig. 3.3. The single and centroid linkage models are selected as the best models by all CVIs except for MSE. Ward linkage also shows good performance having relatively low values for DBI, MIA, and WCBCR, and high values for SIL and Dunn. However, further inspection of the clusters shows that single and centroid methods assign most of the daily load curves to only one cluster. It can be observed by the dendrograms of the single and ward methods as shown in Fig. 3.4. For this reason, ward method which well separates load curves into different clusters is preferred.

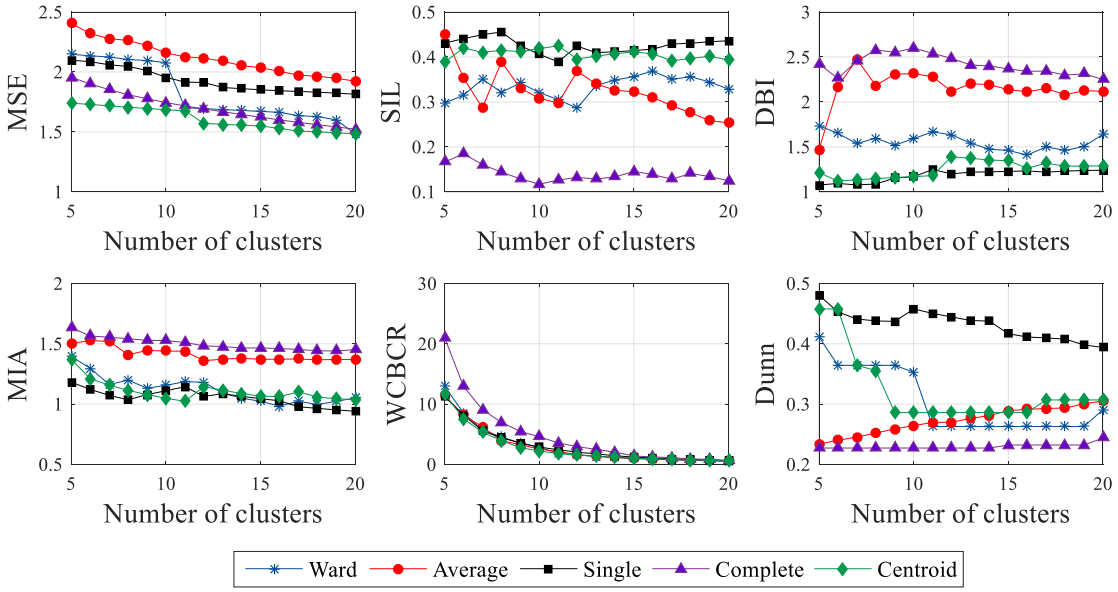


Fig. 3.3. Comparison of hierarchical algorithms

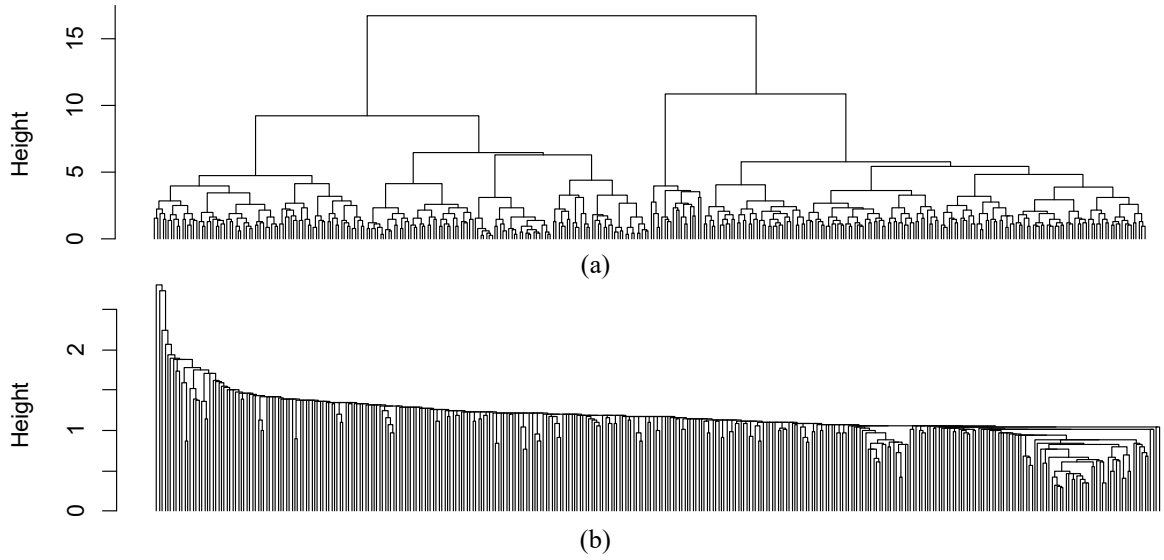


Fig. 3.4. Dendrograms of (a) ward method and (b) single method

3.4.4 SOM

To cluster the load curves, the SOM in conjunction with a hierarchical clustering method is used. The parameters under study are the neurons population and topology of the

neural lattice. For 356 load patterns, the grid size changed accordingly from 10×10 ($5 \times \sqrt{356} \approx 94$) to 20×20 ($20 \times \sqrt{356} \approx 377$). The width, W_1, and height, W_2, of the grid are assumed to have the equal size. For each grid size, the effects of hexagonal and rectangular topologies are studied.

Fig. 3.5 displays a sample 16×16 SOM grid which is divided into 10 clusters after applying the hierarchical algorithm. To compare different configurations, the values of CVI indexes are calculated in each case as shown in Fig. 3.6. In this case, superior results can be observed for the grid size 18×18 with hexagonal topology.

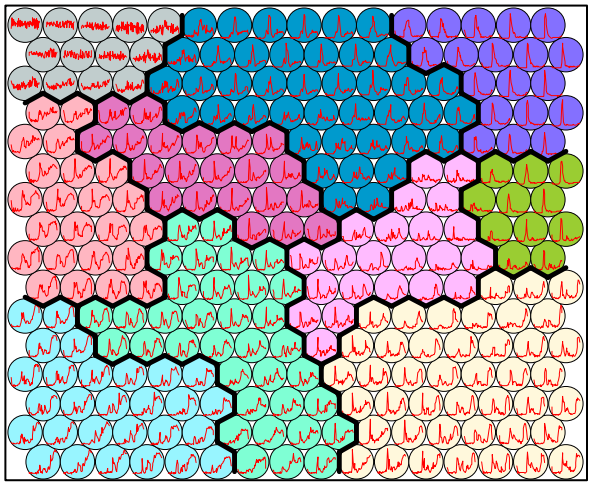


Fig. 3.5. A 16×16 SOM grid and the corresponding clusters after applying the hierarchical method

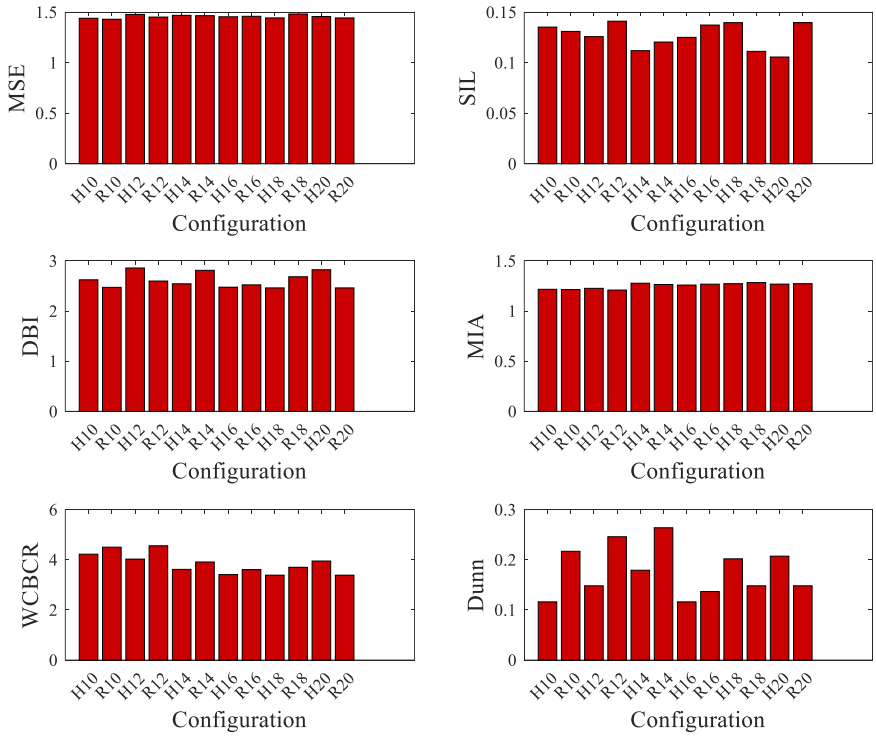


Fig. 3.6. Effect of grid size and topology on the two-level clustering of load curves using SOM and hierarchical method (R: Rectangular, H: Hexagonal)

3.4.5 GMM

Generally, GMM produces the best results when the number of variables is limited. Since the load curves have 48 variables (half-hour recordings) applying GMM might not lead to the promising results. To this end, prior to GMM clustering, use of an indirect clustering approach could be advantageous. For instance, in [27] the authors apply GMM on a set of features which are extracted from the load data.

Here, the principal component analysis (PCA) is used to reduce the size of the input data. Therefore, each load pattern is represented by a limited number of components. The selection of the best number of components is described in the next sections. GMM is applied to the PCA components and the impacts of parameters of covariance matrix (full vs.

diagonal and shared vs. unshared matrixes) on final results are investigated using AIC as depicted in Fig. 3.7. The lowest values of AIC occur for full-unshared method. By increasing the number of clusters, diagonal-shared and full-shared observe a decreasing trend while the AIC for full-unshared increases gradually. The results also suggest the best segmentation with 8 clusters.

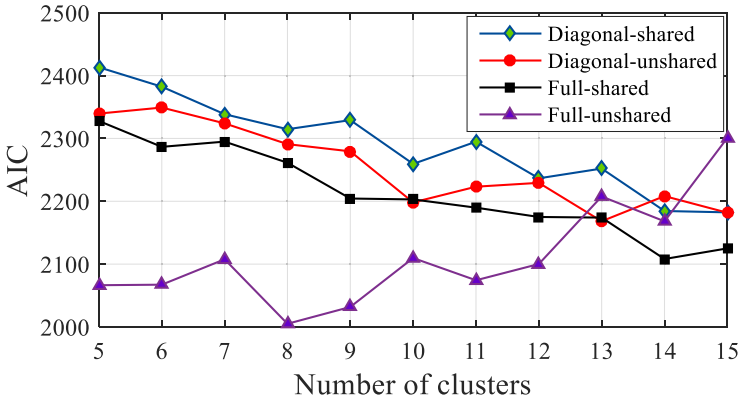


Fig. 3.7. Effects of parameters of covariance matrix on the GMM clustering

3.4.6 Comparing clustering methods

In the previous sections, some of the most important parameters of different clustering methods are illustrated and their effects on clustering results were analyzed. In this section, four major clustering algorithms including K-mean, FCM, hierarchical, and SOM are compared and the formed clusters are analyzed. The final aim is to determine the algorithm which can better reveal the various patterns of consumption behavior and form clusters that are more compact and well separated from each other.

The parameters of the methods are selected based on the analysis in previous sections. Correspondingly, based on the results of section 3.4.2 to 3.4.5, the fuzziness degree is set to 1.9 for FCM, hierarchical clustering with ward linkage is chosen, and SOM is performed for

a grid size of 18×18 with hexagonal topology. Fig. 3.8 shows the CVIs for the selected algorithms.

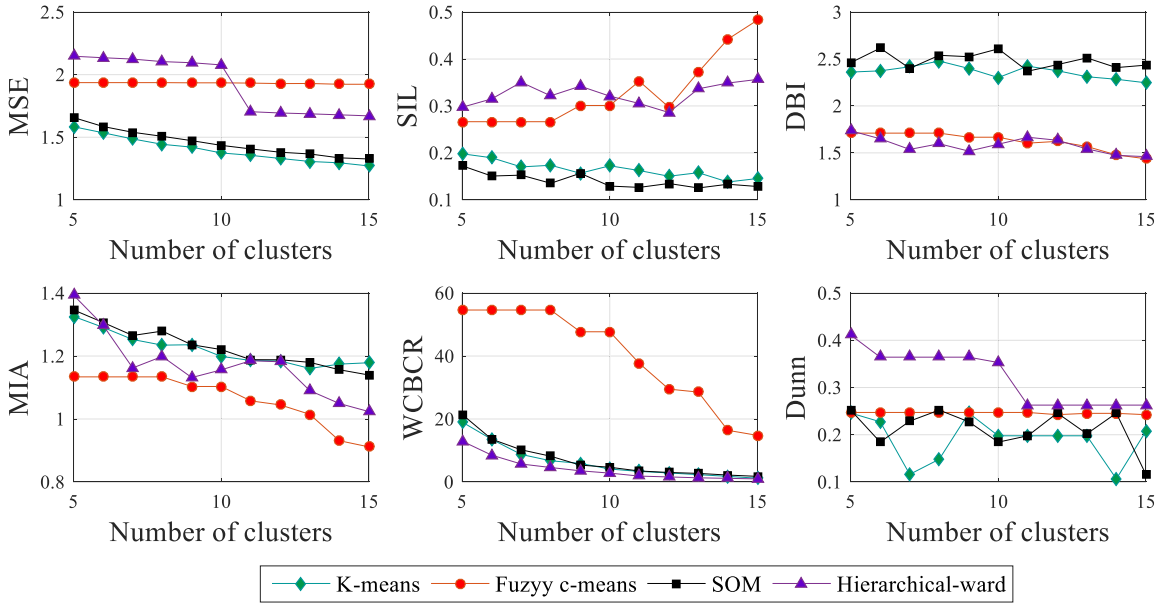


Fig. 3.8. The CVI values for four different clustering algorithms

It can be observed that hierarchical algorithm shows superior results for this special data set which contains the daily load patterns of a customer. In fact, the highest values for SIL and Dunn measures and the lowest values for DBI and WCBCR are obtained for this algorithm. Furthermore, five CVIs indicate SOM as the worst clustering algorithm for this case study. It can be seen from the related curves of SIL (lowest values) and DBI and MIA (highest values).

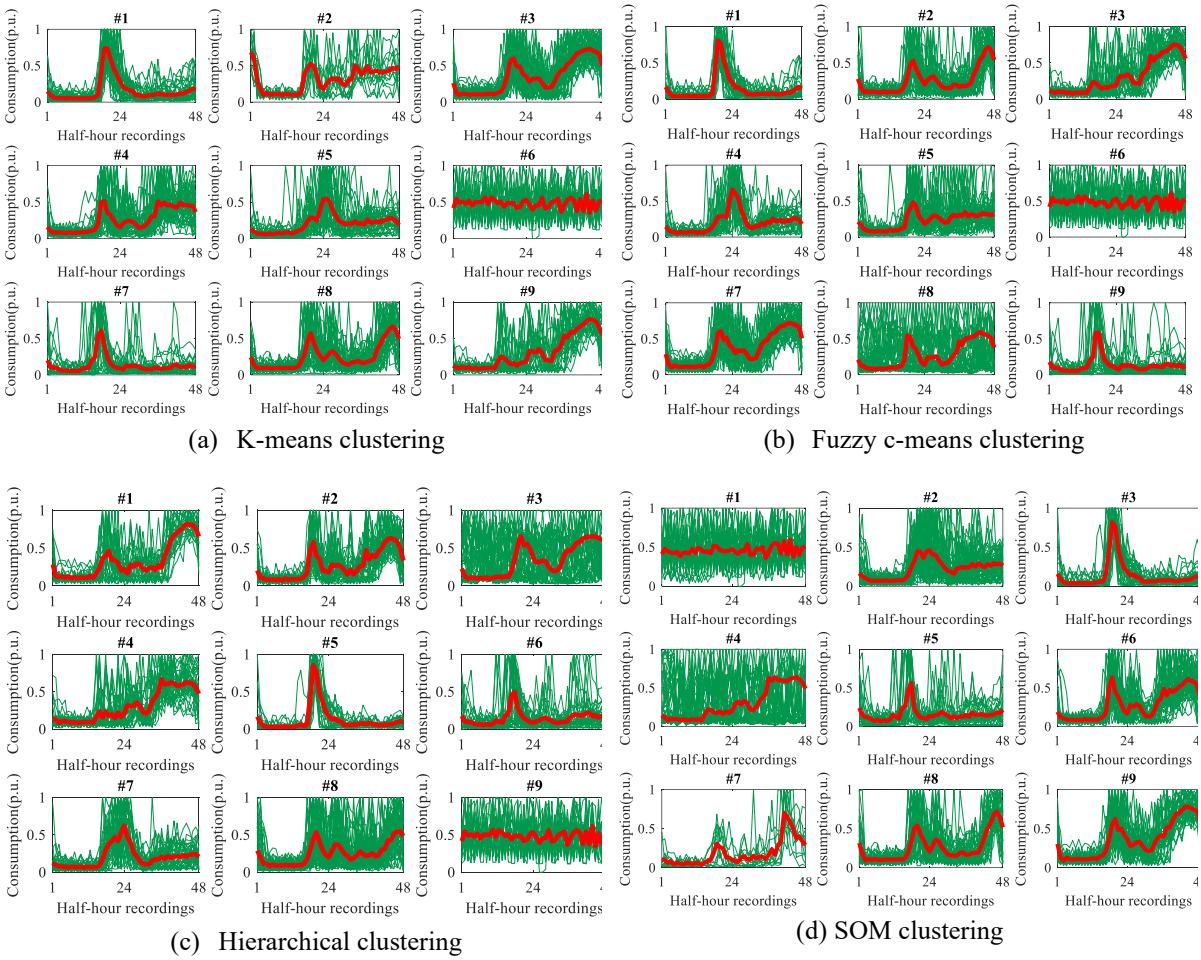
For this specific value of fuzziness degree, the obtained results, for example, for SIL, DBI, and MIA, indicate the good performance for fuzzy clustering. However, it should be noted that in some cases, the results of FCM are sensitive to small changes in degree of fuzziness. Therefore, while using FCM clustering, it is necessary to study various fuzziness degrees for different number of clusters.

Based on the CVI values for all the algorithms, it can be seen that the optimum number of clusters falls into the range of 8 to 10 clusters. Eight clusters appear to produces the satisfactory results since adding more clusters does not improve the results significantly. This finding is in accordance with the GMM clustering outcomes. Generally, the final number of clusters is decided based on the pre-defined objectives and needs. In practice, the outcomes of electricity customer clustering will be used by the utilities for improving different applications such as demand response programs and tariff design. Therefore, typically, the number of clusters cannot be very large.

Fig. 3.9 and Fig. 3.10 display the final clusters which are formed by each method when the number of clusters is set to 9. The center of each cluster is shown by the red line and is computed by averaging on the load patterns belonging to the cluster. It can be observed that clustering can reveal various distinct consumption patterns among the daily load curves of the customer. Particularly, the following patterns are distinguishable (Here, the K-means results are examined. The analysis is similar for the other methods.):

- Morning peak (cluster #1)
- Mid-day peak (cluster #5)
- Morning and afternoon peaks (cluster #4)
- Morning and night peaks (cluster #3)
- Morning and late night peak (cluster #8)
- Late night peak (cluster #9)
- Variable consumption pattern (cluster #6)

Cluster #7 resembles to cluster #1; however, its corresponding peak has less magnitude and happens at earlier hours. Moreover, cluster #2 characterizes the high consumption during midnight and a local peak at around 10 am.



(a) K-means clustering (b) Fuzzy c-means clustering

(c) Hierarchical clustering (d) SOM clustering

Fig. 3.9 Final clusters of 4 different clustering algorithms

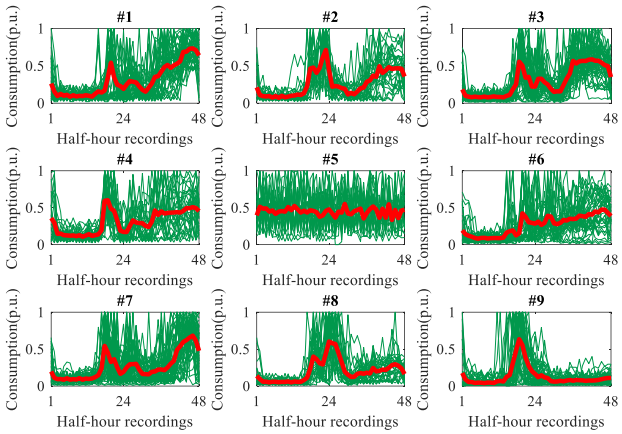


Fig. 3.10. GMM clustering results

3.4.7 Clustering of a large number of electricity customers

Electricity companies desire to segregate their huge number of customers into certain classes based on the daily load patterns. However, as we noted in the previous sections, the daily load patterns of a certain customer might change significantly from a day to another day. This makes the clustering of customers challenging. To overcome this problem, one common approach is to cluster the customers based on their RLPs.

Here, the analyzed data set comprises load data of 4141 customers over a year. Since the customers usually have different consumption behavior on the weekends compared with weekdays, the data set is divided into weekdays and weekends (two loading conditions). Fig. 3.11 and Fig. 3.12 show the final clusters (obtained by a hierarchical algorithm) for weekdays and weekends, respectively. The number of RLPs which belong to each cluster is also displayed in these figures. In order to identify various consumption patterns among customers, a sufficiently big number of clusters is selected.

It can be seen that the difference between the weekday and weekend consumption behavior is significant. For weekday clusters, generally a small peak happens in the morning and the major peak occurs in the evening and nights. Specially, this pattern is clearly visible for clusters #8, # 12, #1, and # 4 that have the highest number of RLPs and totally account for around 40 per cent of load shapes. On the other hand, weekend clusters and particularly, the clusters with the highest number of members i.e. clusters #4, #12, #1, and #6 have a late peak around mid-day or early afternoon. Furthermore, it is also noticeable that the magnitude of the afternoon peak is higher or equal of the night peaks. In addition it can be observed that the consumption level is higher compared with the weekday consumption. Such a difference among usage behavior is predictable since, in the weekends, the residents usually wake up late and spend most of the day in the home while in the weekdays they leave their homes in early mornings and are outside the home for most of the day.

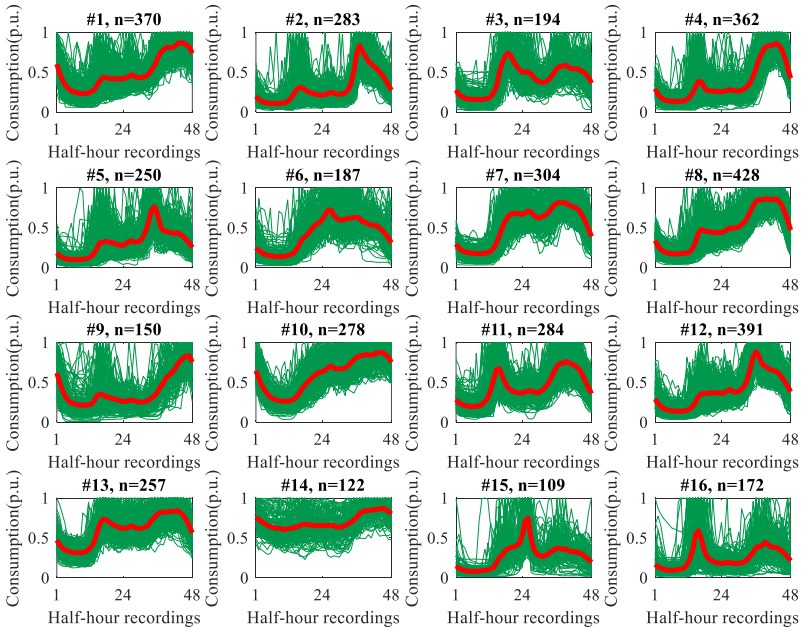


Fig. 3.11 Clusters of the weekday RLPs of 4141 customers

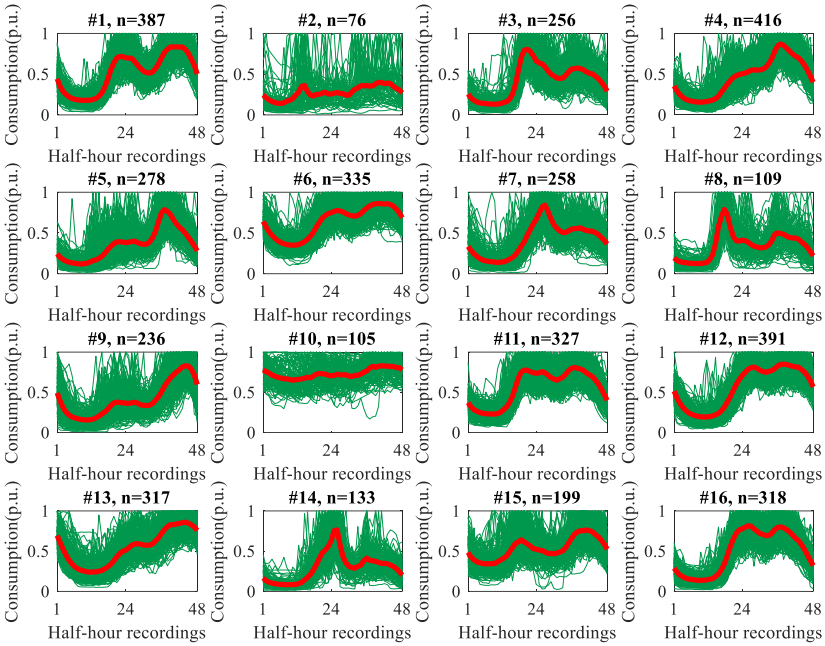


Fig. 3.12. Clusters of the weekend RLPs of 4141 customers

3.4.8 Method comparisons based on the computation time

In this section, the processing time of the algorithms are compared for the clustering of the larger dataset. Table 3-4 and Table 3-5 show the results for the clustering methods and CVI measures, respectively. The results are the average values obtained by running the algorithms 10 different times. The simulations are carried out on a personal computer Intel® CORETMi5 with processors clocking at 2.3 GHz and 8 GB of RAM.

Table 3-4 Comparison of processing time for different clustering algorithms

Clustering method	**Number of clusters**										
	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
K-means	0.128	0.138	0.156	0.235	0.172	0.293	0.227	0.285	0.311	0.265	0.302
FCM-1.2	4.038	9.688	4.762	8.688	8.002	10.390	10.423	15.780	19.909	32.676	25.552
FCM-2	6.343	10.218	9.708	14.455	15.533	14.379	17.375	24.225	17.756	19.410	26.478
FCM-3	0.599	0.614	0.718	0.687	0.693	0.788	0.691	0.709	0.770	0.806	0.825
FCM-4	0.348	0.406	0.369	0.381	0.441	0.457	0.401	0.418	0.389	0.391	0.372
H-Single	0.631	0.598	0.597	0.631	0.623	0.605	0.618	0.597	0.590	0.607	0.602
H-Complete	0.601	0.561	0.572	0.639	0.587	0.582	0.622	0.578	0.593	0.564	0.599

H-Average	0.634	0.603	0.609	0.644	0.631	0.616	0.665	0.594	0.604	0.638	0.635
H-Centroid	0.954	0.943	0.930	1.053	0.986	0.970	1.038	0.934	0.913	0.944	0.981
H-Ward	0.619	0.611	0.610	0.686	0.617	0.610	0.633	0.611	0.625	0.604	0.643
SOM-H12	5.598	5.539	5.556	5.776	5.839	5.652	5.710	5.760	5.763	5.601	5.482
SOM-R12	5.617	5.655	5.675	5.869	5.946	5.697	5.851	5.847	5.730	5.692	5.571
SOM-H18	12.062	12.354	12.293	12.579	12.382	12.630	12.583	12.433	12.488	11.979	12.066
SOM-R18	12.129	12.069	12.337	12.218	12.686	12.341	12.527	12.689	12.422	12.060	12.178

Table 3-5 Comparison of processing time for different CVIs

CVI	**Number of clusters**										
	5	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
MSE	0.017	0.012	0.010	0.011	0.011	0.013	0.012	0.015	0.011	0.011	0.011
SIL	6.916	6.796	6.885	7.017	7.712	7.324	7.192	7.660	7.403	7.314	7.469
DBI	0.033	0.024	0.017	0.018	0.018	0.020	0.019	0.019	0.018	0.020	0.019
MIA	0.022	0.012	0.013	0.013	0.013	0.013	0.013	0.014	0.012	0.013	0.015
WCBCR	0.027	0.017	0.015	0.013	0.019	0.014	0.013	0.012	0.012	0.012	0.014
Dunn	0.420	0.392	0.402	0.394	0.424	0.409	0.399	0.406	0.399	0.404	0.402

It should be noted that the computation time for the algorithms depends on the initialization parameters. For example, in K-means, the method for selection of centers greatly affects the processing time. In addition, if the centers are selected randomly, the time is multiplied by the number of iterations which is needed to repeat the clustering using the new initial centers. Here, the results are reported for only one random selection of centers.

FCM calculation time depends on the fuzziness parameter. In Table 3-4, the processing time for m=1.2, 2, 3 and 4 are reported. It is observed that by increasing the amount of fuzziness degree, the calculation time increases at first and then experiences a steady decrease.

For the hierarchical method, the calculation time only slightly differs for different cluster numbers since the main processing time is dedicated to creating the dendrogram

which can later be cut at any level. Centroid linkage method has the longest time followed by average and Ward methods.

As expected, the time for training the SOM model is relatively high when the number of observations is large. Also, it can be seen that by increasing the grid size from 12×12 to 18×18 the computation time almost doubles.

For CVI measures, MSE, DBI, WCBCR, and MIA calculate the distances between each cluster center with the members of that cluster. Among them, DBI considers all the pairwise combination of clusters which makes it more complicated. Silhouette and Dunn indexes require the calculation of inter-cluster and intra-cluster distances between all the objects. Silhouette has the most complicated formulation which makes it more computationally expensive for large datasets. Table 3-5 also shows that the Silhouette and Dunn indexes represent the longest computation time and the processing time of Silhouette is much higher compared to other CVIs.

3.5 Summary

In this chapter, we comprehensively explored the clustering of electricity customers according to their daily load patterns. The primary aim is to detect different consumption patterns which, subsequently, can be used for improving the other applications in the power system domain such as DR programs. Firstly, the major clustering algorithms were introduced and the main parameters of them are discussed. The case studies were performed to show the effect of these parameters and to compare different clustering methods. Furthermore, the applications of cluster validity indexes were described.

4 A Pattern Recognition Methodology for Analyzing Load Data and Targeting Demand Response Applications

4.1 Background and Motivation

The availability of smart meter data allows defining innovative applications such as demand response (DR) programs for households. However, the dimensionality of data imposes challenges for the data mining of the load patterns. In addition, the inherent variability of residential consumption patterns is a major problem for deciding on the characteristic consumption patterns and implementing proper DR settlements. In this chapter, a data size reduction and clustering approach is utilized in order to analyze the residential customers load data.

It should be noted that the introduction of new technologies can affect residential load curves. The current status of three main technologies including solar PVs, ESSs, and EVs, their possible deployments in short- and mid-terms, and their effects on residential load profiles were discussed in detail in Section 2.3. It was highlighted that the implementation of Solar PVs and batteries has the greatest impact on consumption patterns of dwellings. Using the current data and available studies, it was also shown that the rate of current deployments by households is still limited by the high initial investment costs.

Home energy management systems (HEMSs) are another novel technology that can alter the households' consumption. A HEMS oversees the optimal consumption, generation, storage, and charge/discharge activities in the premises. Therefore, it considers all the affecting factors such as the PV generation, battery status, electricity prices, demand response programs, and EV's charge/discharge and decides on the operation of appliances.

The proposed methodology in this chapter can be used as part of these techniques for identifying DR potentials in different households. The aim of the chapter is twofold: firstly, to develop and apply a symbolic aggregate approximation (SAX) technique as a proper data size reduction method on a large number of daily load shapes of the residential customers, and analyze their underlying consumption patterns, and secondly, to apply the results for DR program targeting for residential households. In spite of its efficacy, the application of SAX in the power system publications, especially for the residential customers and a large number of load curves, has not been explored in detail. To bridge this gap, this chapter offers several contributions as follows:

- It investigates the application of a modified SAX technique on a large dataset comprising hundreds of thousands of daily load curves of residential dwellings. Use of SAX is appropriate since the residential load curves usually display high variability from one day to another day. Therefore, instead of using the daily curves, the relevant SAX representations can be used which brings in more meaningful outcomes.
- To apply the SAX, the main time periods of household activity during the day should be identified. Therefore, an analysis of consumption data is carried out to identify the critical time periods during the day which are used in the SAX to partition the time axis.

- Using a clustering approach, the SAX representations of data are assigned to different clusters. In this stage, two modifications are applied to distance calculation to improve the clustering results. In addition, the effects of the parameters of the SAX technique and cluster numbers are studied by appropriate measures. The obtained results are analyzed which give promising insights about households' consumption patterns.
- The results of the clustering stage are utilized to help in procuring DR from the residential households. The use of SAX is helpful in this stage too. It is in accordance with the needs of retailers or DR aggregators which usually require DR or load changes in specific time periods.

The following sections are organized as follows. In Section 4.2, the preliminary stages before the clustering are briefly explained and the extant literature is summarized and reported. Section 4.3 includes the problem statement and the suggested stages of the method. The theoretical concepts of the methodology including the data size reduction technique, clustering algorithm, and entropy analysis are introduced in Section 4.4. In Section 4.5, the data set is evaluated to extract the main time periods which are used in the SAX method. The case studies and the results are presented in Section 4.6. Finally, conclusions are presented in Section 4.8.

4.2 Preliminary Stages Before the Clustering

In the previous chapter we studied the major clustering algorithms and their applications. However, it should be noted that the volume of recorded electricity consumptions is enormous. For data resolutions of 1 hour, 15 minutes and 1 minute, the length of daily metered data will be 24, 96, 1440 respectively, which clearly shows the

effect of sampling rate on the dimensionality of time series data. Specifically, analyzing these massive sets of data from tens of thousands of smart meters could be a challenging task for electrical utilities. Hence, applying clustering techniques on these data, especially when the number of customers is very high or the time period of study is long, would be a challenging task. Therefore, feature extraction/definition and dimensionality reduction methods are examined in the literature to reduce the size of load data sets. The proper use of these techniques can reduce the input data of clustering algorithms, save computing time, and probably improve the clustering results. In the following, firstly, the applications of feature definition/extraction methods are briefly discussed. Then, the use of principal component analysis (PCA), as a popular data size reduction technique in power system literature, is briefly presented.

4.2.1 Feature definition

Each customer load profile might be represented by a limited number of features. In feature definition approaches some features are defined and employed by the user based on the specific applications. In [23] the authors define seven features and extract them from the raw data including the mean, standard deviation, skewness, kurtosis, chaos, energy, and periodicity. Ref. [79] defines a set of shape indicators, for example, daily average load to maximum load factor, to characterize the load patterns. Haben *et al.* [27] divide each day into four time periods, overnight, breakfast, daytime, and evening periods. Using the consumption values in these periods, seven attributes are defined for each customer. In [73] different variables are derived from the hourly measured energy consumption of customers such as the number of consumption peaks, hourly average consumption, and maximum

consumption per day. Furthermore, [174] defines attributes in such a way that they can represent the flexibility of each household in changing its load. Also, a regression analysis is adopted in [94] which gives eight regression coefficients for the electric load pattern of any customer. These coefficients are different for each customer and are used for the clustering purpose. The proper clustering methods can be applied on these features to distinguish customer classes.

4.2.2 Feature extraction

Feature extraction techniques can also be employed to extract certain features from the load data using techniques such as frequency domain analysis [175], discrete Fourier transform (DFT) [171], and wavelet transform (WT) [176]. DFT is used in [171] to transform time-domain measurements to the frequency domain. Based on the acquired information on amplitude and phase of the harmonic components, a set of features is defined which is used by a modified follow the leader algorithm to cluster customers. Also, frequency analysis of households is performed in [64] to identify the relationship between the load patterns and the lifestyle of households. After calculating the representative frequency for each dwelling, the households are divided into several groups. The application of WT for extraction of features has been discussed in several studies [177], [178], [96], [103]. Ref. [178] proposes two approaches based on WT for clustering one-year load data of a group of French electricity customers. The first method employs discrete WT for feature extraction and feature selection and K-means algorithm for clustering. This approach is very fast and allows the elimination of non-informative features. On the other hand, the second approach is to cluster using a continuous WT and partitioning around medoid (PAM)

algorithm and can result in more refined clusters. Ref. [177] and [96] define a clustering strategy by combining an individual signal pre-processing by wavelet denoising, a dimensionality reduction step by wavelet compression, and a hierarchical clustering algorithm which is applied to a suitably chosen set of wavelet coefficients.

4.2.3 Dimensionality reduction methods

The third approach uses data size reduction techniques to obtain a reduced data set from the primary data set. Various data mining methods have been introduced for decreasing the size of a dataset. In the power system literature, PCA is vastly utilized to convert the load data into a few components which can be further used for clustering

The fundamental idea of PCA is to reduce the dimensionality of a data set consisting of a large number of possibly correlated variables while retaining as much as possible of the variation present in the data set. This is achieved by an orthogonal transformation that converts the data to a new set of variables called principal components (PCs) which are uncorrelated. This transformation transforms the data to a new coordinate system such that the first few PCs retain most of the variation present in all of the original variables [179]. Therefore, the greatest variance by any projection of the data becomes the first coordinate (the first component), the second greatest variance the second coordinate, and so on [180]. Often the number of PCs needed to sufficiently represent the original data is quite small and this makes PCA a suitable tool for dimensionality reduction.

The application of PCA method to the customer data set for different number of PCA components is carried out as shown in Fig. 4.1. Most of the variance is explained by the first six components and its value does not change meaningfully after around 10 PCs. The

adequate number of PCs and the suitable number of clusters can be acquired by CVIs as displayed in this figure. By increasing the number of PCs from 2 to 4, the results improve significantly. However, no considerable change can be observed for more PCs. In this case, the number of final PCs can be selected as 5 or 6.

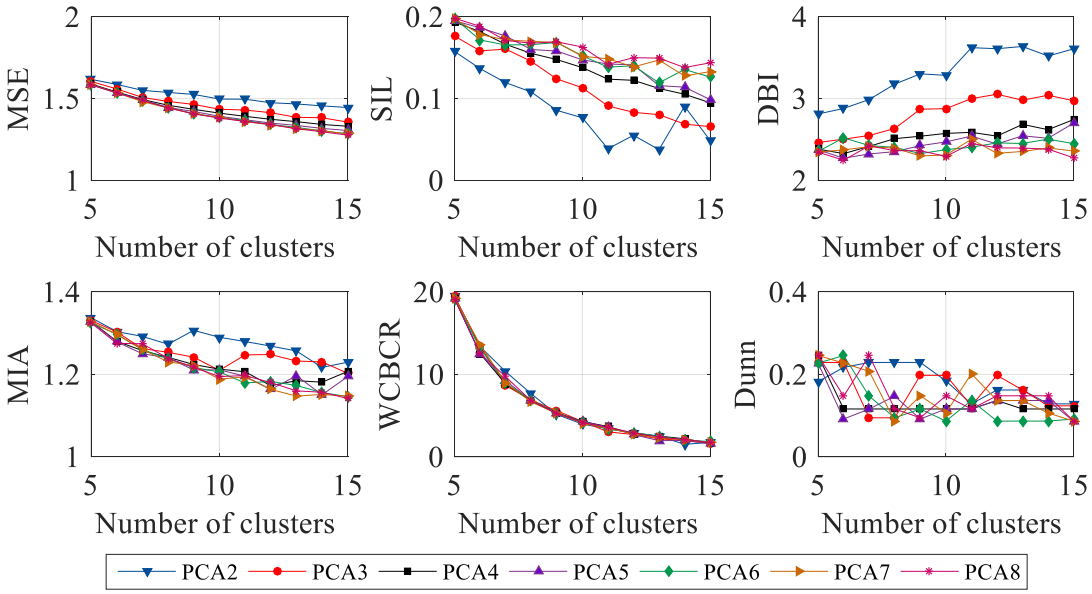


Fig. 4.1. Performance of a combined clustering of PCA and K-means for different number of clusters and PCs

PCA is applied in several studies to characterize customers' consumptions. In [77], PCA is employed to understand and visualize measured consumption data. K-means clustering is then used to cluster the data set based on the first four PCs. In [72], using PCA, 48 half-hour load data are converted to a few PCs and an SOM strategy is applied to reveal a number of distinct behavioral components, for example, high consumptions vs. low consumption. Moreover, in [68], PCA is used to reduce the dimensionality and to detect the existence of seasonality in load curves. Also, the performance of PCA is compared against two other data size reduction techniques in [57].

SAX is a well-known dimensionality reduction technique in the data mining literature [181]. It is a suitable technique for producing a representation of original time series data by a limited number of features; however, its use for residential load data is not studied in detail. In the next sections, an innovative SAX method will be utilized for analyzing the residential customers load patterns and enhancing DR programs.

4.3 Problem Statement

Use of RLPs offers some obvious advantages as the consumption pattern of each customer over a period of time is represented by only one curve. In particular, it reduces the volume of data and makes the clustering achievable. On the other hand, the temporal characteristics of load data, which is important for DR applications, can be missed by using RLPs. For example, as shown in Fig. 4.2, two customers with completely different daily load patterns can still have similar RLPs (shown by the black lines). According to this figure, the first customer has a very unpredictable consumption behavior during a week, while the other follows a very regular pattern. It means that, regardless of the consumption value, the low and high consumption periods of the second customer do not differ considerably from one day to another day. This is especially critical for DR applications when understanding the stability of user's consumption habits over time is of high importance.

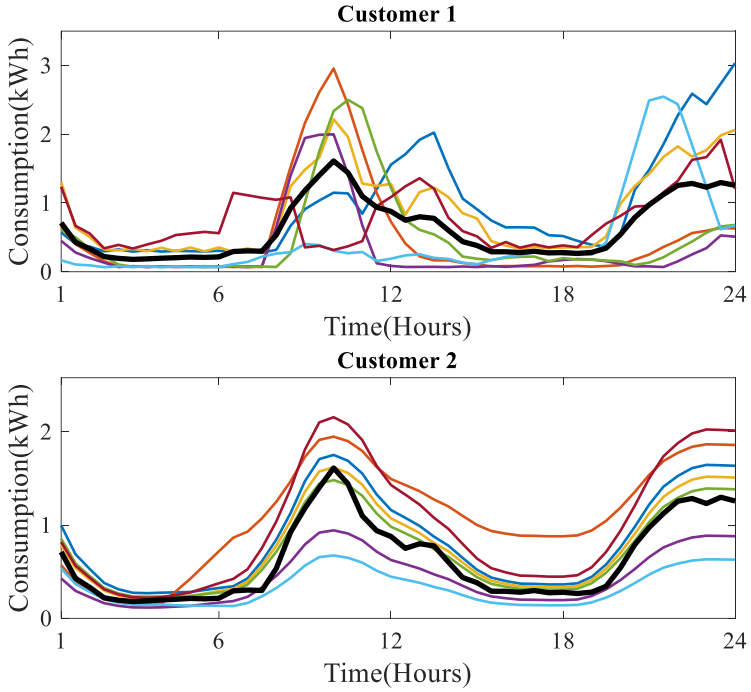


Fig. 4.2. Consumption behavior of two customers

Use of SAX technique can help in settling the aforementioned problems. It has two suitable features which make it a perfect tool for analyzing residential load patterns. First, it considers several important time periods during the day instead of focusing on the exact time of consumption. Secondly, instead of dealing with the infinite number of consumption values, it transforms the load data into a series of symbols. Using these two features, each load pattern can be represented by a limited number of symbols. This allows limiting the number of many daily load curve varieties while retaining the important information of original load data. SAX representations can then be clustered using proper clustering techniques. Furthermore, applying the SAX can facilitate DR applications for example, by enabling the segregation of stable customers from variable ones.

4.3.1 Stages of the method

Based on the aforementioned analysis, the proposed methodology in this paper which includes four main steps is shown in Fig. 4.3.

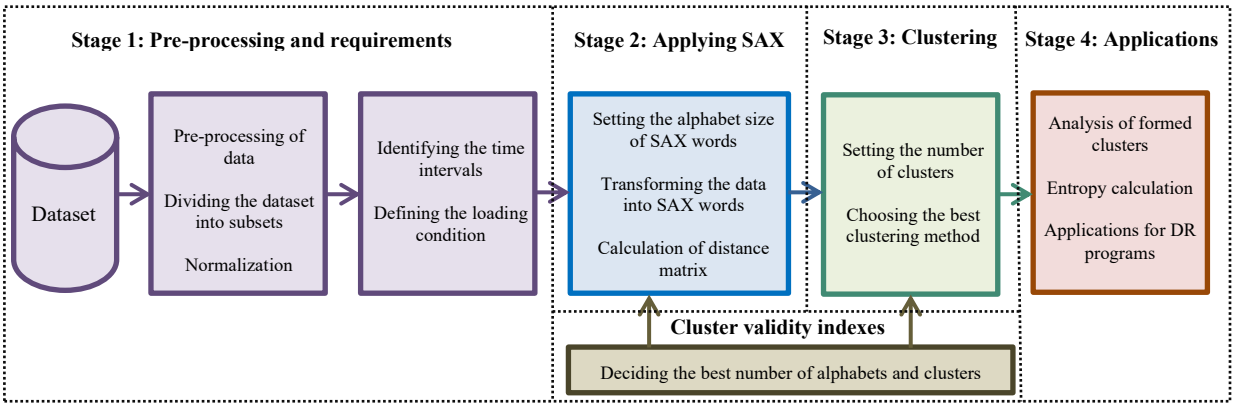


Fig. 4.3. Stages of the methodology

Data preparation: prior to data analysis, load data preparation is carried out in which the data set is divided into weekdays and weekend sets. Public holidays are also extracted from the weekdays and the customers' daily load profiles are normalized based on the daily maximum consumption. Loading conditions are defined in this stage and the data are evaluated to determine the proper time intervals during the day, which are needed as the breakpoints on the time axis in the SAX method.

Implementing the SAX: the second stage applies the SAX on the load patterns and calculates the distance matrix. Two modifications are also applied in this stage to the SAX method to accommodate the underlying assumptions of residential consumption and to achieve better clustering results.

Clustering: clustering using a hierarchical algorithm is conducted in Stage 3. Different hierarchical methods are evaluated and the best one is selected for the segmentation of load

curves. Furthermore, the effects of SAX parameters and number of clusters are examined in Stages 2 and 3.

Applications and DR targeting: The analysis of the results and further applications for DR programs are conducted in Stage 4. In this step, the entropy concept from the information theory is utilized to rank customers based on their stabilities over time which can be further used in applying customized DR programs for different customers. In other words, this measure ranks customers based on their variability in their daily load shapes over time and distinguishes between the customers with stable and regular behavior and customers with irregular and unstable behavior.

4.4 Methodology

In this section, the concepts and mathematical formulations of SAX technique, clustering algorithm, and entropy measure are illustrated.

4.4.1 SAX method

Various methods have been proposed in the data mining literature for the high level representation of time series data including the SAX technique. SAX is a symbolic representation for the time series data which discretizes the original data into symbolic strings. This technique, proposed by Lin *et al.* [181], uses an intermediate transformation of raw data called piecewise aggregate approximation (PAA) for the dimensionality reduction and then, symbolizes the PAA representation into a string composed of some alphabets. PAA partitions the time domain into a specified number of time frames and reduces the length of the original time series by replacing the data falling in the same time frame by

their corresponding mean value. Usually, these time intervals have the same size. In our proposed method, we determine the time intervals based on the approach defined in Section 4.5. Mathematically expressing, PAA transforms the original time series $X = \{x_1, x_2, \dots, x_H\}$ to a vector of $C = \{c_1, c_2, \dots, c_K\}$ $(K < H)$ in which c_i is defined as:

$$c_i = \frac{\sum_{x_j \in T_i} x_j}{k_i} \tag{4.1}$$

where T_i and k_i represent the ith time interval and the number of the data values falling in that time interval, respectively.

Once c_i is obtained, their SAX representation can be realized by defining a series of breakpoints $\{z_1, z_2, \dots, z_{Q-1}\}$ which partition the amplitude axis into Q intervals. These breakpoints can be determined based on the quantile of the statistical distribution that represents the probability density of the amplitudes in the whole data set [168]. A symbol is then assigned to each of these intervals and the PAA values are mapped to these symbols according to the interval they fall in. Therefore, if the set of symbols are defined as $\{\alpha_1, \alpha_2, \dots, \alpha_Q\}$ and $z_{j-1} \leq c_i < z_j$, then c_i will be mapped to α_j. In this way, the original time series can be replaced by a SAX "word".

Fig. 4.4 demonstrates the process of generating SAX word for a typical normalized load curve. Each day is divided into four intervals which have the same length of six and the amplitude breakpoints are selected as $\{0.25, 0.5, 0.75\}$. The data points falling in the same time frame are averaged and mapped to the letters $\{a, b, c, d\}$ hence, resulting in the SAX word "abababacabbcaaacabab".

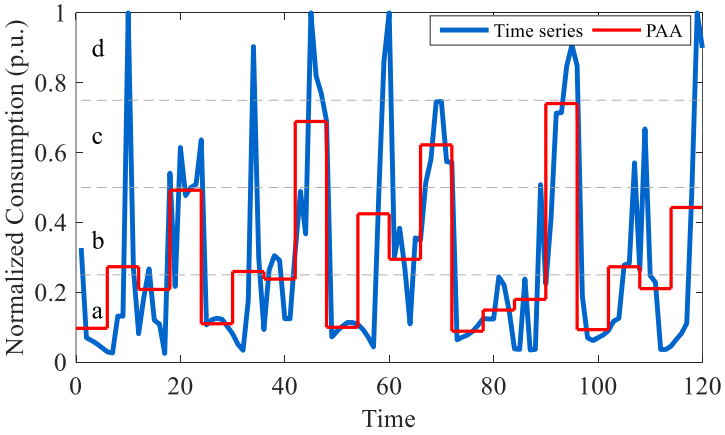


Fig. 4.4. PAA and SAX representation of a load curve for 5 days

In the proposed method, every normalized daily load curve which consists of 48 data points will be replaced by its SAX representation of length five as a day is divided into five periods.

4.4.2 Clustering stage

Considering that the SAX entries are categorical data, certain clustering algorithms such as K-means cannot be used for the clustering. Other algorithms including K-modes, hierarchical, and DBSCAN can be applied to partition the SAX representation of daily load curves into a number of clusters. Hierarchical agglomerative clustering is a good choice which is used here.

To apply the method, firstly, the distance between the pairs of data points (load curves) is calculated and a similarity matrix is built. In order to apply the clustering algorithm on the SAX representations of the data, the distance between two SAX words should be defined. When dealing with the raw data, usually Euclidean distance is used to calculate the distance between two load curves or two RLPs. For the symbolic representation of the load curves a proper distance metric needs to be defined. When the time domain is partitioned into equal

time frames, the following distance definition known as MINIDIST is usually used for calculating the dissimilarity between two SAX words W_1 and W_2 [181]:

$$MINIDIST(W_1, W_2) = \sqrt{k. \sum_{i=1}^{K} (dist(\alpha_i, \beta_i))^2} \tag{4.2}$$

$$dist(\alpha, \beta) = \begin{cases} 0, & if\ |\alpha - \beta| \leq 1 \\ z_{max(\alpha,\beta)-1} - z_{min(\alpha,\beta)}, & otherwise \end{cases} \tag{4.3}$$

where α_i and β_i represent the ith symbols of W_1 and W_2, respectively and $k = H/K$ is the number of data points in each time interval.

Therefore, when calculating the distance between two SAX words, the distance between their individual symbols will be calculated first using (4.3). According to (4.3), if two symbols are adjacent, then the distance between them is zero, otherwise, the distance between them is calculated using the amplitude breakpoints. For instance, in the previous example, $dist(b, c) = 0$ and $dist(b, d) = 0.75 - 0.5 = 0.25$. It can be proved that this definition for the distance between SAX words lower-bounds the Euclidean distance between the two original time series.

We apply two modifications to the above distance calculation. Since the partitions of time axis have different sizes the fixed k in (4.2) is replaced by k_i which is equal to the data points in each time interval. Accordingly, (4.2) should be re-written as:

$$MINIDIST(W_1, W_2) = \sqrt{\sum_{i=1}^{K} k_i. (dist(\alpha_i, \beta_i))^2} \tag{4.4}$$

The definition in (4.3) considers the distance of adjacent symbols as zero and ignores the minimum and maximum points of time series. Therefore, here, the second modification is proposed based on the work in [182] for calculation of the distance in (4.3), which shows better results for clustering purposes. Firstly, for each interval, an indicator can be defined as following:

$$Ind_i = \frac{z_{i-1} + z_i}{2} \tag{4.5}$$

For the lowest region, z_{i-1} is the global minimum and for the highest one z_i is the global maximum. Secondly, the distance between two intervals can be defined as the distance between their corresponding indicators:

$$dist(\alpha, \beta) = \left| Ind_\alpha - Ind_\beta \right| \tag{4.6}$$

Using the above definitions, the distance between the pairs of daily load curves, which are represented by their SAX words, is calculated. This process results in a distance matrix which will be used as the input of the hierarchical clustering algorithm. Based on this matrix, clusters are merged using a linkage criterion. The linkage is an evaluation function which indicates the best candidates for merging. Likewise the dissimilarity measure, the choice of linkage can also have an impact on the final clustering outcomes.

The average linkage is selected for the hierarchical clustering since the obtained results confirmed its superiority to the other methods. Let C_p and C_q be two clusters with n_p and n_q members, respectively. Then, based on the average method the distance between these two clusters is:

$$D(C_p, C_q) = \frac{1}{n_p . n_q} \sum_{\substack{X_i \in C_p \\ Y_i \in C_q}} d(X_i, Y_j) = \frac{1}{n_p . n_q} \sum_{\substack{W_i \in C_p \\ W_j \in C_q}} MINIDIST(W_i, W_j) \quad (4.7)$$

4.4.3 DR application

The consumption of a household can be characterized by several main features: total/average energy consumption [183] [184], load shape [32] [60], the amount and time of the peak [53], and the stability in usage pattern over time. The first three determinants are well studied in the literature. In spite of its importance, the stability of the customer's usage behavior over time has been addressed in very limited studies [54], [52]. Nonetheless, it remains as one of the main challenges of DR implementation. In the following, using the results of clustering in the last subsection, the customers will be ranked based on their stability over time.

By applying the clustering algorithm, the daily load curves of each dwelling will be assigned to different clusters based on the customer's consumption habits. The higher variability of daily consumption patterns means that the daily load curves will belong to the greater number of clusters. This gives an intuitive measure to compare households based on their stability over time. However, this method does not provide a fair basis for comparing customers. For example, the daily curves of two different customers can be assigned to the same number of clusters in spite of having completely different patterns as shown in Table 4-1. Here, there are 100 daily load curves for each customer and the number of final clusters is set to 10. As it can be observed, customer A has a very regular pattern since almost all of

his/her daily curves are assigned to the Cluster # 2, while customer B follows much diverse consumption habits as his/her daily curves are distributed among four different clusters.

Table 4-1 A sample assignment of daily curves of two customers to different clusters

	Cluster Number			
	#2	**#5**	**#7**	**#10**
Customer A	97	1	1	1
Customer B	23	27	19	31

Therefore, a more sophisticated metric is needed to quantify the variability of households and rank them. We borrow the notion of entropy form the information theory to classify customers based on their stability over time. Shannon entropy is a popular entropy measure in various fields and is used in this paper too. It can be expressed as [185]:

$$H = -\sum_{i=1}^{N} p_i \,.\, log_b(p_i) \tag{4.8}$$

where N is the number of classes, and p_i is the probability (relative frequency) of an object from the ith class appearing. b is the base of the logarithm and its common values are 2, e, and 10.

In our context, each class represents a cluster and the relative frequency, p_i, is the number of times the daily load pattern of a household belongs to the cluster i. Therefore, the entropy metric not only considers the number of clusters but also the probability that they appear in the customers' daily load shapes. If a household has a completely stable pattern, which means that all his daily curves belong to just one cluster, the entropy will be 0 ($p_i =$

1). In other extreme case, if a customer follows a completely irregular pattern, which means that all the clusters are equally likely in his daily load curves, the entropy will be the highest.

4.5 Preliminary Analysis of the Dataset

The analyses in this section and the next section are performed for weekday and weekend datasets. The presented concepts are firstly demonstrated using a dataset comprising of 300 customers, while a much larger dataset consisting of 4141 customers is used for further DR analysis.

In this stage, the aim is to find time periods during a day which are distinguished based on the energy consumption levels and the local troughs and peaks. A detailed analysis is performed in order to distinguish the proper periods during the day. These time intervals are the characteristic time periods where certain consumption patterns can happen. These periods correspond with the typical periods of household activity and the changes in the energy usages. They will be used as the time breakpoints in the SAX method.

Fig. 4.5 and Fig. 4.6 show the box and whisker plots of the daily consumption data of all customers for the weekdays and weekends, respectively. For simplicity, outliers are not shown in these plots. The median value of the consumption in every hour is shown by the solid line in the middle of each box. The two ends of the box show the first quartile, q_1, and third quartile, q_3. The bottom and top whiskers are specified as $q_1 - 1.5 \times (q_3 - q_1)$ and $q_3 + 1.5 \times (q_3 - q_1)$ and limit the range outside the box which is shown by the dashed line.

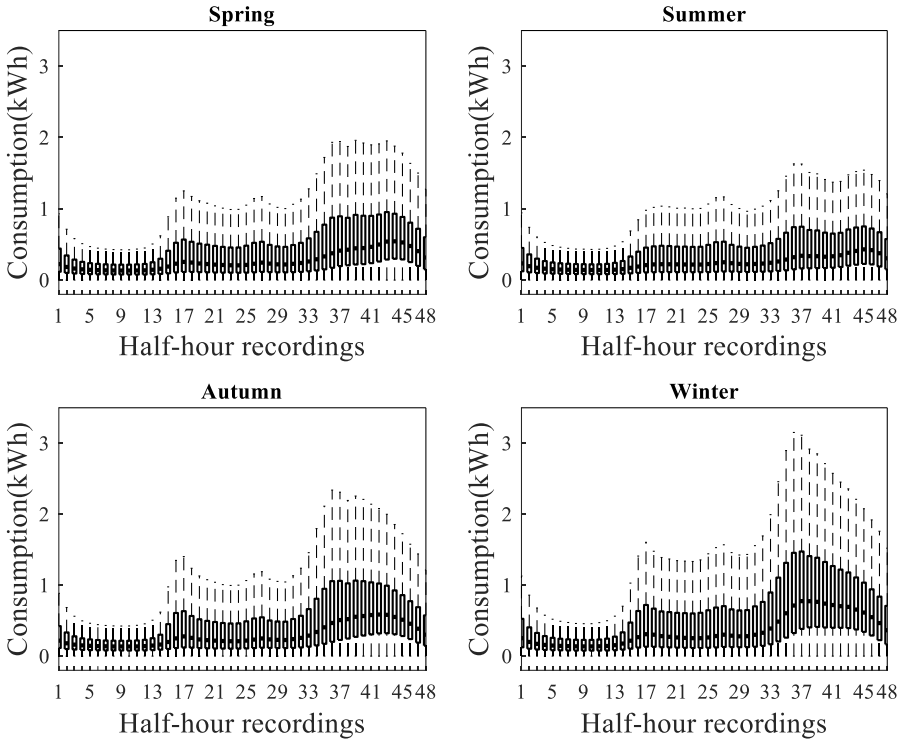


Fig. 4.5. Boxplots of weekday consumption in different seasons

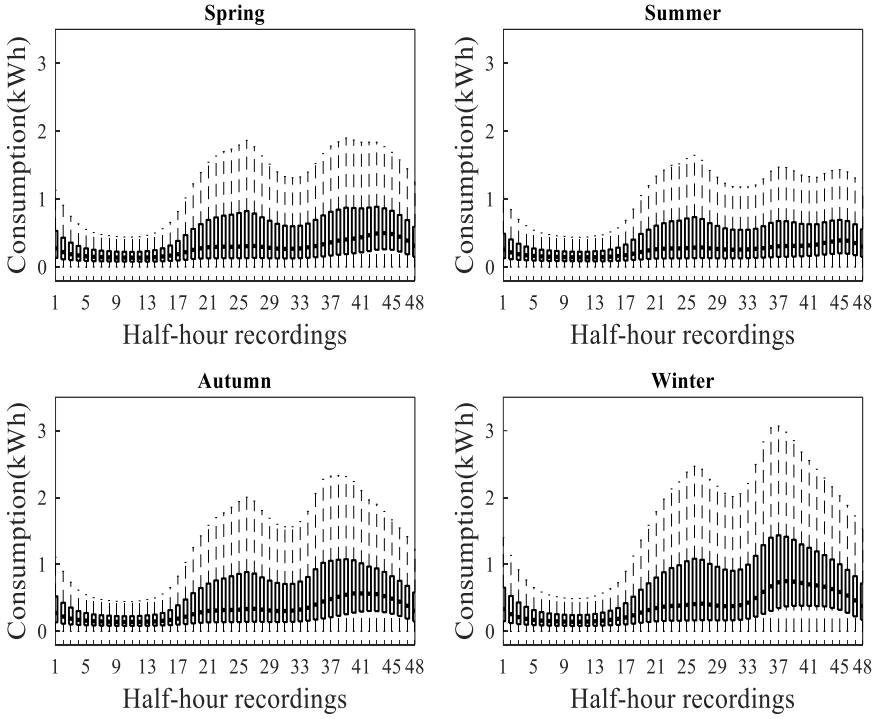


Fig. 4.6. Boxplots of weekend consumption in different seasons

Analysis of these figures reveals important information about the consumption magnitude and pattern during different seasons. In addition, a significant difference can be observed in the energy consumption patterns of weekdays and weekends.

For the weekdays, there are several local peaks which occur in the morning, mid-day, evening, and night periods. Although the occurring times of these peaks slightly differ among the seasons, the general trend is almost the same for all of them. As it is expected, the daytime peak during the weekends is much higher in comparison to weekdays. Specifically, for summer and spring, the daytime peak is the same or higher than the night peak. It is understandable since during the weekdays most of the people are usually at the workplaces, while they spend much more time in the home during the weekends. There is no significant change in the night consumption. Another main difference for the weekends is the shifting of the local peak of the morning to mid-day. Such pattern is also expected since occupants usually wake up at later hours on Saturday and Sunday.

Table 4-2 reports the time intervals which are inferred from these plots which distinguish the periods of household activity and are used as the input of SAX method.

Table 4-2 Characteristic time intervals of the day (used for the SAX method)

Weekdays		
Recording number	**Hours**	**Period**
3- 14	1 am- 7 am	Overnight 2
15-18	7 am- 9 am	Morning
19-33	9am- 4:30 pm	Daytime
34-45	4:30 pm- 10:30 pm	Evening
46-2	10:30 pm- 1 am	Overnight 1

Weekends		
Recording number	**Hours**	**Period**
3- 16	1 am- 8 am	Overnight 2
17-20	8am- 10 am	Morning
21-34	10 am- 5 pm	Daytime
35-46	5 pm- 11 pm	Evening
47-2	11 pm- 1 am	Overnight 1

4.6 Case Study

4.6.1 Application of SAX and clustering algorithms

The time periods that were defined for the SAX method cover the main time intervals during the whole year. However, the usage patterns of customers might slightly vary based on the temperature and seasonal changes. Usually, more similar consumption habits happen during a season. The current practice in the literature is to divide a year into 4 (or sometimes 5) seasons and study the consumption behavior based on the seasons [81] [56] [30]. In our case, since we were studying the customer stability over time, we tried to define the season based on the temperature variations to expand the time period of study compared to the usual practice which considers only one pre-defined season of data. The analysis indicates that there is a good correlation between the temperature values (from Irish Meteorological Service [186]) and the overall consumption pattern as shown in Fig. 4.7. This approach represents a more realistic scenario as well as more number of days, resulting in 93 working days including 11 days from the end of spring, 18 days from the beginning of autumn, and all the working days of summer (64 days).

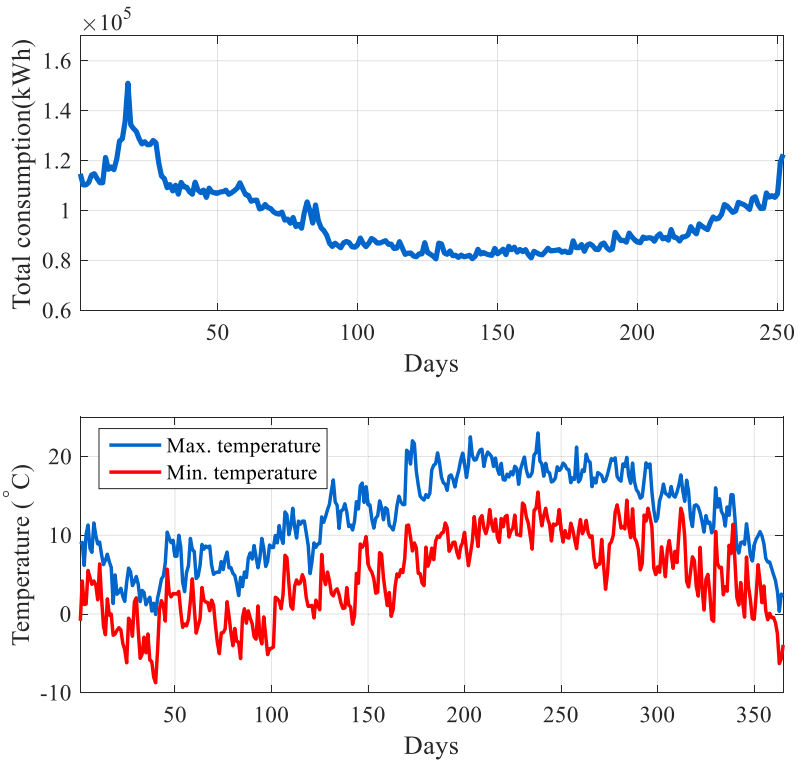


Fig. 4.7. Daily total consumption and temperature variation

The presented concepts are firstly demonstrated using a dataset comprising of 300 customers, while a much larger dataset consisting of 4141 customers is used for further DR analysis. Therefore, 27,900 load curves (300 customers × 93 days) are available for this loading condition. For the weekends, a year of data is considered which consists of 104 days. Consequently, for the time period under study, the total number of 31200 load patterns will be available for the weekends.

The analysis is performed for various conditions mainly, by selecting different number of symbols (amplitude partitions) and different number of clusters. As will be explained in the next subsections, using seven alphabets shows superior clustering performance and hence, it is used for explaining the results.

The amplitude axis is partitioned into seven regions which are defined using the quantile of the PAA values. Accordingly, there are six breakpoints which are determined based on the cumulative density function (CDF) of the whole data. It means that these breakpoints divide the CDF curve into seven equiprobable regions as shown in Fig. 4.8.

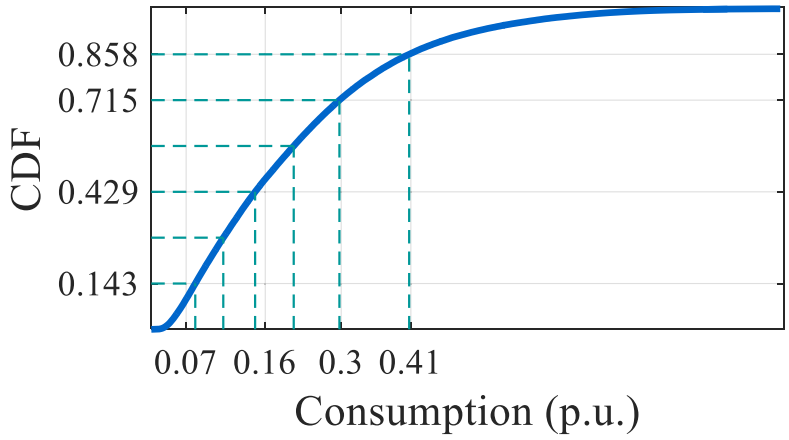


Fig. 4.8. CDF of the whole data

These regions correspond to the letters ‘a’ to ‘g’. The normalized daily load shapes are transformed into their corresponding SAX words. The distance matrix is constructed using the modified MINIDIST function. Finally, the SAX words are clustered using the hierarchical clustering algorithms into clusters. Cophenetic correlation coefficients are calculated for average, single, complete, and Ward clustering methods in which the results demonstrate the better performance for average clustering. Finally, the entropy of each customer is calculated based on the frequency of appearance of clusters in his the load curves.

It should be noted that the maximum number of different SAX words cannot be more than $7^5 = 16708$ as seven possible symbols exist for each period. Consequently, even if

the number of daily load shapes increases considerably, the possible combinations of the symbols “a” to “g” are limited to this number.

4.6.2 Analysis of weekday and weekend clusters

The number of clusters is determined based on the values of DBI and MIA indexes. Following the analysis presented in 4.6.4, the number of clusters is set to 120. Analysis of the formed clusters and their shapes can give valuable information about the consumption patterns. Fig. 4.9 depicts the clustering results for the weekday data set in which the centers for 30 clusters with the largest number of members are shown. The center is calculated by averaging all the curves that belong to the cluster.

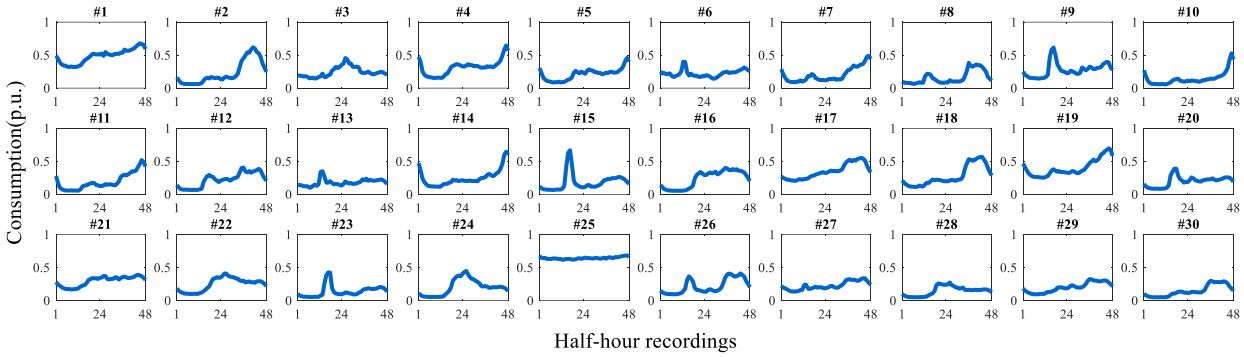


Fig. 4.9. Centers of the clusters with the highest number of daily load curves for the weekday data set

Each cluster characterizes a different consumption pattern. Especially, these patterns are visible:

- Morning peak, clusters #6, #9, #13, and #15, #20, and #23: except for cluster # 6 whose peak happens at early morning (around 6), other clusters show a late peak during the morning. Clusters # 9 and # 15 represent a major peak which shows the energy consumption by various appliances in this period.

- Mid-day peak; clusters #3, #22, #24, and #28: for this category, the peak happens at early afternoon and, except for cluster # 3, they generally follow the same pattern of an extended peak. It means that the consumption slowly increases until the peak point and then again follows a slow decrease pattern.
- Late night peak; clusters #1, #4, #5, #7, #10, #11, #14, and #19: these clusters show a peak near to the midnight. Clusters #1, #7, #11, and #19 show a smooth steady increase in the usage from early afternoon till midnight compared with clusters #10 and #14 which have a sharper rise of consumption.
- Night peak; clusters #2, #17, #18, #27, #29, and #30: compared with other clusters in this category, clusters #2 and #18 display a significant increase in consumption during night time.
- Dual peaks; clusters #8, #12, #26: these clusters show dual peaks happening in the morning and in the evening.

Clusters #16, #21, and #25 classify those days when customer had an almost stable consumption during the day starting from the morning and lasting until midnight.

Clusters #29 and #30 have the largest number of members (4187 and 2406 daily load shapes, respectively) which account for around 24 per cent of all daily curves. Not surprisingly, they represent a consumption pattern similar to the general trend of energy use in summer which was shown in Fig. 4.5.

The same analysis that is performed for the week days can also be applied on weekend data set. The centers of main clusters, that contain the largest number of members, are depicted in Fig. 4.10. It can be seen that weekend clusters generally have a peak in the afternoon and nights rather than in the morning. Especially, clusters #25 to #30 which have

the largest number of members follow this pattern. Cluster #30 has 5541 members and its shape resembles to the consumption patterns shown in Fig. 4.6.

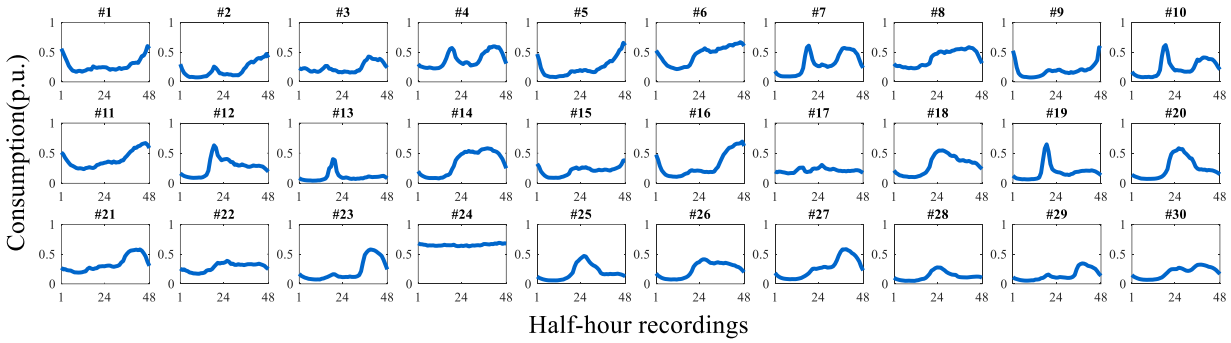


Fig. 4.10. Centers of the clusters with the highest number of daily load curves for the weekend data set

4.6.3 Entropy analysis

The results of the method distinguish the customers based on their variability in the defined periods. The highest entropy that a customer can have is $-log\left(\frac{1}{93}\right) = 1.9685$ which happens when each daily curve belongs to a different cluster. Here, the daily curves of the dwelling with the lowest entropy are assigned to just 4 clusters, while for the customer with the highest entropy they belong to more than 50 clusters. The entropy values for the former and latter cases are 0.2444 and 1.5991, respectively. Fig. 4.11 shows the daily curves of two customers with stable consumption behavior and two customers with variable consumption behavior that are decided based on the entropy analysis. As it can be seen, even the stable customers show slightly different patterns from one day to another day which is an inherent feature of the residential energy usage.

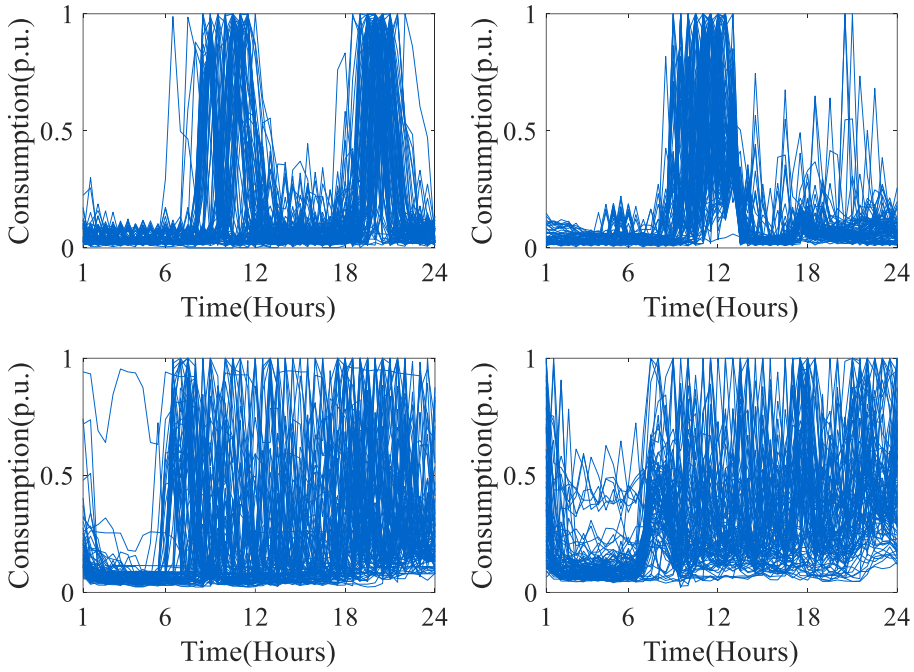


Fig. 4.11. Load curves of two customers with stable consumption behavior (top) and two customers with variable consumption behavior (bottom)

Entropy analysis reveals that the weekend consumption is more variable than the weekdays, which can be associated with the more stable schedule of households during the weekdays. Furthermore, the study shows that those customers who show a stable behavior on the weekdays do not necessarily follow a regular pattern on the weekends.

4.6.4 Effect of amplitude partitioning and number of clusters

The final clustering results are affected by the number of regions of amplitude axis i.e. the number of alphabets. To investigate this effect on clustering results, the number of amplitude breakpoints has been changed from 4 (5 regions) to 7 (8 regions). On the other hand, the effect of the number of clusters is also studied by changing it from 50 to 300.

Fig. 4.12 and Fig. 4.13 display the results for DBI and MIA, respectively. It can be seen that, both DBI and MIA values decrease with the increase of the number of clusters.

However, they do not change significantly after around 120 and for this reason, the number of clusters is set to 120. In addition, it can be observed that the best clustering has been achieved when the alphabet of size 7 is used.

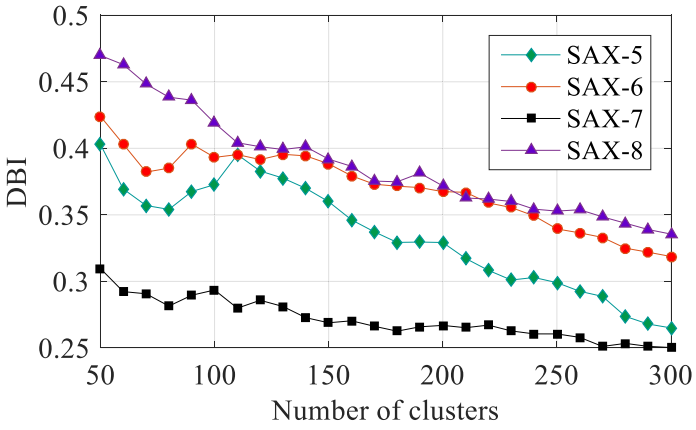


Fig. 4.12. DBI values for different size of alphabets and different number of clusters

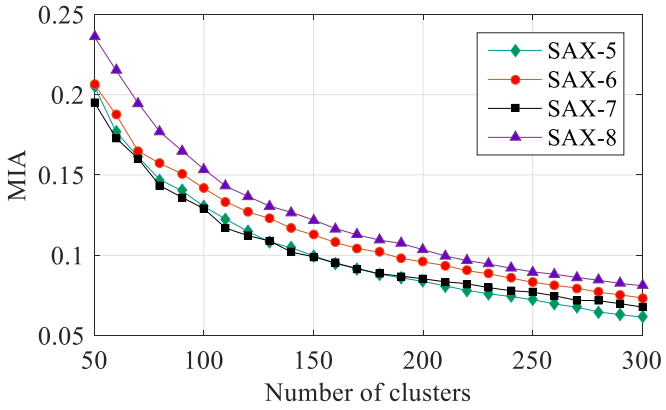


Fig. 4.13. MIA values for different size of alphabets and different number of clusters

It is also interesting to study the effect of these variables on entropy calculations and their ability to distinguish stable customers. To this end, for each number of clusters, the entropies of customers are calculated for a different number of alphabets. Then, 100 customers with the lowest entropies for each of these four scenarios and for the different

number of clusters were obtained. It is observed that there are at least 66 common users among the 100 users for different number of alphabets.

4.6.5 Entropy analysis for a large number of customers

The presented method in the last section is precise, however, with the increase of the number of load curves the computational time increases. The main reason is that SAX words are categorical variables and the distance matrix for the clustering algorithm needs to be constructed. Consequently, if there are n SAX words, the order of distance matrix is $n \times n$. More sophisticated approaches for clustering are needed that is beyond the scope of this study. Therefore, here, we adopt a more straightforward approach in order to be able to compare a large number of customers. The steps of the proposed method are illustrated in the following:

- Step 1: Define the K characteristic load shapes based on which all the load curves can be compared. These characteristic load curves can be defined by the user or for example, as a preliminary stage, they can be found by a simple clustering algorithm like K-means.
- Step 2: Convert these K characteristic load shapes to SAX words (representative SAX words).
- Step 3: Convert all the daily curves of each customer to SAX words.
- Step 4: Compute the distance of each daily load curve of the customer with the representative SAX words and assign it to the most similar one.
- Step 5: Calculate entropy.

The case studies for the working days and weekend days are illustrated in the following.

The simulations are carried out for the working days of one year which comprises 252 days. The partitioning of time and amplitude axis is the same as the previous section. The number of representative load shapes is decided large enough to represent various load patterns. It should be noted that some of the representative SAX words of these K centers might be identical and hence, the replicated ones need to be removed. In our analysis, the final number of representative SAX words is decreased to 39. For each daily curve, the distance of its SAX word with these 39 load shapes is computed and it is assigned to the most similar load shape. Finally, the entropy of each customer is calculated.

The histogram of Fig. 4.14 shows how the daily load shapes of customers are assigned to different clusters. It can be observed that the daily load curves of most of the customers are assigned to around 20 clusters. The daily load curves of the most stable customer belong to only 2 clusters, while the ones of the most variable customer are assigned to 32 clusters. However, as emphasized before, this measure cannot independently reveal the customer's variability. Fig. 4.15 and Fig. 4.16 show the load shapes of some of the stable customers for the weekdays and weekend datasets respectively.

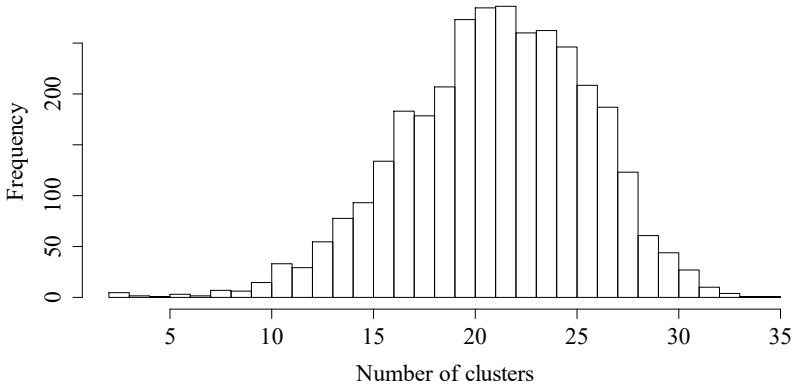


Fig. 4.14. Cluster assignment distribution

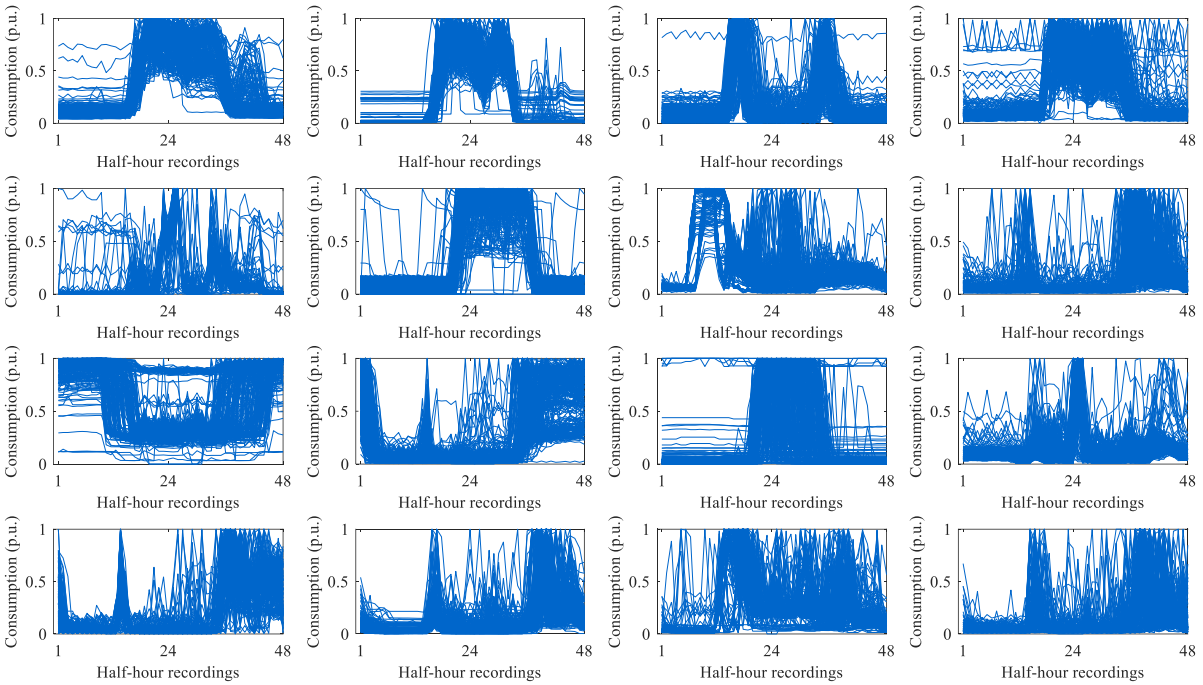


Fig. 4.15. Load curves of sample customers with stable consumption behavior for weekday dataset

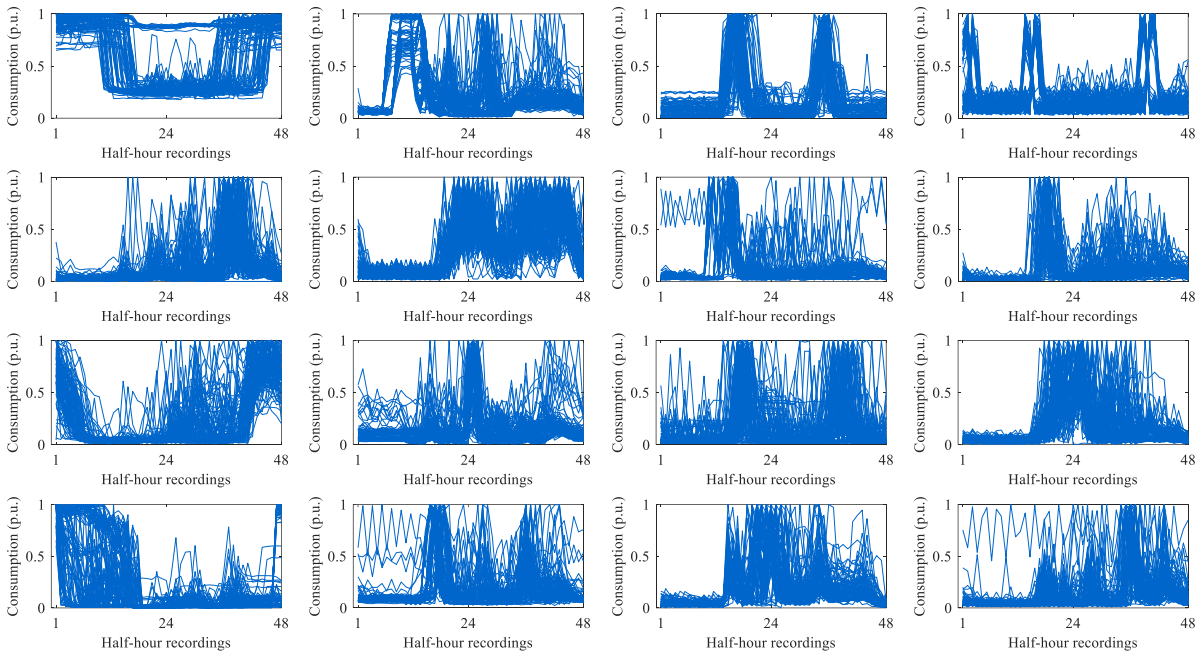


Fig. 4.16. Load curves of sample customers with stable consumption behavior for weekend dataset

There are 104 days for the weekend dataset and the representative load patterns are different for this dataset. The entropy analysis for weekend days reveals that customers show more variable consumption patterns during the weekends.

In the next step, we investigate the stable customers who are common between the two datasets. To this end, the customers are ranked based on their entropies and the top common stable customers in the two datasets are found. The analysis shows that there is not any direct relationship between the stable customers of the weekday dataset and the weekend dataset. For example, among the top 500 stable customers for each of weekday and weekend dataset, only 74 customers are common.

4.7 Comparison with current methods and applications

4.7.1 Current practices for DR aggregation

Increasing the number of participating users (or devices) in DR programs means that it is a higher possibility to have adjustable loads available when the need arises to reduce or increase the demand [187]. In other way, more effective DR can be achieved by having more participants. Besides, current electricity market structures require a minimum amount of bids for participation in the market [188]. Therefore, the DR aggregation can be seen as a promising way for the exposure of small users such as residential customers to wholesale electricity markets. The aggregation of loads can be achieved through an agent called demand response aggregator (DRA). It can schedule DR and will be paid at the relevant regional spot price for this response. It can also offer ancillary services from the load side. The DRAs, on behalf of residential customers, can participate in different markets. Once the market is cleared, the DRAs receive the DR schedule and require the customers with

accepted offers to reduce loads during the contracted DR periods. Load reduction can be achieved through load curtailment, load shifting, utilization of onsite generation and energy storage systems [188].

A basic structure of DR mechanism involving DRA is depicted in Fig. 4.17.

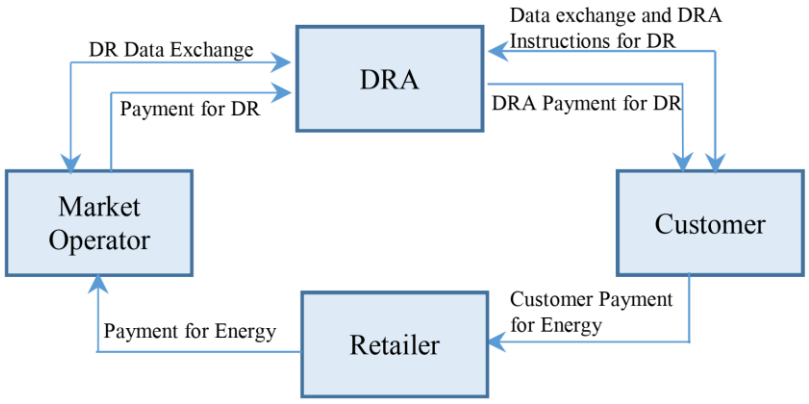


Fig. 4.17. Demand response aggregator model

A great deal of research conducted on DR assumes that there are available (and enough) aggregated DR resources to be utilized for different market/system applications. In this case, it is presumed that aggregation of small DR loads can be performed by a DRA; hence, no attention is paid to the real source of DR, the methodology of aggregation or dynamics of energy use [122] [188-190]. For example, ref. [188] and [189] consider the interaction of DRAs in the market and distribution system, respectively. In [188] DRA offers multiple contracts to customers and seeks to maximize its payoff for participation in the energy market. However, the method does not specify a special kind of appliance. Ref. [189] describes the interaction of several DRAs in a bi-level leader-follower framework in which energy scheduling problem for a load serving entity that serves both inflexible and flexible loads is discussed.

A limited number of studies present frameworks and ideas about the aggregation process of a group of residents. In [191], a conceptual framework called Smartlink is proposed for DR aggregation purposes. This model includes different nodes allowing having as many as DRAs in a hierarchical way. For example, in a system with a two-level hierarchy of nodes, the top DRA interfaces with the utility and three lower-level aggregators. Ref. [192] adopts a multi-agent noncooperative game approach to residential DR aggregation which demonstrates the interaction of self-interested households with a DRA in order to integrate two separate subproblems of home energy management systems with DR aggregation tasks. The aggregation of just homogenous appliances using queuing theory is described in [193] in which the final goal is to determine the aggregation scale that is needed for a specific amount of load reduction.

4.7.2 Challenges of DR aggregation

Few publications address the real requirements and challenges of DR aggregation. Table 4-3 summarizes some of the main issues related to the DR aggregation of residential customers [194-198].

Providing a proper infrastructure for the bidirectional data transfer between the DRA and system operator and between DRA's data centre and the customers is the fundamental step for DR implementation and imposes the main costs to DRAs [194].

Table 4-3 Managerial and technical challenges of DR aggregation

	Problem	System operator	DRA	Users
Management	Design of customer engagement programs	✓	✓	
	Investment in AMIs and SMs	✓	✓	

	Specifications of contracts between DRAs and users		✓	✓
	DRA participation in energy and ancillary markets	✓	✓	
	Investment in smart appliances			✓
	Competition among DRAs	✓	✓	✓
	Approaches for remunerating customers		✓	✓
	Customer behavioural factors/willingness affecting DR		✓	✓
	Customer clustering	✓	✓	✓
Technical	Communication infrastructure with bidirectional data transfer	✓	✓	✓
	Latencies	✓	✓	
	DR distribution over the network/voltage violations due to the asymmetric DR balancing between the phases	✓	✓	
	DRA control over user's appliances		✓	✓
	Cold load pick-up, lead and rebound effects	✓		
	Customer willingness to participate in DR		✓	✓
	Standards and communication protocols	✓	✓	✓

With the deployment of AMIs, it can be a proper solution for various purposes including DR aggregations. However, the cost associated with the investment in such infrastructures and other necessary equipment that DRAs need for the DR management should be considered properly. Besides, many of the work reported in the literature requires an extensive message exchange between the DRA and end-users. Hence, scalability issues may arise in a large scale deployment [190]. Centralized and de-centralized approaches that differ in computational complexity, optimization techniques and execution times may be considered for these purposes [199]. Latencies are another challenge associated with AMIs as they can affect both the grid stability and performance [200]. This is particularly the case when fast response times are required or frequent control updates are issued, for instance when providing spinning reserves by DR resources [201]. Eventually, installed AMIs and

SMs in most cases belong to other entities like distribution system operators or retailers. So, the use of them by DRAs raises questions about privacy and security issues.

DRAs usually have no concept of the network model and completely neglect the distribution of DR across the network and its consequences [196]. One of these problems is the increased consumption before and after the DR event usually referred to as “lead” and “rebound” effects [113]. The lead effect is caused by the expectation of a DR event, for example, a building may be precooled to reduce the afternoon air conditioning load. After the termination of a DR event, the increase in consumption compared to the baseline shows the rebound effect. Also, it is possible that phase unbalance arises during DR periods due to asymmetrical DR distribution on phases. Another physical challenge is that regulation service generally must be capable of both supplying and absorbing power [113].

Beside the system-wide effects, the DRA’s control on user’s appliances or its interaction with smart appliances and HEMS is an important matter that needs to be addressed. As far as DR programs are concerned, both incentive- and price-based DR face challenges if they are to be implemented in large scale. DR programs usually disturb the user’s comfort. Hence, attractive incentive-based programs are needed to persuade users to get involved in DR actions. Many existing peak demand management programs that utilize DLC are disruptive and can have significant effects on the end-users [201]. For instance, in the hottest days the load is usually the highest, so, any interruption in the air conditionings means curtailing them when their services are most in demand. Moreover, price-based programs generally introduce reliability issues and may result in new peaks of demand by shifting power from expensive peak hours to cheaper off-peak hours. Also, end-users may

be hesitant to participate in DR programs with dynamic pricing due to their complexity [202].

Maximizing the profit by offering in different markets is another challenging task for DRAs. Several sources of uncertainty add to the problem complexity. They can also incur penalties if the declared amount of DR cannot be met.

Another issue arises when calculating the accurate amount of load reduction based on a baseline, which demands the proper methods for modelling the baseline. Customer behaviour, usage pattern, and willingness to participate in DR are other important issues that significantly affect DR policies. The load patterns of residential customers usually show high variability and variety, which makes them different from the industrial and commercial load shapes. So, clustering of customers to different groups that show similar patterns is a necessary step for DRAs in the program targeting and customer engagement which can also benefit them in the optimal market participation.

4.7.3 Application of the proposed method

In summary, there is a need to define a proper structure for DR aggregation. The proposed method in this chapter can facilitate the DR aggregation from residential customers and address one of the practical challenges that has been less addressed in the literature. By dividing the customers into different groups based on their variabilities over time, a DRA will be able to offer suitable DR programs to its customers. This approach can lead to proposals that are less disruptive to the customer's comfort and probably achieve the highest overall benefits for the customers and the system.

As mentioned in the previous sections, historically, the lack of data was the main barrier to adopting the clustering methods for categorizing the customers. However, the new proposals demonstrate the ability of smart meter data in clustering of customers for demand response aggregation.

4.8 Summary

In this chapter, using a combination of SAX technique as a dimensionality reduction method and hierarchical clustering algorithm, daily load shapes of customers were segregated into certain clusters. Also, the stability of consumption pattern was investigated using the entropy concept.

The results of clustering and entropy calculation provide insights for offering DR programs to the electricity customers. First of all, the major load curves which represent the main consumption habits can be determined as shown in Fig. 4.9 and Fig. 4.10. This, in turn, will help the utilities to build DR programs or customized time of use tariff structures based on the load patterns. Furthermore, the entropy analysis enables them to divide their customers to various groups which can be targeted differently. For example, low entropy customers whose their load peaks happen at the same period of system peak are good candidates for price-based DR. On the other hand, variable users can be targeted by suitable incentive-based DR.

The case studies using a large number of daily load patterns demonstrate the applicability of the proposed method for DR management. Especially, the method is able to capture the major consumption patterns of the households and overcome the problems of the previous studies [54], which mainly dealt with the original daily load curves of residential

customers that varies a lot on the daily and seasonal bases. Use of SAX method can produce superior results, since the residential consumption patterns, which inherently display a lot of variability, are transformed to a limited number of SAX words.

5 Investigating the Effect of Household Characteristics on Consumption Patterns

5.1 Background and Motivation

In Chapter 2, the importance of survey data in understanding the effect of building characteristics, appliance use, and socio-demographic features on residential consumption pattern/magnitude were briefly explained. Pilot projects such as CER trail, which include both the load data of smart meters and the results of various survey questions, provide this great opportunity to understand the link between the household characteristics and consumption patterns.

In this chapter, using different data mining techniques, a methodology is proposed to investigate the relationship between the household and dwelling features (HDFs) and the consumption patterns of users. The relationship between the HDFs with their total/average/maximum consumption has vastly studied in the literature [183] [184]. On the other hand, the effect of HDFs on consumption patterns of customers has not been investigated in detail and it is still considered as a new area of research. In the literature, a few studies addressed this topic. Ref. [56] uses the survey and consumption data of 103 dwellings and investigates the correlations between the home characteristics and the seasonal representative curves of customers which are obtained by clustering. However, the customers are segmented into only two clusters in each season. McLoughlin *et al.* [75] compare the results of three clustering techniques to extract a set of profile classes which

can represent the households' consumption patterns. Then, these curves are linked to HDFs using a classification technique. There are some ambiguities in their procedure mainly in the formation of profile classes. Also, the existence of relatively high number of clusters produces uncertainties in the interpretation of results. Furthermore, in [74], using clustering, a set of comparison groups are created based on the HDFs in order to compare the electricity usage of each customer with its corresponding comparison group. This can help in encouraging customers to decrease their energy consumption.

The proposed method in this chapter uses a three-stage methodology to analyze the impact of HDFs on usage patterns of electricity customers. In summary, the contributions of the chapter are:

- A three-stage structure is defined to associate the HDFs to the consumption behavior.
- A variable selection methodology along with two statistical tests is employed for the proper selection of HDFs which can affect the consumption patterns.
- The effect of HDFs on consumption patterns is studied through a suitable classification method.

The following sections are organized as follows. Section 5.2 describes stages of the method. Sections 5.3, 5.4, and 5.5 are dedicated to the illustration of modules and accordingly elaborate the clustering method, variable selection procedure, and final variable selection and classification process. Section 5.6 reports the obtained results and their corresponding interpretations. Finally, Section 5.7 summarizes the chapter findings.

5.2 Stages of the Method

Fig. 5.1 explains the structure of the proposed method. It consists of three main parts including the clustering, variable selection and transformation, and final variable selection and classification modules.

Clustering: The first module is dedicated to clustering of customers based on their consumption data. In this chapter, we only use the weekly consumption data. The similar analysis can be applied to the weekend dataset. Firstly, using a feature definition approach, the data of each customer is represented by seven features. Then, customers are clustered by applying a K-means clustering and the best number of clusters is decided based on the values of four different CVIs. The output of this module is the cluster membership for each customer.

Variable selection and transformation: In this module, firstly, the relevant survey questions are selected and are transformed to suitable formats. Some numerical variables are converted into categorical variables and the categories (levels) for each categorical variable are decided. In the next step, based on the distribution of the customers among each level of the variables, the variables are further changed into new ones to correctly represent the nature of survey data. Finally, a statistical test is applied on the variables to determine possible strong association between two different variables.

Final variable selection and classification: This module involves two main stages. At first, the correlations among the survey variables and the cluster memberships are investigated in order to select the final variables. Based on this analysis, some variables are dropped.

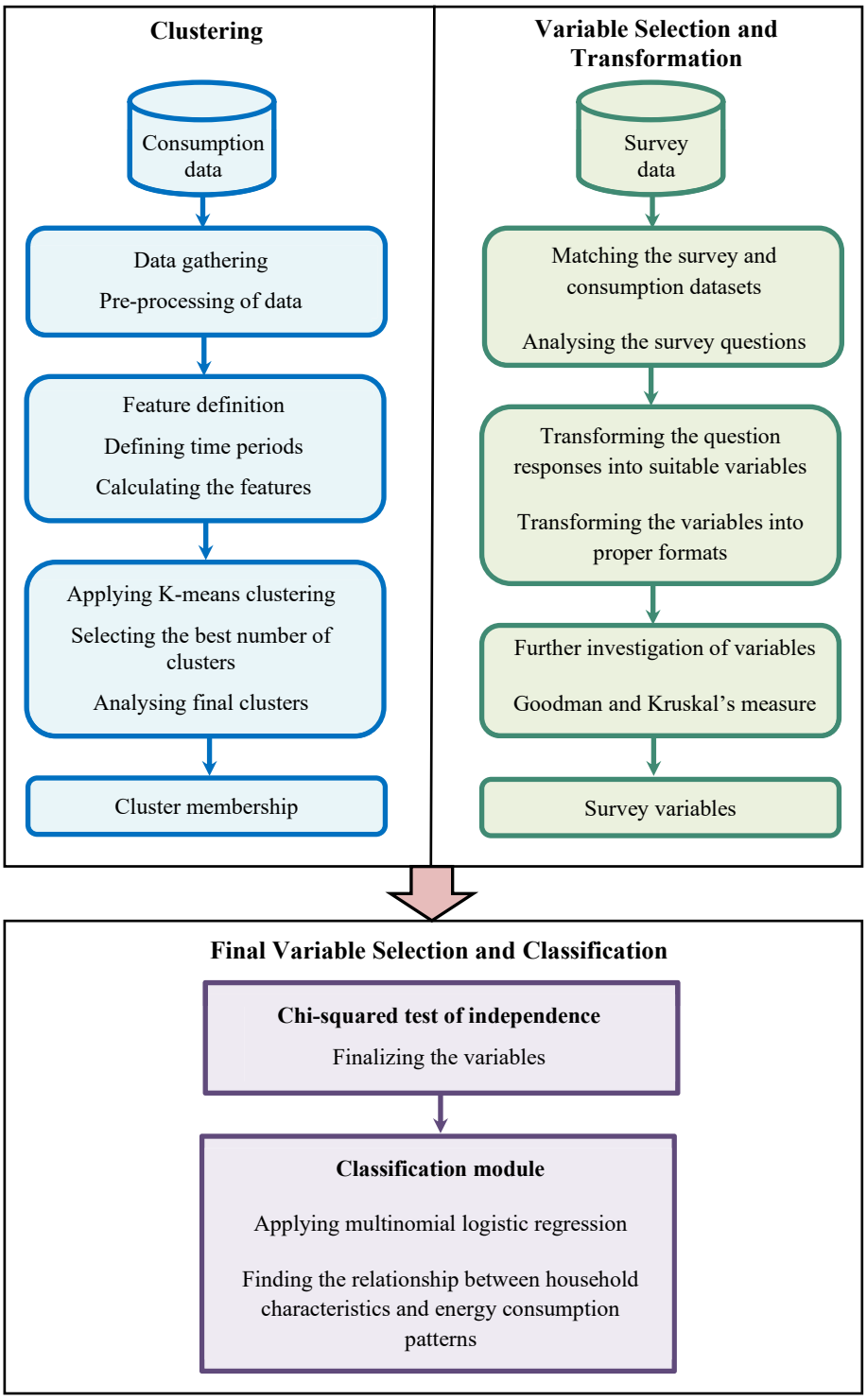


Fig. 5.1 Stages of the method including clustering, variable selection, and classification modules

In the next stage, a multinomial logistic regression (MLR), as a suitable classification technique, is utilized to find out the effect of HDFs on the cluster membership. In our study, the response variable is the cluster membership for each customer and the independent variables are survey variables. Each cluster has unique features and represents specified consumption patterns. Therefore, by using the classification, the effect of HDFs on consumption patterns can be found out.

5.3 Clustering

Before applying the clustering algorithm, using a feature definition approach, the weekly consumption of each customer during the whole year is transformed into a set of features. Motivated by the work in [27] and using the time periods which were found out in the last section (shown again in Table 5-1), seven features are defined as described byTable 5-2.

Table 5-1 The main time periods of household activity during the day

i	**Recording number**	**Corresponding time**
1	3- 14	1 am- 7 am
2	15-18	7 am- 9 am
3	19-33	9am- 4:30 pm
4	34-45	4:30 pm- 10:30 pm
5	46-2	10:30 pm- 1 am

Table 5-2 Defined features for each customer

No.	Attribute	Description
1-5	$P_{RAVG,i} = \frac{P_i}{\hat{P}}$	Relative average power in each time period
6	$\sigma = \frac{\sum_{i=1}^{5} \sigma_i / P_i}{5}$	Mean relative standard deviation
7	$SS = \sum_{i=1}^{5} \left\| P_{W,i} - P_{S,i} \right\| / P_i$	Seasonal score

$\hat{P} =$ Daily mean power over the entire year

$P_i =$ Mean power in time period i over the entire year

$\sigma_i =$ Standard deviation of load data in time period i over the entire year

$P_{W,i} =$ The mean power of time period i over the winter

$P_{S,i} =$ The mean power of time period i over the summer

From the definitions of these features, it can be seen that features 1 to 5, represent the consumption values during each time period which are normalized based on the daily mean power over the entire year. This allows comparing the consumption of a household during different time intervals and understanding the occupants' tendency in using electricity in different periods of the day. Feature 7 accounts for the seasonal changes in the consumption, hence, differentiating between those customers who show a high seasonal change in their usage and those customers with lower seasonal changes. This feature is also normalized since without normalization, the customers with lower demands will display greater changes than the customers with heavier demands. Finally, feature 6 reflects the variations in the

consumption values during the whole year. Although these variations can be caused by seasonal changes, they can also be related to the usage behavior of the customer.

For each customer these seven features are calculated based on the weekly consumption values. In the next step, a K-mean clustering algorithm is applied on these features to segregate customers into separate groups and the best number of clusters is decided using proper CVIs.

5.4 Variable Selection

There are 3418 customers that belong to both consumption dataset and survey data. The variables that are initially selected for the study are summarized in Table 5-3. This table also shows the number of customers that belong to each category (level) for each variable.

5.4.1 Considerations about the variables

The variables in Table 5-3 are extracted from the survey data. However, they cannot be directly used for the analysis. Further assumptions and consideration are needed to transform them into suitable formats. In the following, the descriptions of some of these variables and considerations about them are elaborated.

Income: Almost half of the customers refused to answer the question about their yearly income. In addition, among those that responded the question, many indicated the income as weekly (less than 15,000) which makes it impossible to infer the total yearly income. Therefore, after considering only those customers that reported their yearly income, the number of customers reduced to 932 customers. Regarding this issue, the analysis was performed without the inclusion of income. The income is usually considered as an

Table 5-3 Selected survey variables

Variable	Classes (number of customers)
Dwelling characteristics	
Dwelling type	Apartment (58), Semi-detached (1026), detached (923), terraced (485), bungalow (919)
Dwelling age	Less than 10 years (631), 10-30 years (987), 30-60 years (1205), More than 60 years (595)
Number of bedrooms	1 (38), 2 (289), 3 (1509), 4 (1187), 5 (386), 6 (9)
Heating water by electricity?	Yes (1952), no (1466)
Cooking type in home	Electric (2376), Non-electric (1042)
Proportion of energy-efficient light bulbs (CFL)	None (739), About a quarter (900), About half (576), About three quarter (570), All (633)
Proportion of double glazed windows.	None (279), About a quarter (66), About half (101), About three quarter (91), All (2881)
Insulated walls	Yes (1955), No (1068), Don't know (395)
Socio-demographic information	
Education of head of household	None or primary (440), Secondary to intermediate (1544), Third level (1245), Refused to answer (189)
Social class	AB (501), C1 (908), C2 (569), DE (1316), F(87), refused (37)
Employment status	Employee (1582), Retired (1088), Self-employed (424), unemployed (285), carer (39)
How many people over 15?	1 (817), 2 (1663), 3 (502), 4 (312), 5 (100), 6(20), 7 (4)
Yearly household income	Less than 15,000 (), 15,000 - 30,000 (), 30,000 -50,000 (), 50,000 -75,000 (), More than 75,000 ()
How many people under 15?	0 (2478), 1(407), 2 (322), 3 (155), 4 (45), 5 (9), 6 (2)
Age of respondent	18-35 years (323), 36-55 years (1522), 56-65 years (738), Above 65 years (813)
Number of each appliance	
Washing machine	0 (61), 1 (3333), 2 (24)
Tumble dryer	0 (1095), 1 (2319), 2 (4)
Dishwasher	0 (1162), 1 (2248), 2 (8)
Electric cooker	0 (788), 1 (2618), 2 (11), 3 (1)
Electric heater	0 (2353), 1 (823), 2 (181), 3 (61)
Standalone Freezer	0 (1704), 1 (1649), 2 (62), 3 (3)
Water pump	0 (2757), 1 (645), 2 (16)
TVs< 21 in.	0 (1208), 1 (1334), 2 (614), 3 (194), 4 (68)
TVs > 21 in.	0 (545), 1 (1739), 2 (852), 3 (205), 4 (77)
Desktop computers	0 (1801), 1 (1517), 2 (85), 3 (10), 4 (5)
Laptops	0 (1599), 1 (1428), 2 (288), 3 (70), 4 (33)
Number of game consoles	0 (2292), 1 (757), 2 (275), 3 (70), 4 (24)
Attitudes toward energy savings	
Possibility to make major changes in electricity use	Scales 1 to 5 1: Strongly agree 5: Strongly disagree
We can reduce our electricity bill by changing the way the people we live with use electricity	
It is too inconvenient to reduce our usage of electricity	
I do not have enough time to reduce my electricity usage	

important variable in consumption analysis [203], [204]. However, it generally shows a correlation with the social class and/or the education level [205] and hence, we assume that its effect can be reflected through these variables.

Social class: the customers are assigned to a social class based on the socio-economic classification produced by UK Office for National Statistics (ONS) [206]. The definitions of these classes are given in Table 5-4**Error! Not a valid bookmark self-reference.**. Social classes "C1" and "C2" are considered usually in the same category and hence, they are merged as class "C".

Table 5-4 Definition of social classes

Social class	Description
AB	Higher & intermediate managerial, administrative, professional occupations
C1	Supervisory, clerical & junior managerial, administrative, professional occupations
C2	Skilled manual occupations
DE	Semi-skilled & unskilled manual occupations, Unemployed and lowest grade occupations
F	Farmers

Number of bedrooms: since the number of dwellings with one bedroom is very low, the levels "1" and "2" are merged.

Number of people over 15: three levels defined for this variable including "1", "2", "3 and more".

Number of people under 15: three levels defined for this variable including "none", "1", and "2 and more".

Washing machine/ Tumble dryer/ Dishwasher/ Electric cooker/ Water pump/ Standalone freezer/ Laptops/ Desktop computers/ Game consoles: from Table 5-3 it is obvious that only a few numbers of households have more than one of these items. Therefore, these variables are considered as categorical variables with two levels showing the "presence" or "no presence" of each of these items.

Cooking type: the variables "cooking type" and "presence of electric cookers" are completely associated and so, we drop the former.

Respondent age: respondent age (which is supposed that accords with the age of the head of household) shows a strong correlation with the building age. Thus, this variable is also removed from the analysis.

Proportion of energy-efficient light bulbs: for this variable, three levels are defined: "none", a level comprising of the levels "about a quarter" and "half", and a level including the levels "about three quarters" and "all".

Proportion of double glazed windows: since most of the dwellings have double glazed windows, just two levels are defined for this variable: one consisting of "none" and "about a quarter" and another one including the other three levels.

Number of TVs < 21 inch and number of TVs > 21: Three levels are considered for each of these variables including "none", "1", and "2 and more".

Change/ Reduce Bill/ Inconvenient/ Enough time: three levels are defined for each of these variables including "Agree", "Disagree", and "Neutral".

Employment status: The levels "Employee" and "Self-employed" are merged.

The resulting selected variables are reported in Table 5-5. In this table, the base category is shown in bold.

Table 5-5 Selected variables after the initial evaluation

Variable	Classes (number of customers)
Dwelling characteristics	
Dwelling type	**Apartment (58)**, Semi-detached (1026), detached (923), terraced (485), bungalow (919)
Dwelling age	**Less than 30 years (1618),** 30-60 years (1205), More than 60 years (595)
Number of bedrooms	**1 or 2 (327)**, 3 (1509), 4 (1187), 5 or 6 (395)
Heating water by electricity?	Yes (1952), **no (1466)**
Proportion of energy-efficient light bulbs	**None (739),** A quarter or half (1476), Three quarters or all (1203)
Proportion of double glazed windows.	**None or a quarter (345),** All or three quarters or half (3073)
Insulated walls	Yes (1955), No (1068), **Don't know (395)**
Socio-demographic info	
Education of HoH	**None or primary (440)**, Secondary to intermediate (1544), Third level (1245), Refused to answer (189)
Social class	**AB (501)**, C (1477), DE (1316), F(87), refused (37)
Employment status	Employed (2025), Retired (1098), **unemployed (295)**
How many people over 15?	**1 (817)**, 2 (1663), 3 and more (938)
How many people under 15?	**0 (2478),** 1(407), 2 and more (533)
Appliances	
Washing machine	**No (61)**, Yes (3357)
Tumble dryer	**No (1095)**, Yes (2323)
Dishwasher	**No (1162)**, Yes (2256)
Electric cooker	**No (788)**, Yes (2630)
Number of electric heaters	**0 (2353)**, 1 (823), 2 or more (242)
Standalone Freezer	**No (1704)**, Yes (1714)
Water pump	**No (2757)**, Yes (661)
Number of TVs< 21 in.	**0 (1208)**, 1 (1334), 2 and more (876)
Number of TVs > 21 in.	**0 (545)**, 1 (1739), 2 and more (1134)
Desktop computers	**No (1801)**, Yes (1617)
Laptops	**No (1599)**, Yes (1819)
Game consoles	**0 (2292)**, Yes (1126)
Attitudes toward energy savings	
Change: Possibility to make major changes in electricity use	Agree (1702) , **Neutral (721)**, Disagree (995)
Reduce bill: We can reduce our electricity bill by changing the way the people we live with use electricity	Agree (2655) , **Neutral (509)**, Disagree (254)
Inconvenient: It is too inconvenient to reduce our usage of electricity	Agree (605) , **Neutral (415)**, Disagree (2398)
Enough time: I do not have enough time to reduce my electricity usage	Agree (521), **Neutral (379)**, Disagree (2518)

5.4.2 Determining the association among variables

For numerical variables, the degree of correlation between two variables can be easily calculated using different correlation measures such as Pearson correlation coefficient [207]. For categorical variables, defining the correlation between two variables is more complicated. In this case, the correlation can be understood in terms of "strength of association" and "significant test" [208]. In this study, Goodman and Kruskal's tau (GKT) measure and Chi-squared test are used for the former and latter cases to determine the final variables which will be used for the analyses.

It is better to find the variables that show a strong association with each other since they might affect the classification results. In addition, finding the relative association among various variables can be beneficial in understanding the household characteristics. The GKT measure is an asymmetric measure which decides whether a variable is associated with another variable or not and also measures the strength of the relationship. The asymmetry results from the fact that the measure is based on the fraction of variability in the categorical variable Y that can be explained by the categorical variable X. Here, the mathematical formulations of this measure are not explained. The interested readers can refer to [207] and [209] for theoretical backgrounds and implementation in R, respectively.

By applying the GKT, the set of variables which show an association (above 0.1 in our case) are determined. Some of these relationships, for example, the building type with the number of bedrooms, or the number of people under 15 and the number of game consoles, can be intuitively understood. Table 5-6 shows the values of this measure for the associated variables.

Table 5-6 GKT values for the variables with a relatively high strength of association

Variables	$\tau(x, y)$	$\tau(y, x)$
Building type , Number of bedrooms	0.103	0.059
Building age, Insulating wall	0.129	0.06
Education, Social class	0.071	0.104
Employment, Social class	0.232	0.266
Social class, Respondent age	0.1	0.103
Employment, Respondent age	0.266	0.333
Number of game consoles, Number of people under 15	0.156	0.171
Number of bedrooms, Presence of dishwashers	0.139	0.048
Employment, Number of game consoles	0.106	0.079
Respondent age, Number of game consoles	0.135	0.09

Among these variables, employment shows a relatively strong association with the social class and respondent age. However, the values of the measure are not big enough to omit some of the variables and we keep all of them to study their effects on the cluster membership.

5.5 Final Variable Selection and Classification

5.5.1 Final Variable Selection

In this section, “Chi-squared test of independence” is illustrated which will be used to find out the correlation among the independent variables (survey variables) and response variable (cluster membership).

Chi-squared test of independence is a well-known measure which determines the statistical dependence between two categorical variables. Considering two categorical variables X and Y, it starts with a hypothesis test as the following:

- H_0 (Null hypothesis): The variables X and Y are independent.
- H_1 (Alternative hypothesis): The variables X and Y are dependent.

To analyze the data, firstly a contingency table of two variables is built. For example, for the categorical variables X and Y with three and four levels respectively, the contingency table is:

	$\boldsymbol{Y_1}$	$\boldsymbol{Y_2}$	$\boldsymbol{Y_3}$	$\boldsymbol{Y_4}$
$\boldsymbol{X_1}$	f_{11}	f_{12}	f_{13}	f_{14}
$\boldsymbol{X_2}$	f_{21}	f_{22}	f_{23}	f_{24}
$\boldsymbol{X_3}$	f_{31}	f_{32}	f_{33}	f_{34}

where $f_{i,j}$ represents the observed frequency count of observations belonging to i-th category of X and jth category of Y (the value of ij cell of contingency table).

The Chi-squared test statistics can be defined as:

$$\chi^2 = \sum_{i=1}^{nrows} \sum_{j=1}^{ncols} \frac{(f_{i,j} - e_{i,j})^2}{e_{i,j}} \tag{5.1}$$

where $e_{i,j}$ is the corresponding expected count in ij cell of the contingency table if X and Y were independent. For a given cell, the expected value is calculated as follows:

$$e_{i,j} = \frac{row.sum * col.sum}{Total\ number\ of\ customers} = \frac{\sum_{h=1}^{nrows} f_{h,j} \times \sum_{h=1}^{ncol} f_{i,h}}{N} \tag{5.2}$$

The calculated Chi-squared test is compared with a critical value obtained from Chi-squared table and the resulting p-value is calculated. If the calculated p-value is small enough (usually less than 0.05), the null hypothesis is rejected. It means that the variables are correlated.

This test will be applied on all the variables to decide on the final variables which will be used in the classification procedure.

5.5.2 Classification method

A) Multinomial Logistic Regression

Given a set of m independent variables X, *logistic regression* models the probability that the categorical response variable (Y) belongs to a particular category. When the response variable is dichotomous (binary) or can take only two levels, the logistic function is used to produce the probability that an observation Y_i, is classified as 1:

$$p(X) = Pr(Y_i = 1|X_i) = \frac{1}{1 + e^{-(b_0 + b_1 X_i)}} \tag{5.3}$$

The probability of Y belonging to 0, is $p(Y_i = 0|X_i) = 1 - p(X_i)$.

When there are more than two categories (clusters in our case), multinomial logistic regression (MLR) is used to calculate the probabilities. For K categories, the model contains $K - 1$ logistics regression equations, when $Pr(Y_i = j|X_i)$ describes the probability that a particular observation, Y_i, belongs to the category j when compared against the reference category K [30]:

$$Pr(Y_i = K|X_i) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{b^k X_i}} \tag{5.4}$$

$$Pr(Y_i = j|X_i) = \frac{e^{b^j X_i}}{1 + \sum_{k=1}^{K-1} e^{b^k X_i}} = Pr\,(Y = K|X_i) e^{b^j X_i}, \quad j = 1, 2, \dots, K-1 \tag{5.5}$$

where $b^j = (b_0^j, \dots, b_m^j)$ represents the coefficients for jth category .The observation is allocated to a category with which it has the highest probability.

The sets of equations in (5.5) can also be rewritten as:

$$log\left(\frac{Pr(Y=j|X)}{Pr(Y=K|X)}\right) = b^j X = b_0^j + b_1^j x_1 + \cdots + b_m^j x_m \quad (5.6)$$

where x_1 to x_m are the values of independent variables for the selected observation. The left-hand side of this equation is usually called *logit* or *log-odds*.

The MLR model is usually fitted by the maximum likelihood to estimate the parameters. The intuition behind the maximum likelihood is to estimate the b_j parameters that plugging them into the model yields the closest results to the actual categories that the observations belong to them [210]. The likelihood function is defined as:

$$\ell\ (b) = \prod_{i=1}^{n}\prod_{k=1}^{K} P(Y_i = k)^{g(Y_i=k)} \quad (5.7)$$

where $\ell\ (b)$ is the likelihood function, $b = (b_1, b_2, \dots, b_K)$ is the set of coefficients to be estimated for all classes, and n is the number of observations. $g(Y_i = k)$ is an indicator which is equal to 1 if the observation Y_i belongs to class k, and 0 otherwise. The aim is to maximize $\ell\ (b)$.

Instead of (5.7), usually log-likelihood function is used which is expressed as:

$$log\big(\ell\ (b)\big) = \sum_{i=1}^{n}\sum_{k=1}^{K} g(Y_i = k).log\ (P(Y_i = k)) \quad (5.8)$$

B) Interpreting the results of MLR

According to the explanations in the last section, the coefficients obtained by the MLR should be interpreted based on the reference category. In this case, like the linear regression models, to interpret the coefficient b_i^j for one special variable x_i for the category j of the outcome, it is assumed that the other variables in the model are held constant. Therefore, it

can be said that the multinomial log-odds of preferring the category j to the reference category would be expected to change b_i^j units for the one unit increase in the value of x_i [211]. In our study in this chapter, the variables are categorical. For categorical variables, like the outcome, a reference level needs to be defined for x_i. Consequently, the b_i^j shows the units of change in multinomial log-odds when the corresponding level of x_i is compared against its reference level.

The results can also be interpreted based on the exponentiation of the coefficients which can be calculated as $\exp(b_i^j)$ for each coefficient b_i^j. Again if all the other variables are held constant, for the corresponding level of x_i, the odds of an observation belonging to category j (compared against the reference category of outcome) is $\exp(b_i^j)$ times the odds for the reference level of x_i (refer to (5.5)). It should be noticed that $\exp(b_i^j) < 1$ for $b_i^j < 0$ and $\exp(b_i^j) > 1$ for the $b_i^j > 0$.

By applying MLR on our data, a set of b_i^j values are extracted which shows the effects of survey variables on cluster membership.

5.6 Applications

This section explains the results after applying the proposed method on CER datasets. R software is used for the clustering, classification, and statistical tests.

5.6.1 Clustering results

A) Selecting the number of clusters

After calculating the defined attributes, the customers are clustered using a K-means clustering algorithm. The best number of clusters is decided using four cluster validity measures including: Davies-Bouldin, Dunn, Ball-Hall [212], and Silhouette measures. The decision rule for the Ball-Hall index is based on the "maximum difference". It means that the best value for cluster number is the one which corresponds to the greatest difference between two successive slopes [212].

The final clustering results will be used to associate the household characteristics with their consumption data. In order to be able to draw the meaningful interpretations, the final number of clusters should not be selected too small or too big. Considering this issue, the clustering results were evaluated for the cluster number changing from 5 to 10. The initial analysis of the clustering results indicates that there is a cluster with only 12 members. The members of this cluster are considered as outliers and those 12 customers are removed from the analysis. For the remaining customers, the clustering is carried out again. Fig. 5.2 shows the values of CVIs for different number of clusters. The Silhouette, Davies-Bouldin, and Ball-Hall measures indicate the best number of clusters as six. Only based on the Dunn index the best number of clusters is decided as eight. Therefore, we select the six as the best number of clusters.

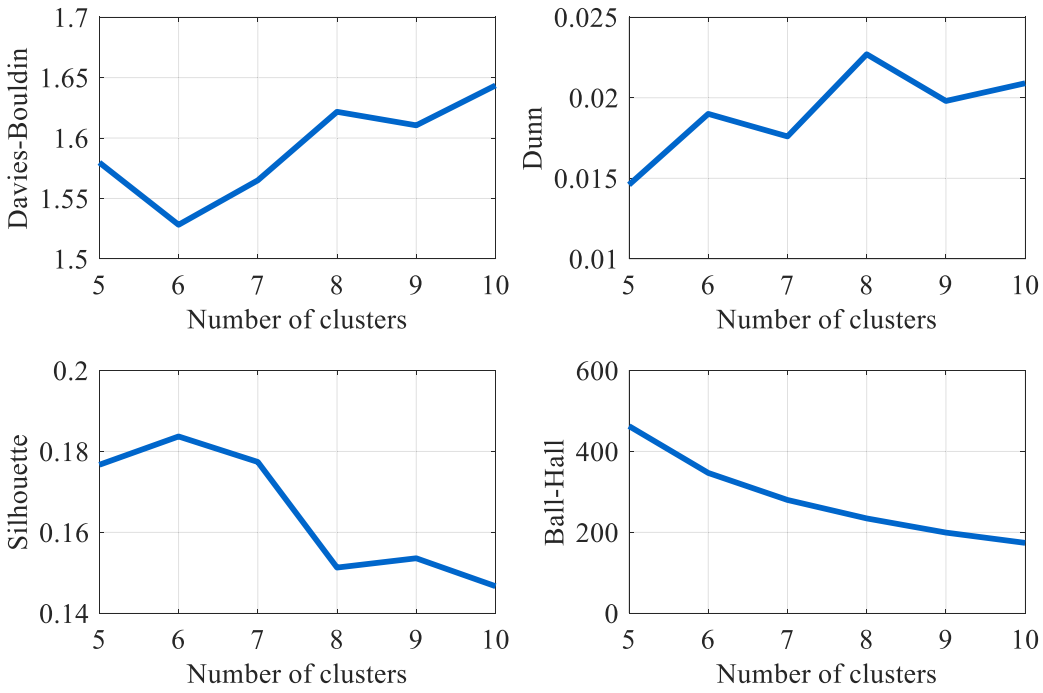


Fig. 5.2 Selecting the best number of clusters

B) Characteristics of the clusters

Table 5-7 shows the mean value of each of the defined features for each cluster.

Table 5-7 Characteristics of formed clusters

Cluster	No. of customers	Features						
		$P_{RAVG,1}$ (Overnight 2)	$P_{RAVG,2}$ (Morning)	$P_{RAVG,3}$ (Daytime)	$P_{RAVG,4}$ (Evening)	$P_{RAVG,5}$ (Overnight 1)	σ	SS
1	578	0.455	0.952	1.008	1.554	0.991	1.089	2.620
2	1037	0.474	0.837	1.035	1.537	0.999	0.905	0.967
3	535	0.423	1.568	0.967	1.529	0.760	1.082	1.424
4	957	0.431	0.760	0.948	1.669	1.109	1.036	1.699
5	208	0.497	1.030	1.007	1.482	1.004	1.364	4.016
6	91	0.594	0.890	0.976	1.448	1.058	1.442	6.352

This table shows that these attributes can characterize the customers based on their consumption values in different time periods and their corresponding seasonal changes as well as mean standard deviation. In particular, some of the characteristics of the clusters can be summarized as:

- Cluster 1: All of the attributes of this cluster have average values compared with other clusters.
- Cluster 2: This cluster has the lowest seasonal score and standard deviation among all clusters. Interestingly, it is the largest cluster by the number of members too.
- Cluster 3: the distinguishing feature of this cluster is the high energy use during the morning period which is the highest among all the clusters. On the other hand, this cluster has the lowest consumption during overnight period. Thus, the corresponding consumers of this cluster can be categorized as the morning users of electricity.
- Cluster 4: Among all the clusters, this group has the highest consumption values during the evening and overnight periods and the lowest consumption values during the morning and daytime. This suggests that the members of this cluster incline to consume more during the later hours of the day.
- Cluster 5: the values of σ and SS are high for this cluster. Nonetheless, similar to the cluster 6, this cluster contains a small number of customers.
- Cluster 6: this cluster displays a relatively high consumption during the 1 am to 7 am. However, the most defining characteristic of this cluster is the high seasonal score and high mean standard deviation. The former shows that the consumption of the customers belonging to this cluster significantly changes during different seasons. In addition, the high standard deviation implies that the consumption of the customers is widely spread.

5.6.2 Chi-squared test

Following the procedure in the last sections, a matrix is built which contains the survey variables for each customer and the corresponding cluster membership. Subsequently, the Chi-squared test of independence is applied to find out which variables show a correlation with cluster membership. The other variables which show no association with the clusters are dropped from the analysis.

Here, we explain the results for only one variable. The analysis is similar for other variables and is not repeated here.

The contingency table for the variable showing the percentage of CFL bulbs in the dwelling and different clusters is shown in Table 5-8.

Table 5-8 Contingency table for the CFL and cluster membership

	Cluster1	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**	**Cluster6**	**Total**
None	128	204	107	224	48	24	735
A quarter or half	254	400	237	443	96	42	1472
3 quarters or all	196	433	191	290	64	25	1199
Total	578	1037	535	957	208	91	3406

For this variable, the calculated p-value from the Chis-squared test is 3.825e-05 which is less than the cut-off value of 0.05. This confirms the correlation among this variable and the output variable.

Fig. 5.3 shows the contribution of each cell of contingency table to the total Chi-squared measure. In this figure, the diameter of the circle is proportional to the amount of cell contribution. It also shows the positive (blue) and negative (red) association between the corresponding row and column levels. For example, there is a strong negative

correlation between the row “Three quarters or all” and column “Cluster 4”. This is an interesting result as the cluster 4 has the highest consumption value among all the clusters during the night periods. Using CFL bulbs apparently has a negative association with the electricity usage in these hours.

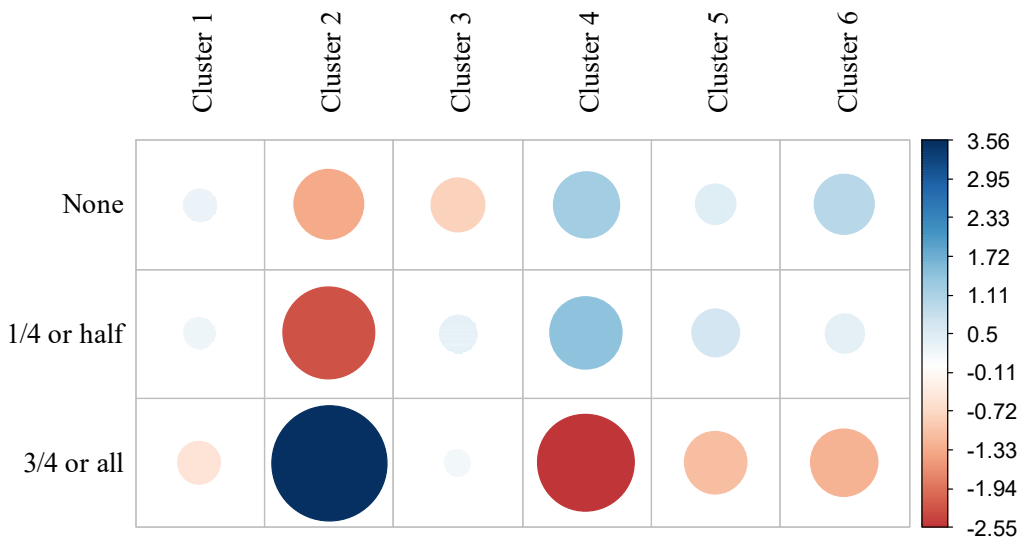


Fig. 5.3 Visualizing the contribution of each cell of contingency table to the Chi-squared value

Finally, the results were analyzed for all variables. The obtained results show that the p-values for these variables are significantly larger than the cut-off value: Change, Enough Time, Inconvenient, Water Pump, and Tumble Dryer. Therefore, these variables are not considered in the next step of the study.

5.6.3 Classification

After all of the pre-processing, selecting the variables, and clustering process, eventually, 23 household variables are remained. Each customer is identified by the values of these variables and the cluster membership.

We are now ready to apply the classification technique to identify the relationship between the variables and the cluster membership. Cluster 1 is selected as the reference class for all the analysis since its corresponding features have average values which make the interpretations and comparisons easier. For household variables, the levels and the base level were previously specified in Table 5-5.

Table 5-9 shows the results of MLR classification. Although the effect of all variables can be discussed based on their corresponding coefficients, we limit the discussions to those variables which are statistically significant (having a p-value less than 0.1).

- Cluster 2: This cluster shows the lowest SS and σ. The older buildings show a strong negative association with this cluster. This can be due to this fact that the older buildings usually have older occupants, who can be retired or have less regular schedules and consequently have higher SS and σ. On the other hand, the higher proportion of CFL and double glazed windows are positively related to this cluster. As it is expected, those households which have more adults and fewer children are more probable to belong to this cluster. It is shown by the positive b values for "Over 15- 2" and "Over 15- 3+" and negative values for "Under 15- 1" and "Under 15- 2+". It seems that those appliances that are considered as the base load (such as electric cooker, freezer, TVs, desktop computers, and game consoles) which their usage do not depend on the seasonal changes are positively correlated to this cluster. On the contrary, the higher number of electric heaters shows a strong negative association to this cluster (the value of b is -1.019 and -0.56 for "Num. Elec. Heater- 2+" and "Num. Elec. Heater- 1" respectively). The parameter "Reduce bill- Agree" also shows a positive effect on this

Table 5-9 Results of MLR classification

Variables	Cluster 2			Cluster 3			Cluster 4			Cluster 5			Cluster 6		
	b	**Exp (*b*)**	**Std. Err**	***b***	**Exp (*b*)**	**Std. Err**	***b***	**Exp (*b*)**	**Std. Err**	***b***	**Exp (*b*)**	**Std. Err**	***b***	**Exp (*b*)**	**Std. Err**
Dwelling Type- Bungalow	-0.029	0.971	0.45	-0.092	0.912	0.543	-0.202	0.817	0.447	0.123	1.131	0.588	1.124	3.076	1.152
Dwelling Type- Detached	-0.011	0.99	0.454	0.313	1.367	0.544	-0.123	0.884	0.451	0.395	1.485	0.591	1.385	3.996	1.151
Dwelling Type- Semi-det.	0.293	1.341	0.447	0.346	1.413	0.537	-0.084	0.919	0.445	-0.007	0.993	0.588	1.185	3.271	1.156
Dwelling Type- Terraced	0.262	1.299	0.453	0.24	1.272	0.545	-0.07	0.933	0.451	-0.284	0.753	0.605	0.849	2.338	1.175
Building Age- 30 to 60	-0.25	0.779*	0.14	-0.181	0.834	0.16	-0.263	0.769*	0.14	-0.16	0.852	0.22	0.303	1.353	0.34
Building Age- Above 60	-0.465	0.628***	0.17	-0.291	0.748	0.195	-0.587	0.556***	0.174	-0.034	0.967	0.249	0.462	1.587	0.373
Num. Bedrooms- 3	-0.183	0.833	0.208	0.118	1.125	0.246	0.159	1.172	0.218	-0.001	0.999	0.287	-0.327	0.721	0.4
Num. Bedrooms- 4	-0.176	0.838	0.23	0.013	1.013	0.271	0.13	1.139	0.239	-0.335	0.716	0.331	0.482	1.619	0.435
Num. Bedrooms- 5or6	-0.398	0.672	0.267	-0.638	0.528**	0.324	-0.131	0.877	0.276	-0.28	0.756	0.388	-0.477	0.621	0.599
CFL- 1/4 or half	-0.013	0.987	0.146	0.084	1.088	0.166	-0.058	0.943	0.142	0.07	1.072	0.216	-0.089	0.915	0.302
CFL- 3/4 or all	0.252	1.286*	0.151	0.105	1.111	0.174	-0.278	0.757*	0.151	0.004	1.004	0.232	-0.251	0.778	0.336
Double G Window- More than 1/4	0.379	1.461**	0.183	0.209	1.232	0.211	0.422	1.525**	0.186	-0.203	0.816	0.235	0.484	1.623	0.357
Ins. Wall- Yes	-0.119	0.888	0.137	0.076	1.079	0.16	-0.073	0.929	0.137	-0.134	0.875	0.208	0.046	1.047	0.284
Heat Water Elec.- Yes	-0.067	0.935	0.112	0.275	1.316**	0.13	-0.016	0.984	0.112	-0.113	0.893	0.173	-0.025	0.975	0.254
Education- Sec. to interm.	-0.04	0.961	0.179	0.182	1.2	0.22	-0.07	0.932	0.181	0.056	1.058	0.265	-0.083	0.921	0.36
Education- Third	-0.465	0.628**	0.198	-0.115	0.891	0.24	-0.495	0.609**	0.2	0.045	1.046	0.295	0.082	1.086	0.406
Social Class- AB	0.468	1.597	0.359	0.672	1.959*	0.402	1.235	3.438***	0.402	-0.194	0.824	0.477	-0.486	0.615	0.697
Social Class- C	0.522	1.686	0.33	0.367	1.444	0.376	1.142	3.134***	0.378	-0.115	0.891	0.422	-0.019	0.981	0.582
Social Class- DE	0.52	1.682	0.346	0.111	1.117	0.398	1.105	3.019***	0.393	-0.371	0.69	0.441	-0.262	0.769	0.604
Over 15- 2	0.312	1.366**	0.148	0.426	1.532**	0.172	0.115	1.122	0.146	-0.048	0.953	0.209	-0.169	0.845	0.294
Over 15- 3+	0.442	1.556**	0.184	0.449	1.567**	0.215	0.269	1.309	0.184	-0.169	0.844	0.29	-0.219	0.803	0.424
Under 15- 1	-0.689	0.502***	0.186	-0.079	0.924	0.2	-0.452	0.636**	0.186	-0.526	0.591	0.355	-0.4	0.67	0.519
Under 15- 2+	-0.66	0.517***	0.192	0.02	1.02	0.208	-0.249	0.78	0.189	0.192	1.212	0.304	-0.277	0.758	0.525
Employment- Retired	0.146	1.157	0.183	-0.13	0.878	0.215	0.035	1.036	0.184	0.315	1.37	0.274	0.202	1.224	0.384
Employment- Unemployed	0.127	1.135	0.239	-0.173	0.841	0.294	0.001	1.001	0.244	0.487	1.628	0.362	-0.806	0.446	0.718
Washing Machine- Yes	0.03	1.03	0.52	-0.283	0.754	0.578	-0.564	0.569	0.485	-1.05	0.35**	0.531	-2.326	0.098***	0.569
Dish Washer- Yes	-0.174	0.84	0.13	-0.02	0.98	0.149	0.187	1.206	0.132	0.044	1.045	0.198	-0.491	0.612*	0.282
Elec. Cooker- Yes	0.293	1.34**	0.128	0.194	1.214	0.148	0.286	1.332**	0.13	0.463	1.59**	0.205	0.053	1.055	0.279
Num. Elec. Heater- 1	-0.56	0.571***	0.131	-0.199	0.819	0.148	-0.224	0.799*	0.129	0.345	1.412*	0.19	0.603	1.827**	0.283
Num. Elec. Heater- 2+	-1.019	0.361***	0.229	-0.333	0.716	0.246	-0.449	0.638**	0.212	0.7	2.014***	0.268	1.515	4.551***	0.343
Freezer- Yes	0.486	1.626***	0.113	-0.048	0.953	0.129	-0.007	0.993	0.113	-0.172	0.842	0.175	-0.222	0.801	0.255
Num. TV less 21- 1	0.252	1.286*	0.132	0.082	1.085	0.152	0.24	1.272	0.131	-0.119	0.887	0.198	-0.266	0.766	0.287
Num. TV less 21- 2+	0.414	1.514***	0.156	0.237	1.267	0.177	0.223	1.25	0.157	-0.601	0.548**	0.268	-0.806	0.447**	0.41
Num. TV bigger 21- 1	0.117	1.124	0.168	-0.316	0.729*	0.184	0.124	1.132	0.169	-0.415	0.66*	0.233	-0.332	0.717	0.329
Num. TV bigger 21- 2+	0.312	1.366	0.194	-0.421	0.656*	0.216	0.366	1.442*	0.195	-0.529	0.589*	0.289	-1.102	0.332**	0.448
Desktop- Yes	0.311	1.365***	0.12	0.118	1.126	0.137	0.227	1.255*	0.121	0.015	1.015	0.189	-0.536	0.585*	0.299

Table 5-9 Results of MLR classification (Cont'd)

Laptop- Yes	0.079	1.082	0.125	-0.081	0.922	0.143	0.008	1.008	0.126	0.025	1. 025	0.193	-0.236	0.79	0.292
Game consoles- Yes	0.367	1.444**	0.154	0.163	1.177	0.172	0.093	1.097	0.155	-0.287	0.751	0.269	0.39	1.478	0.413
Reduce Bill-Agree	0.294	1.342*	0.152	0.257	1.293	0.178	0.183	1.201	0.152	0.005	1.005	0.219	0.541	1.718	0.354
Reduce Bill-Disagree	-0.317	0.728	0.241	0.093	1.098	0.267	0.06	1.062	0.226	-0.253	0.776	0.332	0.16	1.174	0.501

*** P < 0.01, ** P < 0.05, * P < 0.1

cluster, though statistically it is less significant.

- Cluster 3: This cluster manifests the morning usage of electricity. Dwellings with a high number of bedrooms ("Num. Bedrooms- 5 or 6") are less probable to belong to this cluster. This is understandable since such buildings usually tend to use more electricity during the evenings and nights too. Heating water by electricity shows a positive effect on this cluster. The social class AB is positively related to this cluster. Having higher number of adults also positively affects the probability of membership of customers to this cluster. It can be attributed to regular working habits of them during the weekdays. The other important factors are the number of TVs bigger than 21 inch with corresponding negative coefficients which can be associated to the more use of TVs in the evening periods rather than mornings.
- Cluster 4: The customers belonging to this cluster use the highest electricity in the evening and night. Older dwellings show a negative association to this cluster which can be again related to the residents of such buildings. Households with the higher proportion of CFL are less probable to be a member of this cluster. This negative association was shown previously in section 5.6.2.1. All the categories of social classes (which are compared against the base category that is social class F) display a very strong positive association with this cluster. In addition, the higher social classes have larger coefficients. It is an important result which shows the effect of social class on the

electricity usage in these time periods. The presence of children in the households is negatively associated to this cluster. The other important factors are electric cookers, TVs, and desktop computers which indicate a positive association with this cluster and electric heaters that display a negative relationship.

- Cluster 5: This cluster has a relatively high σ and SS. Therefore, not surprisingly, those factors that have a negative association with cluster 2 usually have a positive association with this cluster and vice versa. This is especially true for electric heaters and TVs, which respectively associate positively and negatively to this cluster. The coefficient for the washing machine (as a base load) is also negative.
- Cluster 6: This cluster has the highest σ and SS and a relatively high consumption during 1 am-7 am period. Again, the electric heater is the main factor affecting the membership to this cluster with very high positive coefficients. On the other hand, washing machine, dish washer, TVs, and desktop computers have a negative relationship with this cluster.

5.7 Implications of the study findings

Using the presented methodology in this thesis, we examined the association among various variables with the clusters of customers. The process segments customers into distinctive clusters, each one representing different consumption habits, degree of daily variability, and seasonal changes, thereby, allowing studying the impact of HDFs on consumption patterns. In terms of appliances and electric devices, the outcomes demonstrate the effect of different base load appliances, TVs, and heaters on the time and variability of consumption. The results also reflect the impact of household composition as well as certain physical dwelling characteristics on these factors. While the findings of this study are

beneficial for research activities, they have also several potentials for real-world applications as described in the following.

Demand-side management is usually classified into three main categories including the energy efficiency, conservation (behavioural change), and DR programs [113]. The proposed approach in this thesis can help utilities, policymakers, and customers in each of these directions. Especially, in recent years, research has increasingly considered residential DR programs. Traditionally, DR programs were applied for large users such as industrial and commercial customers. The introduction of advanced metering infrastructures, smart meters, and the concept of smart homes has opened up the possibility of implementing residential DR programs. As discussed in [124], in addition to technical challenges, several managerial challenges need to be addressed for the proper utilization of residential DR settings. Customer clustering, design of customer engagement programs, and customer behavioural factors along with the customer willingness for DR are among the most important ones. The knowledge obtained through this research can help utilities, retailers, or DR aggregators to devise suitable DR programs for their customers. Incentive- and price-based DR programs [112] can be offered to the customers depending on the cluster membership or the household's consumption habits, the household's daily and seasonal variability, and its composition. Eventually, this can lead to the load shifting and peak shaving of the electricity network.

5.8 Summary

The availability of smart meter data allows the analysis of consumption behavior of electricity customers which in turn, can contribute to the more energy efficient and green networks.

In this chapter, using various data mining techniques including statistical tests, feature definition, clustering and classification, the effects of household's socio-economic factors and physical features of dwellings on consumption patterns of customers are investigated. Using this analysis, firstly, those variables that have the highest impact on the formed clusters are identified. In addition, the degree of effect of different categories of each variable on the consumption patterns is found out and compared.

The same methodology can be applied on other datasets with different characteristics to understand the underlying factors of residential consumption behavior.

6 Tariff Design

6.1 Background and Motivation

In smart electricity grids, electricity retailers can offer various tariffs as well as demand response programs to the customers. For instance, in the comprehensive Australian project called “Smart Grid Smart City”, the effectiveness of different options including different combinations of pricing plans (such as dynamic peak price, seasonal TOU, and pre-payment plan), usage feedback options (such as in-home displays and text messaging), and financial incentives (rebates) were tested on participating customer groups [72]. These structures included 8 network trial plans and 12 retail trial plans. Network plans did not change the electricity tariffs and only measured the effectiveness of smart meter-based products such as feedback technologies and DR programs using interruptible loads. On the other hand, retail plans evaluated the effectiveness of different tariff proposals with/without the feedback options. Therefore, the diverse range of new technologies allows retailers to devise customised plans for the customers. One such methodology is to segment the customers into the groups based on the consumption data provided by smart meters.

In this chapter, we propose an approach for designing customized tariff structures using the results of the clustering process. Compared with the limited number of studies which used the clustering results for tariff design (Section 2.6.3), the current approach offers several contributions and advantages:

- The problem is formulated from the perspective of an electricity retailer that intends to maximize its profits.

- The uncertainties of loads and electricity prices are modeled through a stochastic programming model by considering a set of scenarios.
- Loads of customers are divided into the inflexible and flexible parts in which the flexible part can react to retail prices through an elasticity function.
- The nonlinearities in problem formulation are linearized through proper approaches and the problem is modeled as a mixed-integer linear programming (MILP) which can be easily and accurately solved through available software. Usually, finding global solutions for the mixed-integer non-linear programs (MINLP) is difficult, especially, when the problem includes different nonlinearities and a large number of binary variables. This is a shortcoming of some previous works [213] [80].
- In some studies [80], the obtained TOU prices for the periods of cluster peaks show a sudden increase; around ten times higher than the off-peak prices. Such a TOU structure is unrealistic and very improbable to be implemented by the retailers. In this study, the retail prices of different time periods change in a reasonable fashion.
- The risk measure is included in the problem formulation which enables the retailer to hedge against the possible risks.

The rest of this chapter is organized as follows. Section 6.2 describes the overall procedure including the clustering and stochastic programming model. The explanations of different parts of the objective function and their mathematical formulations are given in Section 6.3. Section 6.4 presents the final objective function of the retailer. The two linearization approaches are also explained in this section. The numerical results of the application of the method on the CER dataset are reported in Section 6.5. Finally, Section 6.6 summarizes the findings of this chapter.

6.2 Procedure

6.2.1 Problem statement and stages

The stages of the proposed method for constructing customized TOU structures are shown in Fig. 6.1. These stages are explained in the following sections.

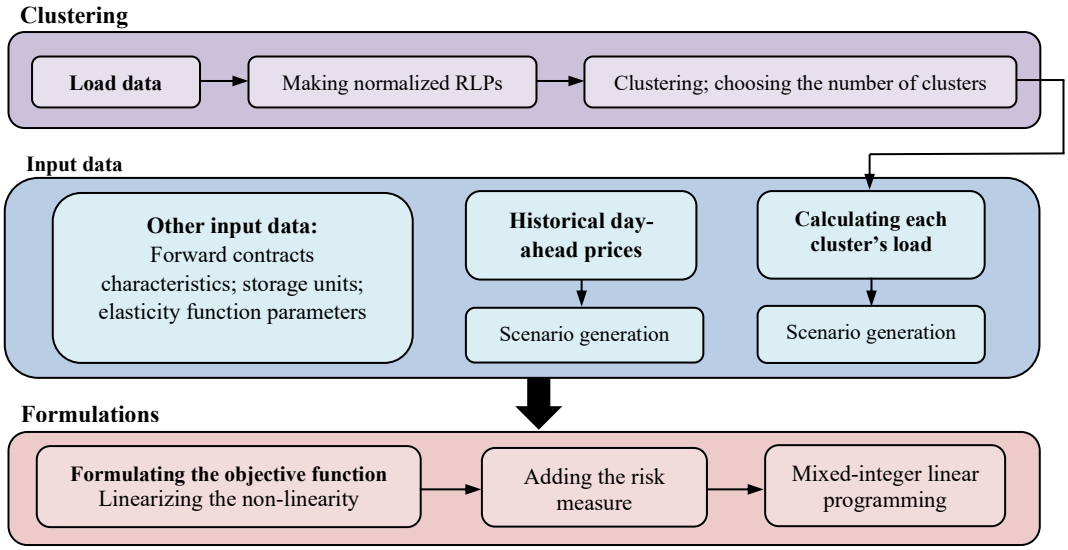


Fig. 6.1 Stages of designing customized TOU structures

6.2.2 Clustering

The clustering of customers is achieved by segmenting them based on their normalized RLPs. Normalization is performed since the aim is to characterize the customers based on their load shapes. While the best number of clusters can be calculated using CVIs, here, we choose the pre-defined number of clusters as six. This is an adequate selection since it can represent the main customers' load shapes and at the same time it keeps the number of groups limited.

Once the customers are assigned to the clusters, the 24-hour load curve of the cluster is determined by averaging non-normalized RLPs of its customers.

6.2.3 Stochastic programming

A stochastic process refers to the random variables, for example, the electricity price or demand, that their values evolve over time [214]. In electricity networks, the stochastic variables such as market prices, customer's demand, and the power generation of renewable resources are the sources of uncertainties. Modeling of the stochastic processes can be done by defining a set of scenarios, in which each scenario represents a single realization of the process. For example, the load of customers for tomorrow, L, is a stochastic variable, and each of its realizations can be represented by $L(\omega)$ where ω is the scenario index. The probability of occurrence of each scenario is specified by $\pi(\omega)$ and $\sum_{\omega=1}^{N_\Omega} \pi(\omega) = 1$ where N_Ω is the number of scenarios.

Since the uncertain data are modeled as stochastic processes, the profit of the retailer will be also a random variable. Therefore, in order to optimize the random objective function, the expected value of the profit is maximized.

A sufficient number of scenarios need to be defined in order to model the stochastic process accurately. However, in practice, defining a very large set of scenarios will result in computational intractability. Usually, scenario reduction methods are utilized to decrease the number of scenarios while retaining most of the information. The technical details of scenario generation and scenario reduction are beyond the scope of this research. Interested readers can refer to [214] for the concepts and formulations.

In the formulations of this chapter, the load of each cluster and the electricity price are modeled through stochastic processes.

6.2.4 Risk measure

As mentioned in the previous section, various sources of uncertainty (for example, market prices or load level) are present in the in the electricity market environment which affect the participants' decision makings. Stochastic programming approach models the uncertain data in which the profits/costs of participants are random variables with a probability distribution. In this way, a common approach to control those variables is to control their expected values.

As mentioned in the previous section, in the optimization processes, the expected value of the objective function is optimized. However, obtaining an acceptable level of expected profit through the optimization process does not mean that, in practice, the retailer will not experience very low profits or even losses. This characterizes a serious shortcoming of such optimization approaches.

Therefore, the conflicting objectives of maximising the value of the random variable and the its desired range should be addressed carefully. To address this issue, a risk measure can be added to the problem formulation to hedge against the risks. Such a measure limits the variation of the random variable and bounds its value to the desired range.

Some of the common risk measures used in the literature include Value-at-Risk (VaR), Tail Value-at-Risk, Expected Shortfall, and Stochastic Dominance [215]. Conditional value-at-risk (CVaR) [214] is a popular risk measure that has been used in different studies [216], [217] and is applied in this research too. Therefore, the final problem is formulated as a risk-averse optimization.

6.3 Formulation

We assume that retailer can procure its needed energy from two market settlements: the forward contracts and the pool (day-ahead market). In forward contracts, constant power is purchased for a specified time period at a fixed price. Since the decisions about the forward contracts are made before the realization of the stochastic process, they are not modeled through scenarios.

6.3.1 Forward contracts costs

The negotiating power of the retailer and the structure of forward contracts is usually modeled through forward contracting curves [218] as shown in Fig. 6.2. As this figure shows, the first block of energy (less than $\bar{P}_{f1}^{F}$) can be procured at the price $\bar{\lambda}_{f1}^{F}$. For the excessive power more than $\bar{P}_{f1}^{F}$ and less than $\bar{P}_{f2}^{F}$, the retailer needs to buy the electricity at the higher price $\bar{\lambda}_{f2}^{F}$. Therefore, the prices increase in a stepwise manner based on the traded quantity.

Let F be the set of forward contracts, N_J be the number of power blocks in the forward contracting curve, and P_{fj}^{F} and λ_{fj}^{F} be the power purchased from the j-block of the forward contract curve of contract f and its corresponding price, respectively. The total cost of buying energy through forward contracts can be calculated as:

$$C_t^F = \sum_{f \in F} \sum_{j=1}^{N_J} \lambda_{fj}^{F} \, P_{fj}^{F} \, d_t \tag{6.1}$$

Eq. (6.1) shows the cost of energy through forward contracts in each time period d_t. d_t is considered as one hour in this study.

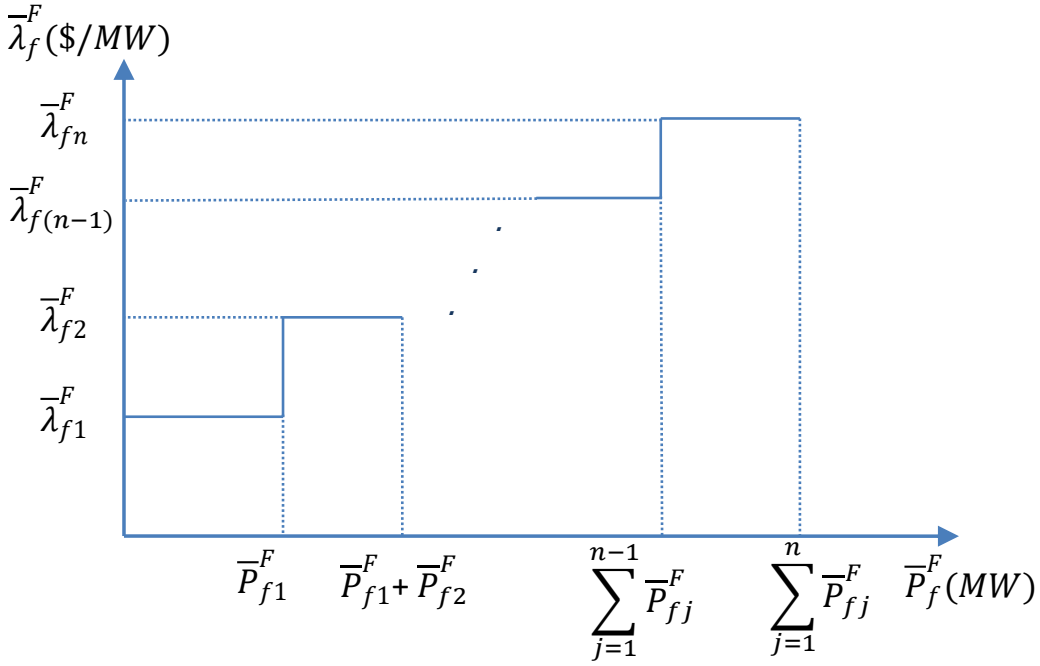


Fig. 6.2 Forward contracts curve

6.3.2 Expected day-ahead market costs/revenues

Electricity market settlements usually consist of several trading markets. For example, three trading floors including day-ahead, intra-day, and real time markets are used in several studies. Other works have considered only one or two of these market in their problem formulation.

Since the actual amount of consumption in each hour is not known, the retailer needs to participate in the pool to purchase the needed energy for its customers. In this research, only the retailer participation in the day-ahead market is considered and other market settlements like real-time markets are not taken into account; however, they can be easily added to the problem formulation.

The operator of the day-ahead market requires all producers and buyers to submit their offers for the corresponding sales and purchases for day $i+1$, on day i . Using these data, the market is cleared and the market clearing price (day-ahead market price) is calculated.

The pool price is a stochastic variable and its values in each hour are modeled through the scenarios. Let $P_{\omega t}^{DA}$ be the amount of energy purchased from/sold to the pool and $\lambda_{\omega t}^{DA}$ be the price of electricity in the day-ahead market in every time period t and each scenario ω. The expected pool cost/revenue can be calculated as:

$$C^{DA} = \sum_{\omega=1}^{N_\Omega} \sum_{t=1}^{24} P_{\omega t}^{DA} \, \lambda_{\omega t}^{DA} \, \pi_\omega \tag{6.2}$$

where Ω and N_Ω represent the scenario set and the number of scenarios, respectively. $\pi(\omega)$ represents the probability of occurrence of each scenario.

6.3.3 Revenues

The retailer earns money by selling electricity to the customers. Usually, the retail prices are higher than the day-ahead prices. As mentioned earlier, the retailer's aim is to determine the optimal TOU prices/structures for each cluster.

Generally, the amount of demand change based on the retail price level. It means that customers consume less energy if they are exposed to higher prices. This relationship between the demand and price is modeled through various mathematical models in different studies, for example, through elasticity function [80], [213], price-demand quota curve [218], [219], the bi-level problem considering the utility function of the customers [220], and the bi-level problem minimizing the violation from comfort for the users [221]. Each of these methods has its own advantages and disadvantages.

In this chapter, a linear elasticity function is considered to model the reaction of customers to retail prices. The load of the customers consists of inflexible (critical) loads that cannot be curtailed or shifted and flexible loads [124]. Flexible loads can be adjusted in response to variations in the electricity price or to incentive payments. Here, we assume that only the flexible part of the loads will change according to price signals.

Let $L_{\omega tk}$ be the load of cluster k and $\lambda^r_{\omega tk}$ the retail price for this cluster in time period t and scenario ω. $L_{\omega tk}$ can be written as:

$$L_{\omega tk} = L^{inf}_{\omega tk} + L^{flex}_{\omega tk} \tag{6.3}$$

where $L^{inf}_{\omega tk}$ and $L^{flex}_{\omega tk}$ refer to the inflexible and flexible parts of the load, respectively.

The elasticity function can be written as [222]:

$$El(\lambda^r_{\omega tk}) = \left[r_1 - \beta\left(\frac{\lambda^r_{\omega tk} - \lambda^r_{0t}}{\lambda^r_{0t}}\right)\right] = r_1[1 - r_2(\lambda^r_{\omega tk} - \lambda^r_{0t})] \tag{6.4}$$

where r_1 and β are the coefficients defining the elasticity function. λ^r_{0t} is the nominal retail price in each time period t. In the case studies, it is considered that the parameters of the elasticity function are the same for all clusters. It is also possible to define cluster-specific elasticity functions. This issue is described briefly in Section 6.4.2.

The expected revenue from the customer-side is given as:

$$R = \sum_{\omega=1}^{N_\Omega} \pi_\omega \sum_{t=1}^{24} \sum_{k=1}^{K} R_{\omega tk} \tag{6.5}$$

where

$$R_{\omega tk} = [L_{\omega tk}^{inf} + L_{\omega tk}^{flex}\, El(\lambda_{\omega tk}^{r})]\lambda_{\omega tk}^{r}$$
$$= (L_{\omega tk}^{inf} + r_1\, L_{\omega tk}^{flex} + r_1\, r_2\, L_{\omega tk}^{flex}\, \lambda_{0t}^{r})\lambda_{\omega tk}^{r} - r_1\, r_2\, L_{\omega tk}^{flex}\, (\lambda_{\omega tk}^{r})^2 \quad (6.6)$$
$$= m_{1,\omega tk}\, \lambda_{\omega tk}^{r} - m_{2,\omega tk}\, (\lambda_{\omega tk}^{r})^2$$

In the above formulations, *K* refers to the total number of clusters.

In addition, we assume that the minimum and maximum bounds are asserted on the retail price. This is a practical assumption since such bounds are either defined by the system operator or imposed as a part of the agreement between the retailer and its own customers.

6.3.4 TOU structure

In this study, it is assumed that the three different prices, high (λ_k^{high}), medium (λ_k^{medium}), and low (λ_k^{low}), are offered to each cluster for different time periods of the day based on the load pattern of the cluster. Therefore, both the price and the price durations are different between different clusters and are determined through the optimization process. The formulation can be easily modified to include more or less prices instead of the 3-step price scheme.

As mentioned, the retail price can take only three values during the day. Mathematically speaking, the retail price can be written as:

$$\lambda_{\omega tk}^{r} = \upsilon_{\omega tk}^{high}\, \lambda_k^{high} + \upsilon_{\omega tk}^{low}\, \lambda_k^{low} + \upsilon_{\omega tk}^{medium}\, \lambda_k^{medium}\,, \forall\, \omega, \forall\, t, \forall\, k \quad (6.7)$$

$$\upsilon_{\omega tk}^{high} + \upsilon_{\omega tk}^{low} + \upsilon_{\omega tk}^{medium} = 1\,, \qquad \forall\, \omega, \forall\, t, \forall\, k \quad (6.8)$$

$$\underline{n} \leq \sum_{t=1}^{24} v_{\omega tk}^{high} \leq \overline{n} \ , \forall\, \omega, \forall\, k \tag{6.9}$$

$$\underline{n} \leq \sum_{t=1}^{24} v_{\omega tk}^{low} \leq \overline{n} \ , \forall\, \omega, \forall\, k \tag{6.10}$$

$$\underline{n} \leq \sum_{t=1}^{24} v_{\omega tk}^{medium} \leq \overline{n} \ , \forall\, \omega, \forall\, k \tag{6.11}$$

In the above formulation, $v_{\omega tk}^{high}$, $v_{\omega tk}^{low}$, and $v_{\omega tk}^{medium}$ are the binary values where only one of them can be 1 at each time period. $\underline{n}$ and $\overline{n}$ are the minimum and the maximum number of hours that each of the high, low, and medium prices can occur. The values of $\underline{n}$ and $\overline{n}$ can be changed based on the preferences of the retailer. For example, both $\underline{n}$ and $\overline{n}$ can be set to 8 if equal time periods for all three prices are intended.

6.3.5 Energy storage units

Retailers can also possess their own facilities to produce electricity [219]. For instance, the retailer can use distributed generation units, renewable energy resources, aggregated electric vehicles, and energy storage units in order to reduce its energy procurement costs. In this chapter, we suppose that the retailer owns several energy storage units that can be discharged during the periods of high electricity prices and charged during the low-price periods. It is assumed that the operation costs of the storage units are negligible. Other complicated models for example, by formulating the aggregation of electric vehicles can also be considered in the problem.

6.3.6 CVaR formulation

The CVaR measure is incorporated into the risk-neutral problem. CVaR is equal to the solution of the following optimization problem:

$$Maximise_{\ \zeta,\eta_\omega} \quad \zeta - \frac{1}{1-\alpha}\sum_{\omega=1}^{N_\Omega} \pi_\omega\, \eta_\omega \tag{6.12}$$

Subject to:

$$\zeta - \sum_{t=1}^{24}\left(\sum_{k=1}^{K} R_{\omega tk} - \lambda_{\omega t}^{DA}\, P_{\omega t}^{DA} - \sum_{f\in F}\sum_{j=1}^{N_J} \lambda_{fj}^{F}\, P_{fj}^{F}\, d_t\right) \le \eta_\omega\ , \forall\ \omega \tag{6.13}$$

$$0 \le \eta_\omega \tag{6.14}$$

In the above formulation, $\alpha \in (0,1)$ is a predetermined value called confidence level. ζ and η_ω are the continuous and scenario-specific auxiliary variables, respectively.

6.4 Objective Function

Finally, the retailer problem can be formulated as the following objective function:

$$\underset{P_{fj}^{F},\ \lambda_{\omega t}^{DA}, v_{\omega tk}^{high}, \lambda_{k}^{high}, v_{\omega tk}^{low}, \lambda_{k}^{low}, v_{\omega tk}^{medium}, \lambda_{k}^{medium}, P_{\omega t}^{DA}, \zeta, \eta_\omega, P_{\omega t}^{ch}, U_{\omega t}^{ch}, U_{\omega t}^{dis}}{Maximize}$$

$$Profit = \sum_{\omega=1}^{N_\Omega} \pi_\omega \sum_{t=1}^{24}\left(\sum_{k=1}^{K} R_{\omega tk} - \lambda_{\omega t}^{DA}\, P_{\omega t}^{DA} - \sum_{f\in F}\sum_{j=1}^{N_J} \lambda_{fj}^{F}\, P_{fj}^{F}\, d_t\right)$$

$$+ \beta\left(\zeta - \frac{1}{1-\alpha}\sum_{\omega=1}^{N_\Omega} \pi_\omega\, \eta_\omega\right) \tag{6.15}$$

Subject to:

forward contracts constraints:

$$0 \le P_{fj}^{F} \le \overline{P}_{fj}^{F}\,, \quad \forall f, \forall j \tag{6.16}$$

retail price constraints:

$$\underline{\lambda}^{r} \le \lambda_{\omega tk}^{r} \le \overline{\lambda}^{r}\,, \quad \forall\, \omega, \forall\, t, \forall\, k \tag{6.17}$$

$$\lambda_{\omega tk}^{r} = \upsilon_{\omega tk}^{high}\, \lambda_{k}^{high} + \upsilon_{\omega tk}^{low}\, \lambda_{k}^{low} + \upsilon_{\omega tk}^{medium}\, \lambda_{k}^{medium}\,, \forall\, \omega, \forall\, t, \forall\, k \tag{6.18}$$

$$\upsilon_{\omega tk}^{high} + \upsilon_{\omega tk}^{low} + \upsilon_{\omega tk}^{medium} = 1\,, \qquad \forall\, \omega, \forall\, t, \forall\, k \tag{6.19}$$

$$\underline{n} \le \sum_{t=1}^{24} \nu_{\omega tk}^{high} \le \overline{n}\,, \forall\, \omega, \forall\, k \tag{6.20}$$

$$\underline{n} \le \sum_{t=1}^{24} \nu_{\omega tk}^{low} \le \overline{n}\,, \forall\, \omega, \forall\, k \tag{6.21}$$

$$\underline{n} \le \sum_{t=1}^{24} \nu_{\omega tk}^{medium} \le \overline{n}\,, \forall\, \omega, \forall\, k \tag{6.22}$$

battery constraints:

$$ES_{\omega t} = ES_{\omega(t-1)} + \eta^{ch} P_{\omega t}^{ch} - \frac{P_{\omega t}^{dis}}{\eta^{dis}}\,, \quad \forall\, \omega, \forall\, t \tag{6.23}$$

$$0 \le P_{\omega t}^{ch} \le P_{max}^{ch}\, U_{\omega t}^{ch}\,, \quad \forall\, \omega, \forall\, t \tag{6.24}$$

$$0 \le P_{\omega t}^{dis} \le P_{max}^{dis}\, U_{\omega t}^{dis}\,, \quad \forall\, \omega, \forall\, t \tag{6.25}$$

$$U_{\omega t}^{ch} + U_{\omega t}^{dis} \le 1\,, \quad \forall\, \omega, \forall\, t \tag{6.26}$$

$$ES_{min} \leq ES_{\omega t} \leq ES_{max}\ ,\ \ \forall\ \omega, \forall\ t \tag{6.27}$$

$$ES_{t=1} = ES_b \tag{6.28}$$

power balance constraint:

$$\sum_{k=1}^{K} L_{\omega tk} + P_{\omega t}^{ch} = P_{\omega t}^{DA} + P_{\omega t}^{dis} + \sum_{f \in F} \sum_{j=1}^{N_J} P_{fj}^{F}\ d_t\ , \qquad \forall\ \omega, \forall\ t \tag{6.29}$$

CVAR risk constraint:

$$\zeta - \sum_{t=1}^{24} \left(\sum_{k=1}^{K} R_{\omega tk} - \lambda_{\omega t}^{DA}\ P_{\omega t}^{DA} - \sum_{f \in F} \sum_{j=1}^{N_J} \lambda_{fj}^{F}\ P_{fj}^{F}\ d_t \right) \leq \eta_\omega\ , \forall\ \omega \tag{6.30}$$

variables conditions:

$$v_{\omega tk}^{high}, v_{\omega tk}^{low}, v_{\omega tk}^{medium}\ \in\ \{0,1\}\ , \qquad \forall\ \omega, \forall\ t, \forall\ k \tag{6.31}$$

$$U_{\omega t}^{ch}\ , U_{\omega t}^{dis}\ \in \{0,1\}\ ,\ \ \forall\ \omega, \forall\ t \tag{6.32}$$

$$0 \leq P_{\omega t}^{DA}\ , \forall\ \omega, \forall\ t \tag{6.33}$$

$$0 \leq \eta_\omega\ , \forall\ \omega \tag{6.34}$$

The first part of the objective function (6.15) shows the expected profit of the retailer and the second part characterizes the CVaR risk measure. Parameter $\beta \in [0, \infty)$ is a weighting factor that determines the tradeoff between the expected profit and risk. If $\beta = 0$, the formulation will reduce to a risk-neutral problem. On the other hand, by increasing the value of β, the problem will be more risk-averse. Therefore, in practice, the retailer can

select the value of β based on its preferences. For example, a retailer who is seeking higher profits while accepting the higher risks will set the value of β close to zero.

Constraints (6.23) to (6.28) model the charging and discharging of the storage unit. $P_{\omega t}^{ch}$ and $P_{\omega t}^{dis}$ indicate the charge and discharge power, respectively. $U_{\omega t}^{ch}/U_{\omega t}^{dis}$ are the binary variables that determine the charge/discharge status in each time frame and η^{ch}/η^{dis} are the charge/discharge efficiency. Also, the state of charge of storage unit is shown by *ES*.

6.4.1 Linearization

There are two nonlinearities in the objective function. The first one stems from the product of the binary variables $\upsilon_{\omega tk}^{high}$, $\nu_{\omega tk}^{low}$, and $\nu_{\omega tk}^{medium}$ with the corresponding continuous variables λ_k^{high}, λ_k^{low}, and λ_k^{medium}. The second nonlinearity is introduced by the term $(\lambda_{\omega tk}^{r})^2$ in the revenue function.

In the following, proper solutions (by introducing the new variables and constraints) are proposed to overcome these nonlinearities and make the problem linear. These sets of equations represent exactly the original formulations. However, the linearization is achieved by adding a significant number of variables and constraints.

i) Linearization of first nonlinearity

The first nonlinearity is linearized by introducing three auxiliary variables $\phi_{\omega tk}^{high} = \upsilon_{\omega tk}^{high}\ \lambda_k^{high}$, $\phi_{\omega tk}^{low} = \upsilon_{\omega tk}^{low}\ \lambda_k^{low}$, and $\phi_{\omega tk}^{medium} = \upsilon_{\omega tk}^{medium}\ \lambda_k^{medium}$.

Therefore, the retail price can be written as:

$$\lambda_{\omega tk}^{r} = \phi_{\omega tk}^{peak} + \phi_{\omega tk}^{low} + \phi_{\omega tk}^{medium}\ , \forall\ \omega, \forall\ t, \forall\ k \tag{6.35}$$

And the following constraints are added to the problem:

$$\phi_{\omega tk}^{high} \leq \overline{\lambda}^{r} v_{\omega tk}^{high} \tag{6.36}$$

$$\phi_{\omega tk}^{high} \geq \underline{\lambda}^{r} v_{\omega tk}^{high} \tag{6.37}$$

$$\phi_{\omega tk}^{high} \leq \lambda_{k}^{high} - \underline{\lambda}^{r}(1 - v_{\omega tk}^{high}) \tag{6.38}$$

$$\phi_{\omega tk}^{high} \geq \lambda_{k}^{high} - \overline{\lambda}^{r}(1 - v_{\omega tk}^{high}) \tag{6.39}$$

$$\phi_{\omega tk}^{low} \leq \overline{\lambda}^{r} v_{\omega tk}^{low} \tag{6.40}$$

$$\phi_{\omega tk}^{low} \geq \underline{\lambda}^{r} v_{\omega tk}^{low} \tag{6.41}$$

$$\phi_{\omega tk}^{low} \leq \lambda_{k}^{low} - \underline{\lambda}^{r}(1 - v_{\omega tk}^{low}) \tag{6.42}$$

$$\phi_{\omega tk}^{low} \geq \lambda_{k}^{low} - \overline{\lambda}^{r}(1 - v_{\omega tk}^{low}) \tag{6.43}$$

$$\phi_{\omega tk}^{medium} \leq \overline{\lambda}^{r} v_{\omega tk}^{medium} \tag{6.44}$$

$$\phi_{\omega tk}^{medium} \geq \underline{\lambda}^{r} v_{\omega tk}^{medium} \tag{6.45}$$

$$\phi_{\omega tk}^{medium} \leq \lambda_{k}^{medium} - \underline{\lambda}^{r}(1 - v_{\omega tk}^{medium}) \tag{6.46}$$

$$\phi_{\omega tk}^{medium} \geq \lambda_{k}^{medium} - \overline{\lambda}^{r}(1 - v_{\omega tk}^{medium}) \tag{6.47}$$

ii) Linearization of second nonlinearity

The linearization of the revenue function is more complicated. Revenue function has a quadratic form which can be accurately approximated by a set of piecewise blocks. The

linearization of the quadratic production cost of the generators is done in different unit commitment studies [223], [224], [225]. Here, based on the work in [223], a piecewise approximation of revenue function is formulated.

Recalling the revenue function as:

$$R_{\omega tk} = m_{1,\omega tk}\, \lambda^{r}_{\omega tk} - m_{2,\omega tk}\, (\lambda^{r}_{\omega tk})^{2} \tag{6.48}$$

As shown in Fig. 6.3, the quadratic function is approximated by several linear segments.

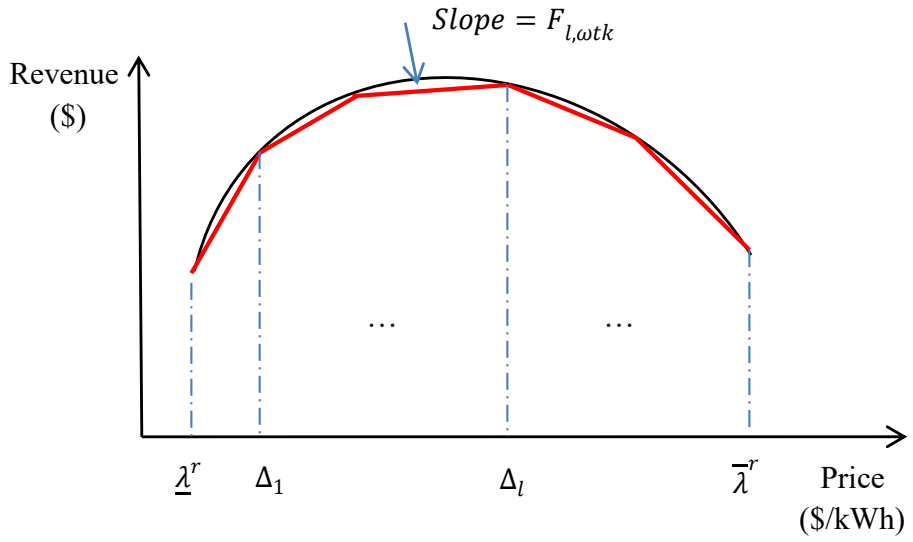


Fig. 6.3 Revenue function and approximated piecewise revenue function

The original function and its piecewise approximate will be indistinguishable if the number of these segments is considered large enough. Each segment l is defined by the lower and upper prices, $\Delta_{(l-1),k}$ and Δ_{lk}. $F_{l,\omega tk}$ is the slope of segment l and $\delta_{l,\omega tk}$ represents the incremental retail price which is limited between the zero and the $\Delta_{lk} - \Delta_{(l-1),k}$. Therefore, the retail price and the linearized revenue function can be expressed as:

$$\lambda^{r}_{\omega tk} = \underline{\lambda}^{r} + \sum_{l=1}^{NL_k} \delta_{l,\omega tk}\,, \qquad \forall\, t, \forall\, k \tag{6.49}$$

$$R_{\omega tk}^{linear} = A_{\omega tk} + \sum_{l=1}^{NL_k} F_{l,\omega tk} . \delta_{l,\omega tk} \tag{6.50}$$

where:

$$A_{\omega tk} = R_{\omega tk}\left(\underline{\lambda}^r\right) \tag{6.51}$$

$$\Delta\Delta_k = \left[(\overline{\lambda}^r - \underline{\lambda}^r)/NL_k\right], \qquad \forall\, k \tag{6.52}$$

$$\Delta_{lk} = \underline{\lambda}^r + l.\Delta\Delta_k \ , \ \ \forall\, k, \forall\, l = 1, \dots, NL_k - 1 \tag{6.53}$$

$$0 \le \delta_{1,\omega tk} \le \Delta_{1,k} - \underline{\lambda}^r \ , \qquad \forall\, t, \forall\, k \tag{6.54}$$

$$0 \le \delta_{l,\omega tk} \le \Delta_{lk} - \Delta_{(l-1),k} \ , \qquad \forall\, t, \forall\, k, \forall\, l = 2, \dots, NL_k - 1 \tag{6.55}$$

$$0 \le \delta_{NL_k,\omega tk} \le \overline{\lambda}^r - \Delta_{(NL_k-1),k} \ , \ \ \forall\, t, \forall\, k \tag{6.56}$$

$$F_{1,\omega tk} = \left[R_{\omega tk}(\Delta_{1k}) - R_{\omega tk}\left(\underline{\lambda}^r\right)\right]/\Delta\Delta_k \ \ , \ \ \forall\, t, \forall\, k \tag{6.57}$$

$$F_{l,\omega tk} = \left[R_{\omega tk}(\Delta_{lk}) - R_{\omega tk}\left(\Delta_{(l-1),k}\right)\right]/\Delta\Delta_k, \qquad \forall\, t, \forall\, k, \forall\, l = 2, \dots, NL_k - 1 \tag{6.58}$$

$$F_{NL,\omega tk} = \left[R_{\omega tk}\left(\overline{\lambda}^r\right) - R_{\omega tk}\left(\Delta_{(NL_k-1),k}\right)\right]/\Delta\Delta_k \ , \qquad \forall\, t, \forall\, k \tag{6.59}$$

NL represents the number of segments and its value can be different for each cluster (NL_k).

It should be noticed that in the above formulations, the variable set only consists of $\delta_{l,\omega tk}$. Other parameters are determined based on the available data. For example, the slopes ($F_{l,\omega tk}$) can be calculated using (6.48) for the intersection points. Therefore, (6.50) represents a linearized form of the revenue function.

6.4.2 Other considerations

i) TOU structure

Usually, TOU price structures are designed in a way that each of the high, low, and medium prices happen at consecutive hours. This is an assumption to assure that the customers can follow the TOU prices. On the other hand, the availability of newer technologies such as smart meters and home energy management systems, which automatically manage the home consumption, allows defining more complicated pricing schemes. In this regard, recent publications on electricity pricing or demand response consider schemes such as the real-time pricing in which the price of electricity varies at each hour and for each day.

In this chapter, we consider that only three prices are offered to each cluster and these prices and the price structure are the same for at least several months. In order to avoid the nonlinearity in the problem formulation, the condition of consecutive hours for each retail price is relaxed. However, the aforementioned condition can be fulfilled by adding the following non-linear constraints to the problem [80]:

$$\left|\upsilon_{\omega,1,k}^{high} - \upsilon_{\omega,24,k}^{high}\right| + \sum_{t=2}^{24} \left|\upsilon_{\omega tk}^{high} - \upsilon_{\omega(t-1)k}^{high}\right| = 2 \qquad (6.60)$$

$$\left|\upsilon_{\omega,1,k}^{low} - \upsilon_{\omega,24,k}^{low}\right| + \sum_{t=2}^{24}\left|\upsilon_{\omega tk}^{low} - \upsilon_{\omega(t-1)k}^{low}\right| = 2 \quad (6.61)$$

$$\left|\upsilon_{\omega,1,k}^{medium} - \upsilon_{\omega,24,k}^{medium}\right| + \sum_{t=2}^{24}\left|\upsilon_{\omega tk}^{medium} - \upsilon_{\omega(t-1)k}^{medium}\right| = 2 \quad (6.62)$$

ii) Cluster-specific elasticity function

In this research, we consider the same elasticity function for all clusters. Based on the available data, it might be possible to define the cluster-specific elasticity functions. In the following, a possible approach is explained.

The CER dataset does not contain any specific questions or information regarding the reaction of customers to retail prices. The questions that ask about the attitudes of customers toward energy savings or energy reductions can be used to identify the potential of each customer for the change in its energy use. These set of questions are shown in Table 6-1.

Table 6-1 Attitudes toward energy saving and knowledge about energy reduction: the responses are on a scale of 1 to 5 where 1 is "strongly agree" and 5 is "strongly disagree".

Questions	
Capability	
1	Possibility to make major changes in electricity use?
2	We can reduce our electricity bill by changing the way the people we live with use electricity.
3	It is too inconvenient to reduce our usage of electricity.
4	I am not be able to get the people I live with to reduce their electricity usage.
5	I do not have enough time to reduce my electricity usage.
Willingness	
6	Interested in changing electricity use if it reduces the bill.
7	Interested in changing electricity use if it helps the environment.
8	We would like to do more to reduce electricity usage.
9	I do not want to be told how much electricity I can use.

In this table, questions are divided into two groups, one set of questions that specify the capability for changes and the other set of questions which designate the willingness of customers. A set of rules can be defined to indicate either a customer belongs to one of these sets from the elasticity perspective and sensitivity to energy prices: “very sensitive”, “sensitive”, “slightly sensitive”, and “not sensitive”. Finally, the values of elasticity function for each cluster can be defined based on the results of all customers who belong to that cluster.

6.5 Numerical Results

The problem is formulated as a MILP and coded in the General Algebraic Modeling System (GAMS) environment [226]. The server facilities and optimization solvers of Network-Enabled Optimization System (NEOS) [227], [228] are used for solving the problem.

6.5.1 Data

The values of different parameters which are used in the case studies are listed in Table 6-2.

The forward contracts are modeled through three different kinds of contracts including the peak (9:00-22:00), off-peak (1:00-8:00 and 23:00-24:00), and round-the-clock (all day). The characteristics of these contracts are illustrated in Table 6-3.

Table 6-2 The values of different parameters for numerical studies

K: Number of clusters	6
N_F: Number of forward contracts	3
N_J: Number of blocks in each forward contract	3
N_Ω: Number of scenarios	10
NL_k: Number of segments for the approximated revenue function	100
ES_{max}: Maximum energy of energy storage system	1500 kWh
ES_{min}: Minimum energy of energy storage system	200 kWh
ES_b: State of charge at the beginning of the period	300 kWh
P_{max}^{ch}: Maximum rate of charge	200 kW
P_{max}^{dis}: Maximum rate of discharge	200 kW
η^{ch}: Charge efficiency coefficient	0.95
η^{dis}: Discharge efficiency coefficient	0.95
$\underline{\lambda}^r$:Minimum retail price	0.040 \$/kWh
$\overline{\lambda}^r$: Maximum retail price	0.075 \$/kWh
α: Confidence level for CVaR risk measure	0.95
r_1: First constant in elasticity function	1
r_2: Second constant in elasticity function	15.5
λ_{0t}^r: The threshold price (nominal retail price) in elasticity function	0.052 \$/kW

Table 6-3 Characteristics of forward contracts

	Block 1		Block 2		Block 3	
	Price ($/kWh)	Upper limit (kWh)	Price ($/kWh)	Upper limit (kWh)	Price ($/kWh)	Upper limit (kWh)
F1 (peak)	0.051	1500	0.053	1200	0.056	800
F2 (off-peak)	0.044	1300	0.046	850	0.050	750
F3 (round-the-clock)	0.049	1400	0.052	1000	0.055	1000

The historical day-ahead prices of Iberian electricity market [229] are used for estimating the day-ahead prices and generating the price scenarios.

It should be noticed that by increasing the number of scenarios the number of variables and the processing time increases significantly. Therefore, only 10 scenarios are considered for case studies. In practice, a higher number of scenarios can be selected.

Scenario generation and reduction are performed using the method in [217]. The technical descriptions of this process are beyond the scope of this thesis and are not mentioned here.

6.5.2 Characteristics of clusters

Different tariffs can be set for different seasons or loading conditions such as weekdays or weekends. Here, only the weekdays belonging to summer and winter are analyzed. The analysis is similar for other seasons or weekends. Using normalized RLPs of customers, six clusters are formed for each season as shown in Fig. 6.4 and Fig. 6.5. As described in the previous chapters, each cluster characterizes different lifestyle and consumption pattern.

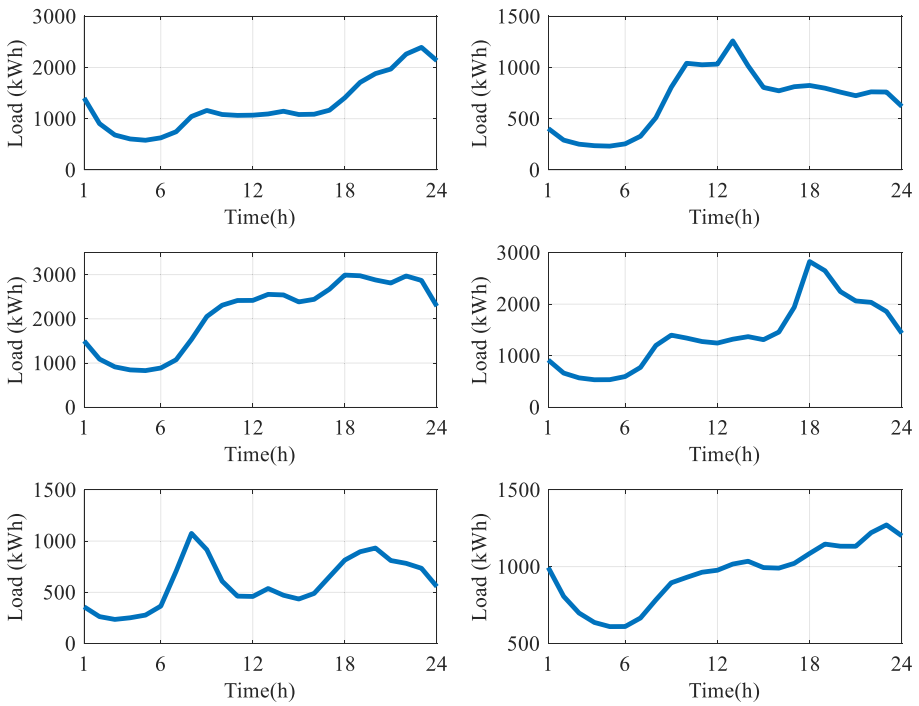


Fig. 6.4 Six formed clusters for summer

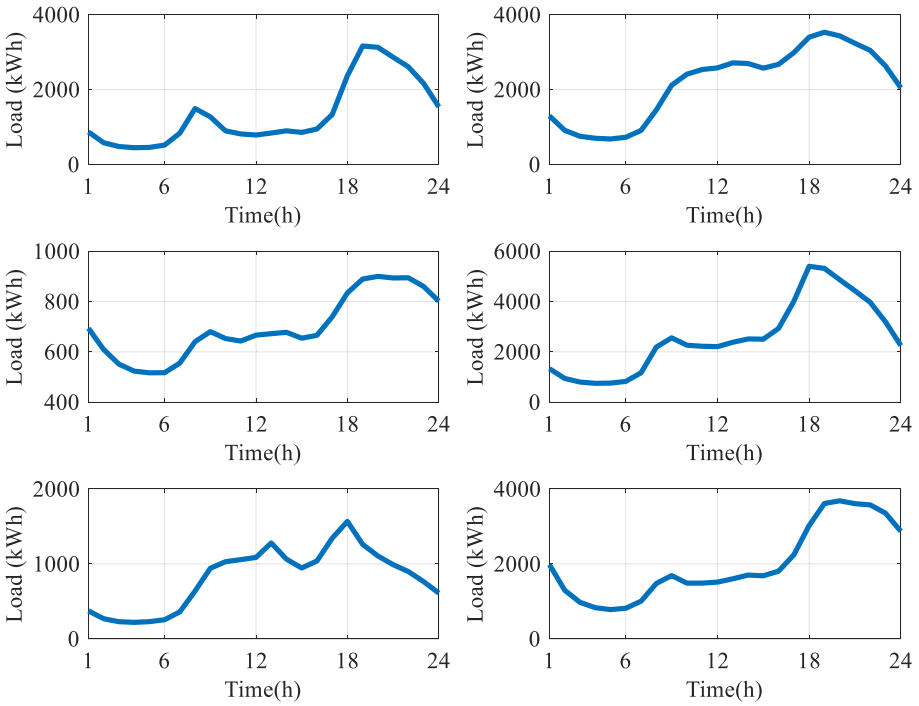


Fig. 6.5 Six formed clusters for winter

6.5.3 Results

I) Benchmark 1: Fixed retail price for all customers

In this case, it is assumed that the retailer offers just a fixed price to all customers. The optimum value of this price is determined through the optimization process. The optimum value of the retail price and the profit of the retailer for the summer are 0.071 $/kWh and $ 805.6 respectively. The corresponding values for the winter are 0.069 $/kWh and $ 1104.5 respectively.

II) Benchmark 2: Calculation of TOU prices without the clustering

In this case, it is assumed that the retailer offers unique TOU price structures to all the customers. The values of these prices are calculated through the optimization process.

Table 6-4 and Table 6-5 report the values of high, medium, and low prices for summer and winter respectively. Furthermore, Fig. 6.6 and Fig. 6.7 show the TOU structures for the summer and winter respectively. In this figures the black curve shows the cluster load and the blue line displays the TOU structure. The attained profit of the retailer for the summer increases from $ 805.6 in the previous case to $ 821.3. The winter profit also observes an increase from $ 1104.5 to $ 1135.

Table 6-4 Summer retail prices for benchmark scheme (without clustering)

$\boldsymbol{\lambda_k^{high}}$ ($\$/kWh$)	$\boldsymbol{\lambda_k^{medium}}$ ($\$/kWh$)	$\boldsymbol{\lambda_k^{low}}$ ($\$/kWh$)
0.074	0.0705	0.0649

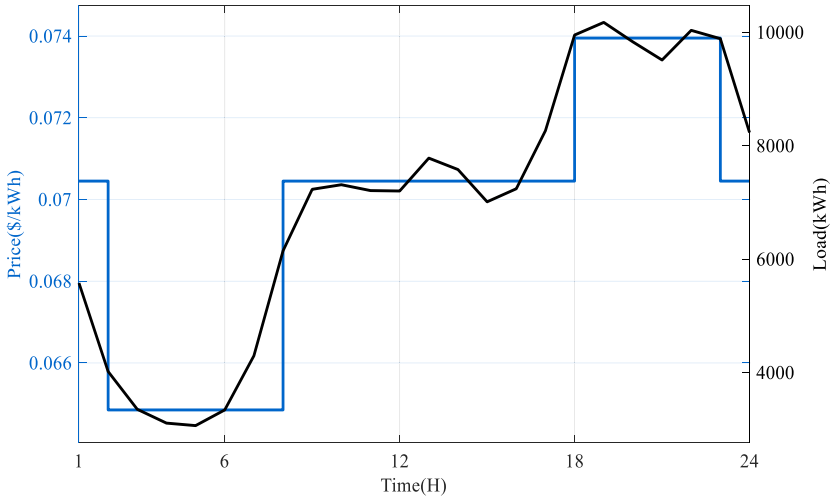


Fig. 6.6 The summer retail structure for the benchmark scheme (without clustering)

Table 6-5 Winter retail prices for benchmark scheme (without clustering)

$\boldsymbol{\lambda_k^{high}}$ ($\$/kWh$)	$\boldsymbol{\lambda_k^{medium}}$ ($\$/kWh$)	$\boldsymbol{\lambda_k^{low}}$ ($\$/kWh$)
0.0726	0.0677	0.0631

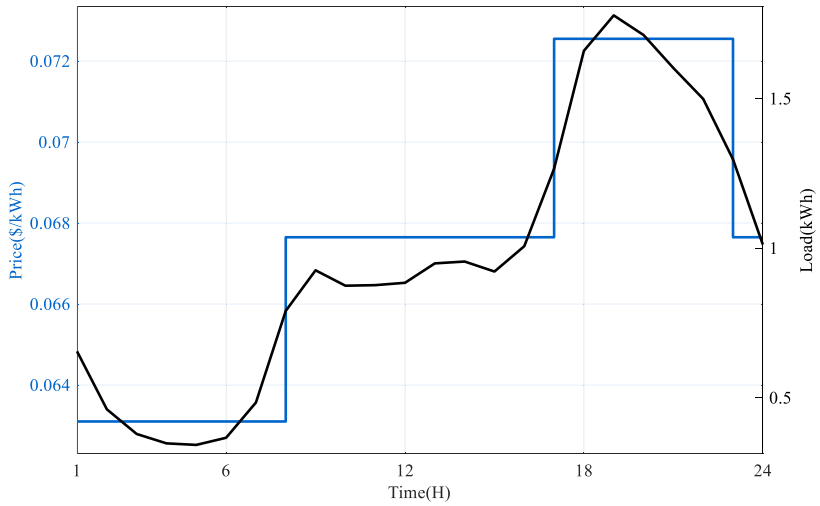


Fig. 6.7 The winter retail structure for the benchmark scheme (without clustering)

III) Calculation of TOU prices considering the clustering

In this case, the customers are segregated into separate clusters and the optimum retail prices are decided for each cluster. The values of these prices for the summer and winter are reported in Table 6-6 and Table 6-7 respectively. The TOU structure for each cluster in these seasons is also shown in Fig. 6.8 and Fig. 6.9. In these figures, the black and blue lines represent the load of the cluster and the calculated TOU structure for that cluster, respectively.

In general, the TOU structure for each cluster reflects the load shape of that cluster. In most cases, 4 or 5 time periods during the day are defined for the TOU tariffs. As mentioned earlier, we can limit this to, for example, three consecutive time periods with the cost of the nonlinearity of the optimization problem. In practice, the retailer is able to merge some of these intervals based on its preferences.

Table 6-6 Obtained high, medium, and low retail prices for each cluster for summer

Clusters	Retail prices		
	$\boldsymbol{\lambda_k^{high}}$ $(\$/kWh)$	$\boldsymbol{\lambda_k^{medium}}$ $(\$/kWh)$	$\boldsymbol{\lambda_k^{low}}$ $(\$/kWh)$
Cluster 1	0.0722	0.0684	0.0652
Cluster 2	0.0708	0.0701	0.0635
Cluster 3	0.0729	0.0726	0.0656
Cluster 4	0.0719	0.0670	0.0645
Cluster 5	0.0712	0.0677	0.0638
Cluster 6	0.0733	0.0715	0.0684

Table 6-7 Obtained high, medium, and low retail prices for each cluster for winter

Clusters	Retail prices		
	λ_k^{high} ($/kWh)	λ_k^{medium} ($/kWh)	λ_k^{low} ($/kWh)
Cluster 1	0.0726	0.0666	0.0631
Cluster 2	0.0726	0.0705	0.0638
Cluster 3	0.0736	0.0712	0.0701
Cluster 4	0.0729	0.0673	0.0617
Cluster 5	0.0715	0.0691	0.0635
Cluster 6	0.0733	0.0677	0.0652

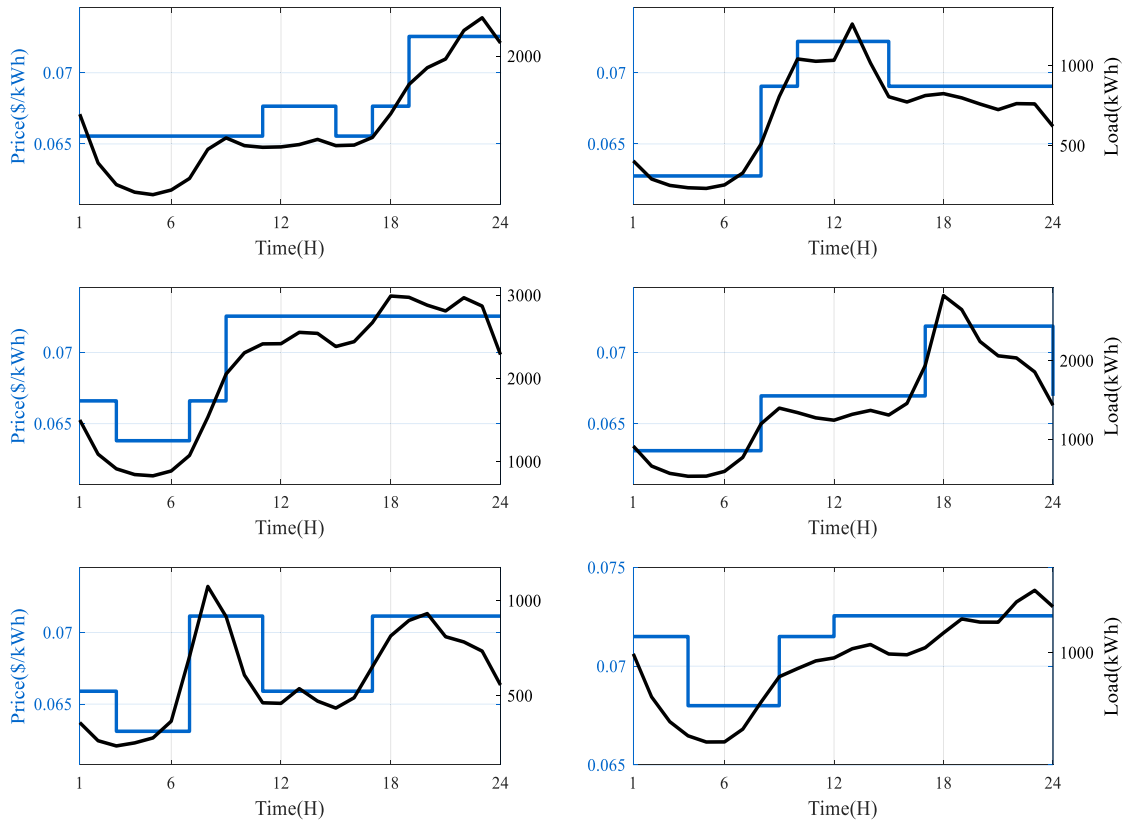


Fig. 6.8 TOU price structure for each cluster for summer

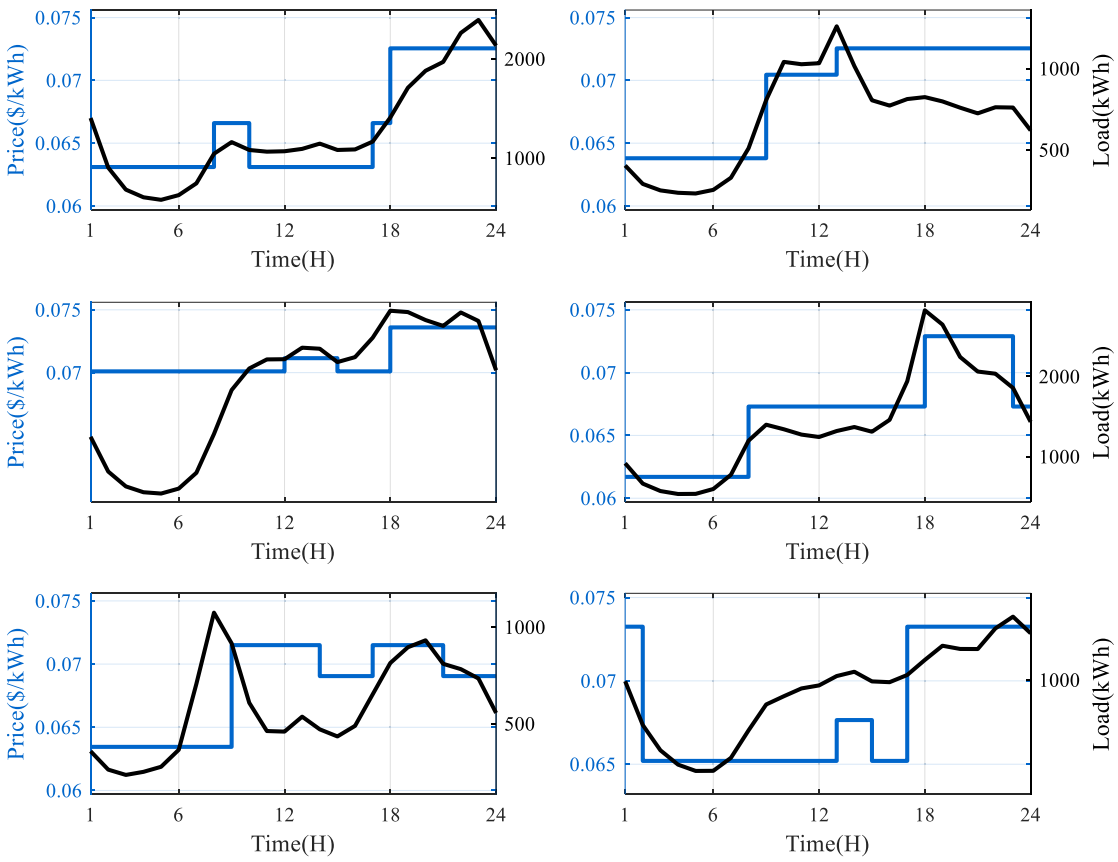


Fig. 6.9 TOU price structure for each cluster for winter

Fig. 6.10 and Fig. 6.11 depict the purchased energy from the day-ahead market in each time period. On the other hand, those forward contracts that are selected and the amount of energy which is purchased through them are shown in Table 6-8 and Table 6-9. While the optimal decision for the retailer is to sign all three kinds of forward contracts for the summer, for the winter it is suggested to purchase the energy only for the peak hours (forward contract F1).

Fig. 6.12 and Fig. 6.13 illustrate the charge/discharge status of the storage unit in each hour. As expected, the charging happens at off-peak or the lower price periods. On the other hand, the discharge mostly occurs at peak periods. The charge/discharge patterns are

different for the summer and winter due to the different load patterns of the households in these seasons.

An interesting result is observed for time period 3 am to 6 am of summer in which the energy procurement through the day-ahead market is very low. These time periods also correspond to the charging status of the storage unit. It means that the bought energy through forward contracts is enough to supply the load and at the same time charge the battery.

Finally, the overall profits for the summer and winter are $ 885.7 and $ 1178.4 respectively which exhibit an increase compared with the previous benchmark cases.

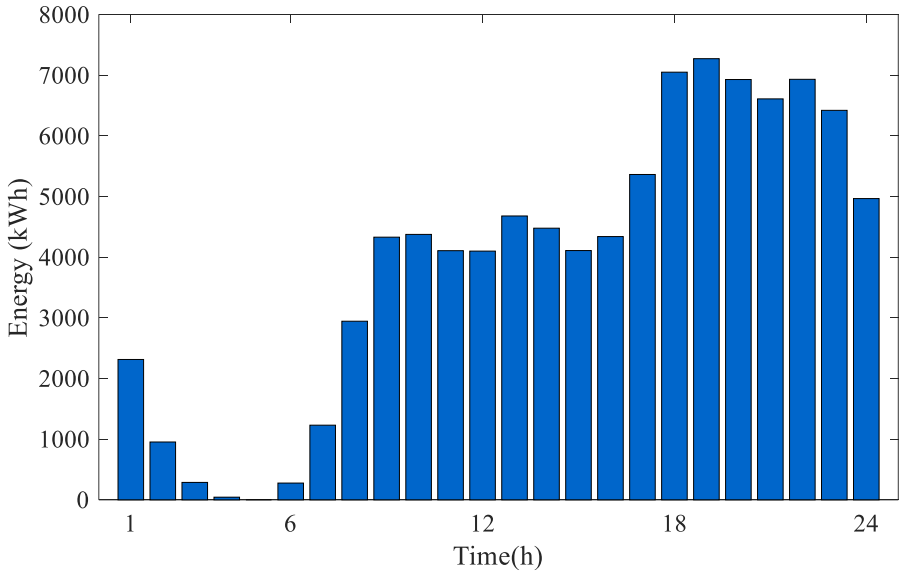


Fig. 6.10 Purchased energy from day-ahead market (summer)

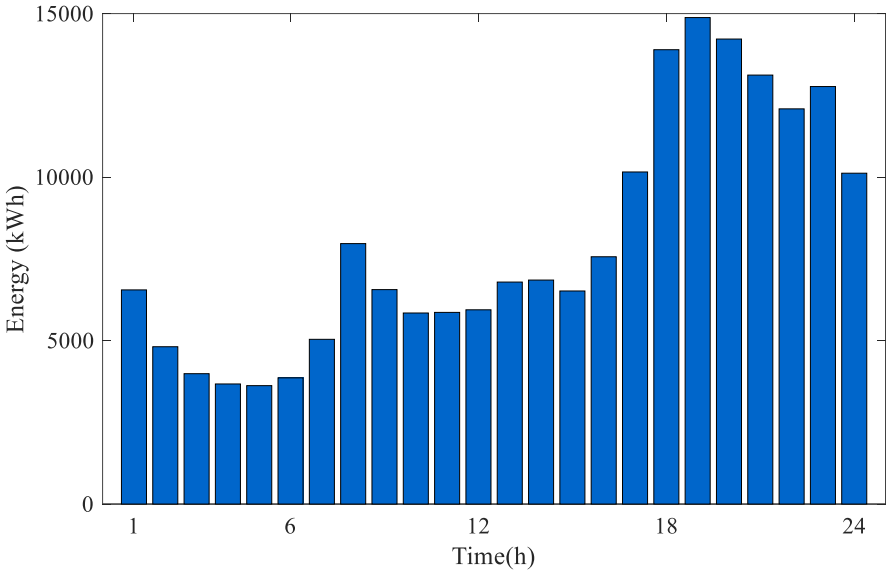


Fig. 6.11 Purchased energy from day-ahead market (winter)

Table 6-8 Purchased energy from forward contracts (summer)

	Block 1 (kWh)	Block 2 (kWh)	Block 3 (kWh)
F1	1500	--	--
F2	1300	564.4	--
F3	1400	--	--

Table 6-9 Purchased energy from forward contracts (winter)

	Block 1 (kWh)	Block 2 (kWh)	Block 3 (kWh)
F1	1500	1200	--
F2	--	--	--
F3	--	--	--

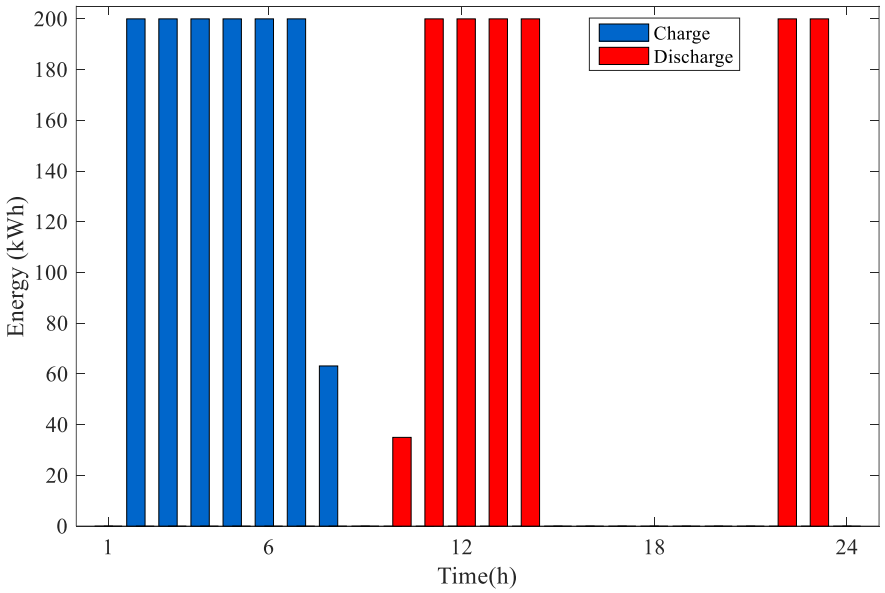


Fig. 6.12 Charge/discharge states of the storage unit (summer)

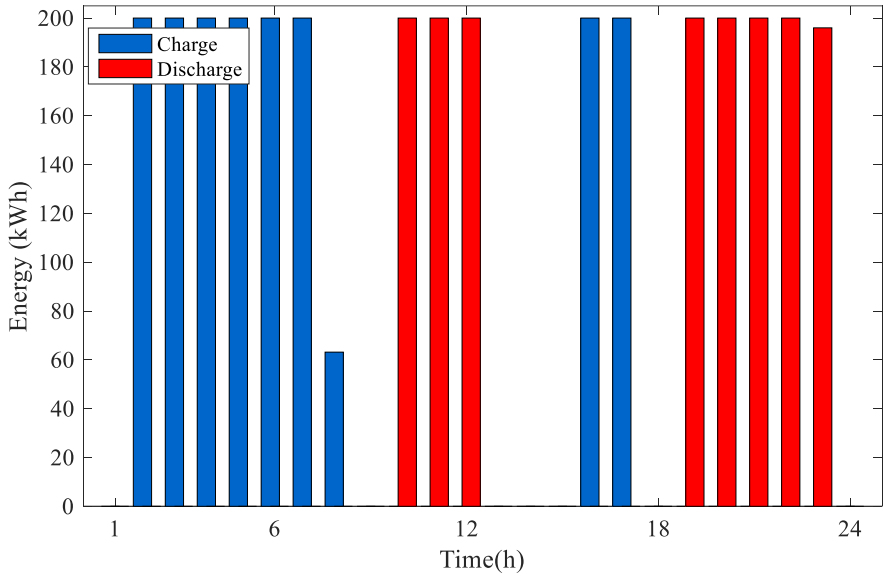


Fig. 6.13 Charge/discharge states of the storage unit (winter)

IV) Calculation of TOU prices considering the clustering/ without the forward contracts

In this section, it is assumed that the retailer does not have access to any forward contracts and procures the total demanded energy from the day-ahead market. This scenario affects not only the profits but also the TOU structures as shown by Fig. 6.14 and Fig. 6.15. For the summer and winter, the profit declines to $ 800.1 and $ 1134 respectively.

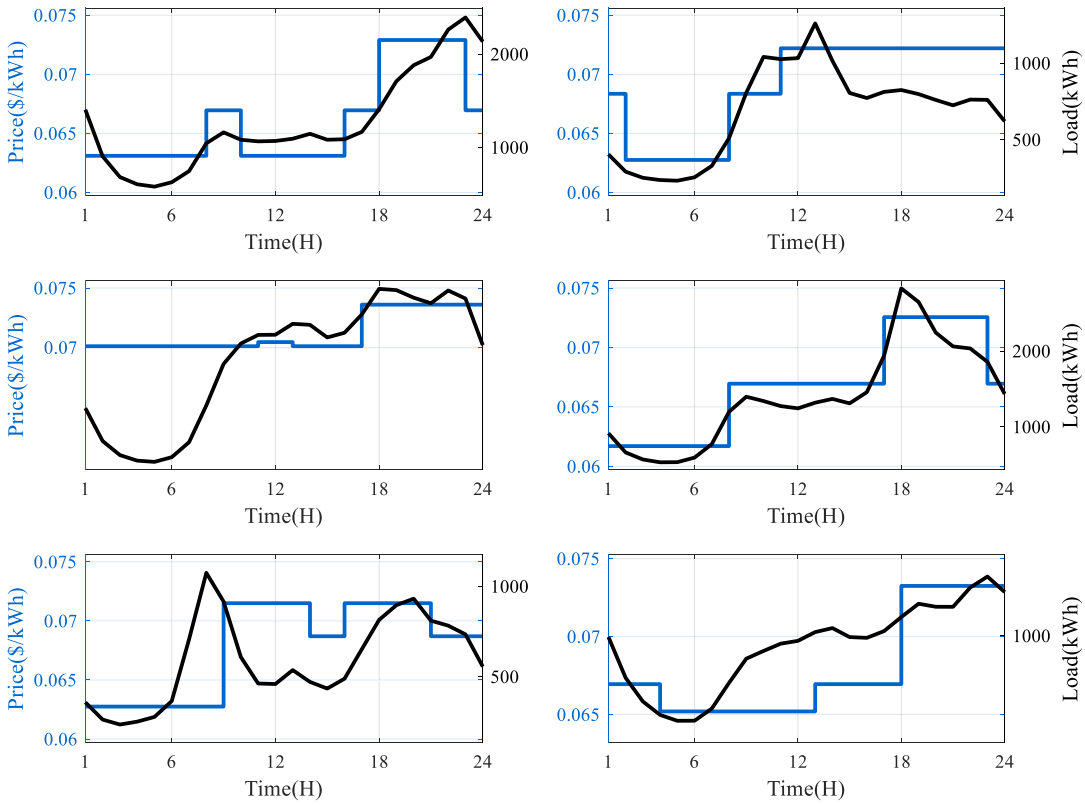


Fig. 6.14 Summer TOUs without forward contract

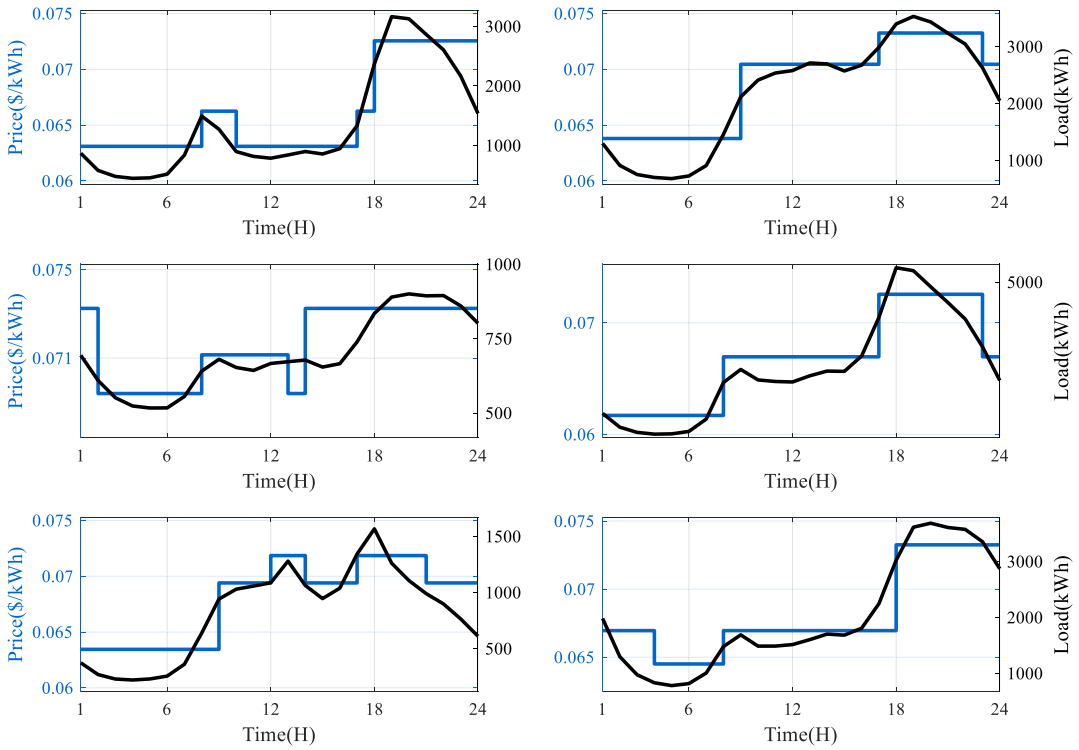


Fig. 6.15 Winter TOUs without forward contract

V) Calculation of TOU prices considering the clustering/ without the battery

Removing the battery from the analysis decreases the retailer benefits; however, it is less effective on TOU structures. It also affects the forward contracts as can be seen by Table 6-10. In this case, the retailer buys a reduced amount of energy from the second block of forward contract F2 (off-peak). It means that the energy which was previously used for charging the batteries is not purchased through the forward contracts anymore. The forward contracts for the winter are not affected in this case study.

In addition, the total profits for the summer and winter drop to $ 871.4 and $ 1150 respectively.

Table 6-10 Forward contracts when the battery is not considered (summer)

	Block 1 (kWh)	Block 2 (kWh)	Block 3 (kWh)
F1	1500	--	--
F2	1300	364.4	--
F3	1400	--	--

VI) Calculation of TOU prices/ with both selling and purchasing in day-ahead market

The retailer may also be able to sell some of its excess energy in the day-ahead market. This configuration enables the retailer to increase its profits.

For winter, the analysis shows that the retailer only purchases energy from the day-ahead market. This can be attributed to the higher level of consumption in this season.

In summer, for two hours during the off-peak period, the retailer sells some energy in the day-ahead market as shown in Fig. 6.16.

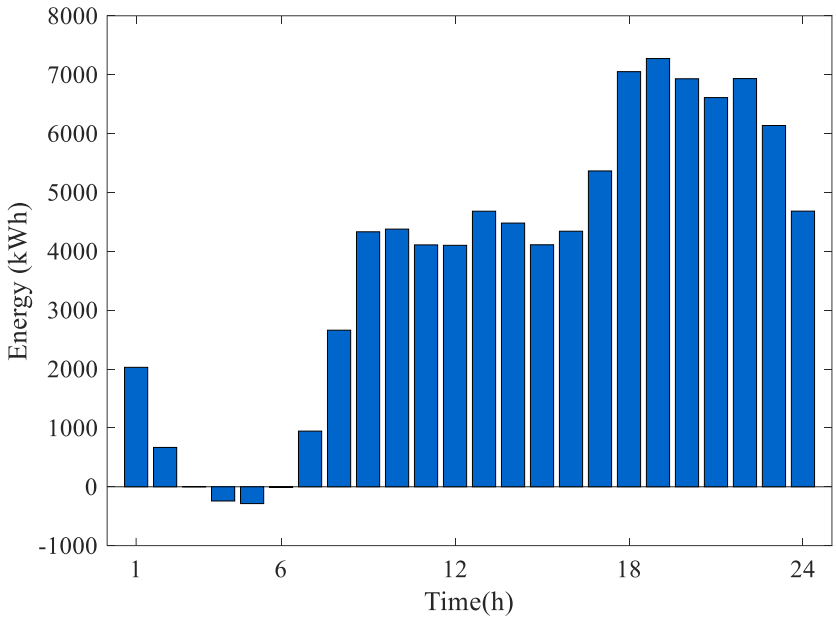


Fig. 6.16 Purchased/sold energy in day-ahead market (summer)

The charge/discharge patterns of the energy storage are not affected by this new arrangement and remain the same as the previous configuration. There is an increase in the purchased forward contracts as shown in Table 6-11. It demonstrates that the retailer prefers to buy some energy at lower prices and sells it later in the day-ahead market at higher prices. Moreover, the values of retail prices are changed for this case as displayed in Table 6-12. The structures of TOU prices change slightly. The total attained profit of retailer for the summer is $ 909.5 that shows an increase from $ 885.7 when only purchasing energy from the day-ahead market was considered.

Table 6-11 Forward contracts when selling in day-ahead market is allowed (summer)

	Block 1 (kWh)	Block 2 (kWh)	Block 3 (kWh)
F1	1500	--	--
F2	1300	850	--
F3	1400	--	--

Table 6-12 Retail prices when selling in day-ahead market is allowed (summer)

Clusters	Retail prices		
	$\boldsymbol{\lambda_k^{high}}$ ($/kWh$)	$\boldsymbol{\lambda_k^{medium}}$ ($/kWh$)	$\boldsymbol{\lambda_k^{low}}$ ($/kWh$)
Cluster 1	0.0722	0.0684	0.0652
Cluster 2	0.0708	0.0701	0.0635
Cluster 3	0.0729	0.0726	0.0656
Cluster 4	0.0719	0.0670	0.0645
Cluster 5	0.0712	0.0677	0.0638
Cluster 6	0.0733	0.0715	0.0684

VII) Impact of incorporating risk into the problem

In the previous case studies, the value of β was set to zero. To study the risk-averse behavior of the retailer, the value of β is changed from 0 to 0.7 in steps of 0.1. Table 6-13 and Table 6-14 report the results for the profit and the CVaR measure for the summer and winter respectively. As can be seen from these tables, by increasing β from 0 to 0.7, the expected profit of the retailer decreases only 3.7% and 4.1% for the summer and winter respectively. These results demonstrate the effectiveness of CVaR measure in hedging the risks. It means that by a small decrease in the profit, the retailer is able to significantly diminish the risk of experiencing low profits.

Table 6-13 Comparison of retailer's profit and risk for different values of β (summer)

β	CVaR ($)	Expected profit ($)
0	--	885.7
0.1	148.8	882.8
0.2	224.1	879.2
0.3	298.1	873.5
0.4	381.3	870.1
0.5	457.8	863.4
0.6	533.8	857.8
0.7	610.1	852.9

Table 6-14 Comparison of retailer's profit and risk for different values of β (winter)

β	CVaR ($)	Expected profit ($)
0	--	1178.4
0.1	92.5	1168.6
0.2	185.1	1160.7
0.3	276.0	1153.7
0.4	371.5	1142.6
0.5	464.2	1136.8
0.6	555.5	1130.9
0.7	650.8	1129.8

6.6 Summary

In this chapter, a comprehensive formulation for modeling the behavior of an electricity retailer in the smart grid environment was developed. It is assumed that the retailer segregates its customers into several groups and offers them cluster-specific TOUs. The retailer's decision making was considered to be constrained by the maximum and minimum retail prices and the reaction of customers to the offered prices. On the other hand, clustering allows the retailer to maximize its profits. Case studies were done for different configurations to analyse the effect of clustering, forward contracts, storage units, selling energy in day-ahead markets, and the risk measure. The obtained results can help the electricity retailers to use the smart meter data in more efficient ways to design innovative customized tariff structures for their customers.

7 Conclusions and Suggestions for Future Work

This chapter summarizes the thesis and explains the possible future directions for further research.

In Section 7.1, a general summary of the achievements and contributions of this thesis is presented. In Section 7.3, the novel and more complex approaches to the load data clustering are briefly reviewed to inform the interested readers. Finally, based on the presented concepts in these two sections, Section 7.4 identifies important areas for further research.

7.1 Summary and Conclusions

The clustering of electricity customers load data and its applications for smart grids were explored in this thesis. We tried to tackle the real-world problems with the efficient solutions using proper data mining techniques. The proposed approaches and the achieved outcomes have practical merits and can be used by either researchers or utilities for analyzing load data and offering new services to customers.

Chapter 2 was dedicated to a broad introduction of smart metering, residential consumption characteristics, effects of new technologies, clustering in power system domain and its advantages for power networks. We discussed extensively the adoption of new technologies including solar PVs, storage systems, and EVs in residential buildings and their impacts on daily consumptions. Overall, two major trends related to the use of EVs and the combined solar PV and battery were observed. While EVs can greatly change the

demand, it is expected that their penetration levels into the system would take much longer time compared with the PVs and ESSs. Besides, the complexity of EV pattern makes it harder to propose an ultimate prediction of the changes that it might cause in the consumer demand. Solar PVs combined with storage systems have the highest impact on the load shapes of the customers. They enable the owners to meet the zero-energy import from the grid in some periods and earn benefits by devising a strategy based on the TOU tariffs. However, with the current battery prices and the slow rate of the reduction of prices in recent years (Fig. 2.6), their high penetrations in the short term would not be possible.

It was also highlighted that, in spite of the great opportunities that smart meter data can provide, the route to extracting value from these data is still unclear. Therefore, we emphasized the inclusion of data mining techniques in business intelligence systems of electricity companies. As a special case, the services and advantages that clustering of load data can offer were comprehensively reviewed and described.

Following the presented concepts in Chapter 2, Chapter 3 provided an extensive comparison of the main clustering methods for segmenting the daily load curves of customers. Firstly, the mathematical formulations of five major clustering techniques along with six cluster validity indexes were presented and the parameters that can affect the outcomes of each method were illustrated in detail. In the next step, these clustering methods were applied on two different datasets and their performances in forming the clusters were compared.

Chapter 4 addressed the variability of residential load data and proposed an approach based on the SAX method to transform the variable load data into symbolic representations. The rationale behind this approach is the suitability of SAX in division of a day into several

time periods and its ability to transform the data into a limited number of symbols. After identifying the time periods from the aggregated load data, a modified SAX method was applied on a large data set of daily load curves of customers. Subsequently, the SAX words were clustered using a hierarchical clustering method. In the next part of this chapter, the notion of entropy was utilized to rank the customers based on their stability in consumption over time that can help DR aggregators in interacting with different users.

In Chapter 5, a three-stage methodology was elaborated to investigate the relationship between the socio-demographic characteristics of households and buildings features with the consumption patterns of customers. At first, the survey dataset was analyzed in different ways to sort the questions and to extract the suitable variables. Secondly, the load data of customers were represented by seven features and the users were clustered into six groups using a K-means method. Finally, the MLR was employed to find out the effect of each variable on consumption patterns. The results give insights into the affecting parameters on household energy consumption.

Chapter 6 aimed at designing TOU tariffs for residential users based on their consumption patterns. The problem was formulated as an optimization problem in which a retailer maximizes its profits. The optimization process decides the TOU structures, the values of retail prices, the charge/discharge patterns of battery units, the needed forward contracts, and the day-ahead market energy purchases. The CVaR measure was also included in the formulation to study the risk-averse problem and its effects on the profits.

7.2 Future Trends in Data Analytics of Smart Grids

In recent years, the volume of generated data in electricity grids has increased significantly. Various sources of data such as smart meters, phasor measurement units (PMUs), and distribution power quality monitors record the online data of the network and customers. The adequate analysis of these valuable data is essential for the efficient performance of smart grids.

Considering the availability of these data, several major trends can be tracked in the industry and academia as outlined in the following:

- **Data sets:** More datasets are now publicly available for research studies. In addition to the well-known CER dataset that is also used in this thesis, other consumption datasets have been published as parts of prototype projects by industry and academia. Examples include: i) Pecan Street dataset which is an ongoing project that continuously measures and collects the circuit-level electricity use of around 1000 homes (Austin, Texas) [230], ii) Smart Grid Smart City project (SGSC) that was a comprehensive project conducted from 2010 to 2014 and collected various data of electricity users including their electricity consumptions (Australia) [231], iii) UMass Start* datasets including a high-resolution dataset of 400 homes and a lower resolution dataset at the appliance level for three dwellings (Western Massachusetts, US) [232], and iv) Ausgrid Solar Home dataset containing the three-year data of 300 customers of Ausgrid company (Australia) [233].
- **Standards:** Several organizations including IEEE (Institute of Electrical and Electronics Engineers), IEC (International Electrotechnical Commission), and NIST (National Institute of Standards and Technology) have initiated standards

which directly or indirectly address the generation, collection, transfer, storage, and analysis of data in smart grids [234].

- **Projects and joint initiatives:** Various projects are currently performed worldwide by various organizations and utilities that, entirely or partly, concentrate on data science solutions for power systems [235]. Examples of projects and joints initiatives include ESSnet Big Data project [236], the Bits to Energy Lab [237], and CITIES Innovation Center [238] in Europe, projects funded by National Science Foundation (NSF) in the United States, and the increasing number of projects supported by the National Key R&D Program of China and the National Science Foundation of China. Also, leading companies in the energy sector have started working on data analysis techniques for power systems to control and monitor the network and increase their profits [239], [240].
- **Big data issues:** Big data technologies are necessary for the analysis of data in the future smart grids. Big data are usually characterized by three main features [234] [241]: i) volume which refers to the size of generated data, ii) variety that reflects that heterogeneity of data types and the existence of data in various formats, and iii) velocity that is the speed of data collection, transfer, process, and so on. A recent definition by International Data Corporation (IDC) explains big data as [242]: "big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and/or analysis". This definition specifies the most important problem in big data, which is how to discover values from datasets with an enormous scale, rapid generation, and various types [243]. To deal with the big data problem in the smart grid domain, a wide range of technologies, tools, and platforms

such as the machine learning, deep learning, cloud computing, and internet of things (IoT) can be utilized. Various issues need to be addressed including the establishment of adequate processing and computing infrastructure, collaborative efforts between the professionals from different sectors such as the engineers and data scientists, and consideration of cyber-security issues including the privacy, security, and confidentiality.

7.3 Future Directions of Load Data Clustering

The changes which are gradually happening in power systems and the advancements in data mining techniques will affect customer segmentation in various ways. The improvements in algorithms of time series clustering, advancements in the parallel, distributed, and on-line clustering, and introduction of other novel technologies in smart grids such as cloud computing [244], [245] will have a great impact on clustering of electricity customers. Along with these advances, more innovative applications can be defined for power systems. In the following, some of these approaches are briefly described.

In recent years, new methods have been evolved and applied for clustering of customers. One of these approaches is *time series clustering*. A time series is defined as a series of data points indexed in time order. These data can be the values of a quantity obtained at successive times (a collection of observations made chronologically), often with an equal interval between them [246], [150]. In a wide variety of real-world cases like the sensor data, stock markets, environmental applications, fault monitoring and medical data, the obtained data are in the form of a time-series. The measured data by smart meters also represent a time series data in which the electricity consumptions of the customers are the desired values that are continuously recorded over time.

Time series data are generally specified by some basic characteristics including the necessity to be acquired continuously and the large data size [246]. The same goals that are set for all other clustering applications are also applicable to the clustering of time series data, however, the nature of time series data poses unique challenges for applying any efficient clustering algorithm.

For time series clustering, the use of *dynamic time warping (DTW)* [247] as a similarity measure can be beneficial. The Minkowski similarity measures such as Euclidean distance are only defined for series of equal length and are sensitive to scale and time shifts [248]. They also reflect similarity in time by performing a one-to-one mapping between the data instances of the time series under comparison. On the other hand, DTW distance reflects similarity in shape by performing a one-to-many mapping, hence allowing time shifting, and thus matches similar shapes even if they have a time-phase difference [150].

It should be noticed that calculating DTW is computationally expensive. If the lengths of time series X and Y are n and m, respectively, the DTW distance between them can be computed in $O(nm)$ time, which is almost quadratic if n and m are similar [249]. For a comprehensive discussion on the DTW distance and its implementation in R, one can refer to [250], [251], and [252].

In spite of current achievements, the sheer quantity of data from smart meters poses challenges for traditional data analysis tools of utility companies. In order to deal with this big data, new infrastructure and tools are required. Companies in the energy sector, facing this challenge of big data, need to implement more powerful analysis tools to extract value

from the collected data. A few studies in the literature have addressed the "dynamic"/"online" clustering of load data and the problem of big data.

Dynamic clustering of time series data is considered in [253] and [254] to deal with the dynamic evolution of the consumption data through time. The presented framework in [253] for dynamic clustering of load curves compares the performance of K-means and FCM algorithms with different similarity measures.

Ref. [78] proposes an online clustering method for high dimensional load data. The principle behind this online time series clustering is a batch divide-and-conquer scheme in which the clustering is applied on chunks of data points and once the entire data set is scanned, it combines the results to find the final clustering. Moreover, to tackle the problem of big data, a fully distributed clustering framework is introduced in [22]. The procedure starts with dividing the data set into k parts and applying an adaptive K-means to each individual part to obtain the cluster centers. Then, these cluster centers are selected as the inputs to another novel clustering algorithm to obtain the global clustering results.

In another approach, a novel encoding engine based on an artificial neural network is developed in [133] which encodes and clusters load profiles in real-time by a distributed approach. The advantage of this neural network-based auto-encoder is that it neither needs to know anything a priori about the input, nor uses any fixed distance metrics like Euclidean distance.

Deep learning-based clustering methods are other novel trends in the clustering of smart meter data. In deep learning, multiple layers are used to extract higher-level features from raw input. These methods can be divided into two categories: two-stage approaches and integrated approaches [255]. While the former performs the feature extraction and clustering

in two stages, the latter combines the representation learning process and the clustering stage into one model. A probabilistic baseline estimation framework is proposed in [255] for DR applications. It employs a deep embedded clustering which is able to extract new features and forms the clusters jointly. A combination of deep neural networks and K-shape clustering is used in [256] for load forecasting and a joint deep learning and clustering process that captures daily and seasonal variations is proposed in [257]. Deep learning techniques are used in other studies, for example, for identifying the socio-demographic information from the load data [258] and designing incentive DR programs [259].

In addition to the application of novel clustering algorithms, the innovative techniques can be used for the improvement of other aspects of load data clustering [235]. As illustrated in this thesis, numerous CVIs are utilized to evaluate the results of clustering methods. Instead of these measures, application-oriented indexes such as the accuracy of load forecasting can be used for selecting the optimal clustering algorithm or the best results. Definition/extraction of proper features before the clustering can also improve the clustering results.

7.4 Future Work

The current research can be extended in various ways, mainly:

- Regarding DR programs, based on the goals of the end user (for example, DR aggregator or system operator), different procedures can be followed to make the use of clustering results. For example, if the stability of customers through time is considered, the methodology described in Chapter 4 can be helpful. On the other hand, if the consumption values of customers for a long period are considered, a

feature definition approach such as the one in Chapter 5 would be more beneficial. Finally, the clustering with DTW measure is more suitable when one is interested in the characteristic consumption behavior of the customer which can be shifted through time from one day to another day. This latter case is especially important when the general consumption attitudes such as the appliance uses, regardless of the time of use, are under consideration.

- In this thesis, the reaction of customers to TOU prices was formulated using an elasticity function. Upper and lower bounds were also imposed on the retail price to ensure the fair prices for the customers. The reaction of customers to the retail prices can be modeled using more sophisticated approaches. For example, a bi-level formulation can be formulated to maximize the retailer's profits and minimize the users' costs at the same time.
- The innovative clustering approaches can be pursued for the analysis and segmentation of electricity customers. These methods can also be employed for the improvement of different applications such as the enhanced load forecasting.
- With the advancements in smart meter technologies, DMS tools, and data transfer standards and protocols, fine grained electricity data with shorter time resolutions can be gathered. The immediate impact will be on the real-time operation of power networks. Until now, most of the clustering studies consider the offline data of customers. On-line clustering of load data can improve the real-time management of power systems. Possible applications include very short term load forecasting and dynamic demand response. For instance, system operators, load serving entities, and DR aggregators will be able to analyze the load consumption data at very short time

scales to forecast the electricity demand and to initiate DR programs such as load curtailments. Studying the on-line clustering of time series data and its impacts on power systems are interesting areas of research.

References

[1] S. S. S. R. Depuru, L. Wang, and V. Devabhaktuni, "Smart meters for power grid: Challenges, issues, advantages and status," *Renewable and sustainable energy reviews,* vol. 15, no. 6, pp. 2736-2742, 2011.

[2] B. Yildiz, J. I. Bilbao, J. Dore, and A. B. Sproul, "Recent advances in the analysis of residential electricity consumption and applications of smart meter data," *Applied Energy,* vol. 208, pp. 402-427, 2017/12/15/ 2017, doi: https://doi.org/10.1016/j.apenergy.2017.10.014.

[3] G. R. Barai, S. Krishnan, and B. Venkatesh, "Smart metering and functionalities of smart meters in smart grid-a review," in *Electrical Power and Energy Conference (EPEC), 2015 IEEE*, 2015: IEEE, pp. 138-145.

[4] J. Leiva, A. Palacios, and J. A. Aguado, "Smart metering trends, implications and necessities: A policy review," *Renewable and Sustainable Energy Reviews,* vol. 55, pp. 227-233, 2016.

[5] "Realizing the Full Potential of Smart Metering," *Accenture's Digitally Enabled Grid Program, Available online: https://www.accenture.com/us-en/~/media/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Industries_9/Accenture-Smart-Metering-Report-Digitally-Enabled-Grid.pdf*.

[6] T. Zhang, G. Zhang, J. Lu, X. Feng, and W. Yang, "A new index and classification approach for load pattern analysis of large electricity customers," *IEEE Transactions on Power Systems,* vol. 27, no. 1, pp. 153-160, 2012.

[7] M. Espinoza, C. Joye, R. Belmans, and B. De Moor, "Short-term load forecasting, profile identification, and customer segmentation: a methodology based on periodic time series," *IEEE Transactions on Power Systems,* vol. 20, no. 3, pp. 1622-1630, 2005.

[8] F. L. Quilumba, W.-J. Lee, H. Huang, D. Y. Wang, and R. L. Szabados, "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities," *IEEE Transactions on Smart Grid,* vol. 6, no. 2, pp. 911-918, 2015.

[9] X. Fang, S. Misra, G. Xue, and D. Yang, "Managing smart grid information in the cloud: opportunities, model, and applications," *IEEE network,* vol. 26, no. 4, 2012.

[10] C. Flath, D. Nicolay, T. Conte, C. van Dinther, and L. Filipova-Neumann, "Cluster analysis of smart metering data," *Business & Information Systems Engineering,* vol. 4, no. 1, pp. 31-39, 2012.

[11] P. Cichosz, *Data mining algorithms: explained using R*. John Wiley & Sons, 2014.

[12] "Commission for Energy Regulation (CER)- Electricity Smart Metering Customer Behaviour Trials (CBT)- Findings Report [Online] http://www.cer.ie/docs/000340/cer11080(a)(i).pdf." [Online]. Available: http://www.cer.ie/docs/000340/cer11080(a)(i).pdf.

[13] C. Muscas, M. Pau, P. A. Pegoraro, and S. Sulis, "Smart electric energy measurements in power distribution grids," *IEEE Instrumentation & Measurement Magazine,* vol. 18, no. 1, pp. 17-21, 2015.

[14] Y. Kabalci, "A survey on smart metering and smart grid communication," *Renewable and Sustainable Energy Reviews,* vol. 57, pp. 302-318, 2016.

[15] "2009 annual metering report: part a – evolving technology and process.," *Available online: www.aemo.com.au/media/Files/Other/electricityops/0600-0007%20pdf.pdf*. [Online]. Available: www.aemo.com.au/media/Files/Other/electricityops/0600-0007%20pdf.pdf.

[16] N. Uribe-Pérez, L. Hernández, D. de la Vega, and I. Angulo, "State of the Art and Trends Review of Smart Metering in Electricity Grids," *Applied Sciences,* vol. 6, no. 3, p. 68, 2016.

[17] L. Alejandro *et al.*, "Global market for smart electricity meters: Government policies driving strong growth," *Office of Industries, US International Trade Commission, Tech. Rep. ID-037,* 2014.

[18] "US Energy Information Administration (EIA)," *Available online: http://www.eia.gov/*.

[19] "Information provided by Victorian Government about smart meters. Available online: http://www.smartmeters.vic.gov.au/."

[20] "Essential Services Commission of Victoria, Available online: http://www.esc.vic.gov.au/."

[21] "ANSI C12, Smart Grid Meter Package," *American National Standard Institute, ,* 2014.

[22] Y. Wang, Q. Chen, C. Kang, and Q. Xia, "Clustering of electricity consumption behavior dynamics toward big data applications," *IEEE Transactions on Smart Grid,* vol. 7, no. 5, pp. 2437-2447, 2016.

[23] T. Räsänen and M. Kolehmainen, "Feature-based clustering for electricity use time series data," in *International Conference on Adaptive and Natural Computing Algorithms*, 2009: Springer, pp. 401-412.

[24] F. Elkarmi, "Load research as a tool in electric power system planning, operation, and control—The case of Jordan," *Energy Policy,* vol. 36, no. 5, pp. 1757-1763, 2008.

[25] B. Stephen, A. J. Mutanen, S. Galloway, G. Burt, and P. Järventausta, "Enhanced load profiling for residential network customers," *IEEE Transactions on Power Delivery,* vol. 29, no. 1, pp. 88-96, 2014.

[26] F. McLoughlin, A. Duffy, and M. Conlon, "Evaluation of time series techniques to characterise domestic electricity demand," *Energy,* vol. 50, pp. 120-130, 2013.

[27] S. Haben, C. Singleton, and P. Grindrod, "Analysis and clustering of residential customers energy behavioral demand using smart meter data," *IEEE Transactions on Smart Grid,* vol. 7, no. 1, pp. 136-144, 2016.

[28] K. Gram‐Hanssen, "Standby consumption in households analyzed with a practice theory approach," *Journal of Industrial Ecology,* vol. 14, no. 1, pp. 150-165, 2010.

[29] T. Teeraratkul, D. O'neilly, and S. Lallz, "Condensed representation and individual prediction of consumer demand," in *Smart Energy Grid Engineering, 2016 IEEE*, 2016: IEEE, pp. 11-16.

[30] R. Li, C. Gu, F. Li, G. Shaddick, and M. Dale, "Development of low voltage network templates—Part I: Substation clustering and classification," *IEEE Transactions on Power Systems,* vol. 30, no. 6, pp. 3036-3044, 2015.

[31] R. Li, C. Gu, F. Li, G. Shaddick, and M. Dale, "Development of low voltage network templates—Part II: Peak load estimation by clusterwise regression," *IEEE Transactions on Power Systems,* vol. 30, no. 6, pp. 3045-3052, 2015.

[32] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy,* vol. 42, no. 1, pp. 68-80, 2012.

[33] G. Chicco, R. Napoli, and F. Piglione, "Application of clustering algorithms and self organising maps to classify electricity customers," in *Power Tech Conference Proceedings, 2003 IEEE Bologna*, 2003, vol. 1: IEEE, p. 7 pp. Vol. 1.

[34] "Renewables 2020 Global Status Report," *REN21,* 2020. [Online]. Available: https://www.ren21.net/gsr-2020/.

[35] "Global Market Outlook For Solar Power / 2019 - 2023," *SolarPower Europe,* 2019. [Online]. Available: https://www.solarpowereurope.org/global-market-outlook-2019-2023.

[36] "Trends in Photovoltaic Applications 2019," *International Energy Agency Photovoltaic Power Systems Programme (IEA PVSP)*. [Online]. Available: https://iea-pvps.org/trends_reports/2019-edition/.

[37] M. Hannan, M. Hoque, A. Mohamed, and A. Ayob, "Review of energy storage systems for electric vehicle applications: Issues and challenges," *Renewable and Sustainable Energy Reviews,* vol. 69, pp. 771-789, 2017.

[38] "Emerging Technologies Information Paper," *Australian Energy Market Operator (AEMO),* 2015.

[39] "Better Batteries," *BloombergNEF (BNEF)*. [Online]. Available: https://www.bloomberg.com/quicktake/batteries.

[40] "2020 Battery Market Report " *SunWiz* 2020. [Online]. Available: https://www.sunwiz.com.au/battery-market-report-australia-2020/.

[41] M. Campbell, A. Miller, and N. Watson, "Impacts of new technologies on load profiles," 2016.

[42] M. Jack and K. Suomalainen, "Potential future changes to residential electricity load profiles–findings from the GridSpy dataset," 2018.

[43] J. Suppers, "Impacts of new technologies on household electricity demand: From an individual household, a community, and a national perspective," PhD thesis, The University of Waikato, 2018.

[44] "Electric cars, solar panels, and batteries in New Zealand: The benefits and costs to consumers and society," Concept Consulting Group Ltd, 2016.

[45] P. Grahn, "Electric vehicle charging impact on load profile," PhD thesis, KTH Royal Institute of Technology, 2013.

[46] "Is Home Battery Storage Worth It? ." https://www.solarchoice.net.au/is-home-battery-storage-worth-it (accessed July 2020).

[47] "Virtual Power Plant in South Australia- Stages 1 & 2 Reports," AGL Energy Limited, 2017.

[48] W. Labeeuw, J. Stragier, and G. Deconinck, "Potential of active demand reduction with residential wet appliances: A case study for Belgium," *IEEE Transactions on Smart Grid,* vol. 6, no. 1, pp. 315-323, 2015.

[49] L. Lutzenhiser, "Behavioral assumptions underlying California residential energy efficiency programs," *CIEE Energy & Behavior Program,* 2009.

[50] L. Dethman and D. Thomley, "Comparison of Segmentation Plans for Residential Customers," *Energy Trust,* 2009.

[51] M. Nachreiner, B. Mack, E. Matthies, and K. Tampe-Mai, "An analysis of smart metering information systems: A psychological model of self-regulated behavioural change," *Energy research & social science,* vol. 9, pp. 85-97, 2015.

[52] A. Albert and R. Rajagopal, "Smart meter driven segmentation: What your consumption says about you," *IEEE Transactions on power systems,* vol. 28, no. 4, pp. 4019-4030, 2013.

[53] B. A. Smith, J. Wong, and R. Rajagopal, "A simple way to use interval data to segment residential customers for energy efficiency and demand response program targeting," in *ACEEE Summer Study on Energy Efficiency in Buildings*, 2012, vol. 5, pp. 374-386.

[54] J. Kwac, J. Flora, and R. Rajagopal, "Household energy consumption segmentation using hourly data," *IEEE Transactions on Smart Grid,* vol. 5, no. 1, pp. 420-430, 2014.

[55] Y. Wang, Q. Chen, C. Kang, M. Zhang, K. Wang, and Y. Zhao, "Load profiling and its application to demand response: A review," *Tsinghua Science and Technology,* vol. 20, no. 2, pp. 117-129, 2015.

[56] J. D. Rhodes, W. J. Cole, C. R. Upshaw, T. F. Edgar, and M. E. Webber, "Clustering analysis of residential electricity demand profiles," *Applied Energy,* vol. 135, pp. 461-471, 2014.

[57] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu, and C. Toader, "Comparisons among clustering techniques for electricity customer classification," *IEEE TRANSACTIONS ON POWER SYSTEMS PWRS,* vol. 21, no. 2, p. 933, 2006.

[58] F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," *Applied energy,* vol. 141, pp. 190-199, 2015.

[59] Y.-I. Kim, J.-M. Ko, and S.-H. Choi, "Methods for generating TLPs (typical load profiles) for smart grid-based energy programs," in *Computational Intelligence Applications In Smart Grid (CIASG), 2011 IEEE Symposium on*, 2011: IEEE, pp. 1-6.

[60] G. Tsekouras, P. Kotoulas, C. Tsirekis, E. Dialynas, and N. Hatziargyriou, "A pattern recognition methodology for evaluation of load profiles and typical days of large electricity customers," *Electric Power Systems Research,* vol. 78, no. 9, pp. 1494-1510, 2008.

[61] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," *IEEE Transactions on power systems,* vol. 20, no. 2, pp. 596-602, 2005.

[62] J. L. Viegas, S. M. Vieira, R. Melício, V. Mendes, and J. M. Sousa, "Classification of new electricity customers based on surveys and smart metering data," *Energy,* vol. 107, pp. 804-817, 2016.

[63] S. Ramos, J. M. Duarte, F. J. Duarte, and Z. Vale, "A data-mining-based methodology to support MV electricity customers' characterization," *Energy and Buildings,* vol. 91, pp. 16-25, 2015.

[64] A. Ozawa, R. Furusato, and Y. Yoshida, "Determining the relationship between a household's lifestyle and its electricity consumption in Japan by analyzing measured electric load profiles," *Energy and Buildings,* vol. 119, pp. 200-210, 2016.

[65] F. Fahiman, S. M. Erfani, S. Rajasegarar, M. Palaniswami, and C. Leckie, "Improving load forecasting based on deep learning and K-shape clustering," in *Neural Networks (IJCNN), 2017 International Joint Conference on*, 2017: IEEE, pp. 4134-4141.

[66] D. Zhou, M. Balandat, and C. Tomlin, "Residential demand response targeting using machine learning with observational data," in *Decision and Control (CDC), 2016 IEEE 55th Conference on*, 2016: IEEE, pp. 6663-6668.

[67] K. Zhou, S. Yang, and Z. Shao, "Household monthly electricity consumption pattern mining: A fuzzy clustering-based model and a case study," *Journal of Cleaner Production,* vol. 141, pp. 900-908, 2017.

[68] A. M. Ferreira, C. A. Cavalcante, C. H. Fontes, and J. E. Marambio, "A new method for pattern recognition in load profiles to support decision-making in the management of the electric sector," *International Journal of Electrical Power & Energy Systems,* vol. 53, pp. 824-831, 2013.

[69] K. Zhou, C. Yang, and J. Shen, "Discovering residential electricity consumption patterns through smart-meter data mining: A case study from China," *Utilities Policy,* vol. 44, pp. 73-84, 2017.

[70] M. Crow, "Clustering-based methodology for optimal residential time of use design structure," in *North American Power Symposium (NAPS), 2014*, 2014: IEEE, pp. 1-6.

[71] H. Liu, Z. Yao, T. Eklund, and B. Back, "From Smart Meter Data to Pricing Intelligence--Visual Data Mining towards Real-Time BI," *Turku Centre for Comput. Sci., Finland, TUCS Tech. Rep. No. 1035,* 2012.

[72] O. Motlagh, G. Foliente, and G. Grozev, "Knowledge-mining the Australian smart grid smart city data: A statistical-neural approach to demand-response analysis," in *Planning Support Systems and Smart Cities*: Springer, 2015, pp. 189-207.

[73] I. B. Sanchez, I. D. Espinós, L. M. Sarrión, A. Q. López, and I. N. Burgos, "Clients segmentation according to their domestic energy consumption by the use of self-organizing maps," in *Proc. 6th Int. Conf. Eur. Energy Market (EEM'09)*, May 2009: IEEE, pp. 1-6.

[74] T. Räsänen, J. Ruuskanen, and M. Kolehmainen, "Reducing energy consumption by using self-organizing maps to create more personalized electricity use information," *Applied Energy,* vol. 85, no. 9, pp. 830-840, 2008.

[75] F. Mc Loughlin, A. Duffy, and M. Conlon, "Analysing domestic electricity smart metering data using self organising maps," in *Integration of Renewables into the Distribution Grid, CIRED 2012 Workshop*, 2012: IET, pp. 1-4.

[76] S. V. Verdú, M. O. Garcia, C. Senabre, A. G. Marín, and F. J. G. Franco, "Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps," *IEEE Transactions on Power Systems,* vol. 21, no. 4, pp. 1672-1682, 2006.

[77] M. Koivisto, P. Heine, I. Mellin, and M. Lehtonen, "Clustering of connection points and load modeling in distribution systems," *IEEE Transactions on Power Systems,* vol. 28, no. 2, pp. 1255-1265, 2013.

[78] R. Pal, C. Chelmis, M. Frincu, and V. Prasanna, "Time Series Clustering for Demand Response, An Online Algorithmic Approach," *Online: www.cs.usc.edu/assets/007/93954.pdf*.

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

[80] J. Yang, J. Zhao, F. Wen, and Z. Dong, "A Model of Customizing Electricity Retail Prices Based on Load Profile Clustering Analysis," *IEEE Transactions on Smart Grid,* 2018.

[81] F. Wang *et al.*, "Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns," *Energy conversion and management,* vol. 171, pp. 839-854, 2018.

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

[83] L. A. P. Júnior *et al.*, "Unsupervised non-technical losses identification through optimum-path forest," *Electric Power Systems Research,* vol. 140, pp. 413-423, 2016.

[84] J. L. Viegas, P. R. Esteves, and S. M. Vieira, "Clustering-based novelty detection for identification of non-technical losses," *International Journal of Electrical Power &*

Energy Systems, vol. 101, pp. 301-310, 2018/10/01/ 2018, doi: https://doi.org/10.1016/j.ijepes.2018.03.031.

[85] S. K. Bhatia, "Adaptive K-Means Clustering," in *Proc. Int. Florida Artif. Intell. Res. Soc. Conf.*, 2004.

[86] J. Paparrizos and L. Gravano, "k-shape: Efficient and accurate clustering of time series," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 2015: ACM, pp. 1855-1870.

[87] F. Divina, A. F. Goméz Vela, and M. García Torres, "Biclustering of Smart Building Electric Energy Consumption Data," *Applied Sciences,* vol. 9, no. 2, 2019, doi: 10.3390/app9020222.

[88] D. Alahakoon and X. Yu, "Smart electricity meter data intelligence for future energy systems: A survey," *IEEE Transactions on Industrial Informatics,* vol. 12, no. 1, pp. 425-436, 2016.

[89] B. Wixom and H. Watson, "The BI-based organization," 2012.

[90] S.-l. Yang and C. Shen, "A review of electric load classification in smart grid environment," *Renewable and Sustainable Energy Reviews,* vol. 24, pp. 103-110, 2013.

[91] Q. C. Yi Wang, Tao Hong, Chongqing Kang, "Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges," *IEEE Transactions on Smart Grid,* 2018.

[92] M. S. Al-Musaylh, R. C. Deo, J. F. Adamowski, and Y. Li, "Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated

demand data in Queensland, Australia," *Advanced Engineering Informatics,* vol. 35, pp. 1-16, 2018/01/01/ 2018, doi: https://doi.org/10.1016/j.aei.2017.11.002.

[93] M. Piao, H. G. Lee, J. H. Park, and K. H. Ryu, "Application of Classification Methods for Forecasting Mid-Term Power Load Patterns," *Advanced Intelligent Computing Theories and Applications,* p. 47, 2008.

[94] A. Shahzadeh, A. Khosravi, and S. Nahavandi, "Improving load forecast accuracy by clustering consumers using smart meter data," in *Neural Networks (IJCNN), 2015 International Joint Conference on*, 2015: IEEE, pp. 1-7.

[95] W. Li and Z.-H. Han, "Short-term power load forecasting using improved ant colony clustering," in *Knowledge Discovery and Data Mining, 2008. WKDD 2008. First International Workshop on*, 2008: IEEE, pp. 221-224.

[96] M. Misiti, Y. Misiti, G. Oppenheim, and J.-M. Poggi, "Optimized clusters for disaggregated electricity load forecasting," *REVSTAT – Statistical Journal,* vol. 8, no. 2, pp. 105-124, 2010.

[97] T. Hong, *Short-term electric load forecasting*. Ph.D.dissertation, North Carolina State University, 2010.

[98] J. W. Taylor, "Triple seasonal methods for short-term electricity demand forecasting," *European Journal of Operational Research,* vol. 204, no. 1, pp. 139-152, 2010.

[99] X. Pan and B. Lee, "A comparison of support vector machines and artificial neural networks for mid-term load forecasting," in *Industrial Technology (ICIT), 2012 IEEE International Conference on*, 2012: IEEE, pp. 95-101.

[100] N. An, W. Zhao, J. Wang, D. Shang, and E. Zhao, "Using multi-output feedforward neural network with empirical mode decomposition based signal filtering for electricity demand forecasting," *Energy,* vol. 49, pp. 279-288, 2013.

[101] T. K. Wijaya, M. Vasirani, S. Humeau, and K. Aberer, "Cluster-based aggregate forecasting for residential electricity demand using smart meter data," in *Big Data (Big Data), 2015 IEEE International Conference on*, 2015: IEEE, pp. 879-887.

[102] X. Bai *et al.*, "A spatial load forecasting method based on the theory of clustering analysis," *Physics Procedia,* vol. 24, pp. 176-183, 2012.

[103] C. Alzate and M. Sinn, "Improved electricity load forecasting via kernel spectral clustering of smart meters," in *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, 2013: IEEE, pp. 943-948.

[104] P. Laurinec and M. Lucká, "Interpretable multiple data streams clustering with clipped streams representation for the improvement of electricity consumption forecasting," *Data Mining and Knowledge Discovery,* journal article November 16 2018, doi: 10.1007/s10618-018-0598-2.

[105] P. Laurinec and M. Lucká, "Clustering-based forecasting method for individual consumers electricity load using time series representations," *Open Computer Science,* vol. 8, no. 1, pp. 38-50.

[106] T. Jarábek, P. Laurinec, and M. Lucká, "Energy load forecast using S2S deep neural networks with k-Shape clustering," in *Informatics, 2017 IEEE 14th International Scientific Conference on*, 2017: IEEE, pp. 140-145.

[107] Y. Chen, H. Tan, and U. Berardi, "Day-ahead prediction of hourly electric demand in non-stationary operated commercial buildings: A clustering-based hybrid approach," *Energy and Buildings,* vol. 148, pp. 228-237, 2017.

[108] E. Y. Shchetinin, "Cluster-Based Energy Consumption Forecasting in Smart Grids," Cham, 2018: Springer International Publishing, in Distributed Computer and Communication Networks, pp. 445-456.

[109] Y.-S. Huang and J.-J. Deng, "Short-Term Load Forecasting Based on Ant Colony Fuzzy Clustering and SVM Algorithm," in *Fourth International Conference on Natural Computation*, 2008: IEEE, pp. 162-166.

[110] J. Hao, D. Liu, Z. Li, Z. Chen, and L. Kong, "Power system load forecasting based on fuzzy clustering and gray target theory," *Energy Procedia,* vol. 16, pp. 1852-1859, 2012.

[111] Y.-S. Huang and J.-J. Deng, "Short-term load forecasting based on ant colony fuzzy clustering and SVM algorithm," in *Natural Computation, 2008. ICNC'08. Fourth International Conference on*, 2008, vol. 2: IEEE, pp. 162-166.

[112] J. S. Vardakas, N. Zorba, and C. V. Verikoukis, "A survey on demand response programs in smart grids: pricing methods and optimization algorithms," *IEEE Communications Surveys & Tutorials,* vol. 17, no. 1, pp. 152-178, 2015.

[113] T. Samad, E. Koch, and P. Stluka, "Automated demand response for smart buildings and microgrids: The state of the practice and research challenges," *Proceedings of the IEEE,* vol. 104, no. 4, pp. 726-744, 2016.

[114] "Benefits of demand response in electricity markets and recommendations for achieving them," *US Dept. Energy, Washington, DC, USA, Tech. Rep,* 2006.

[115] L. Hancher, X. He, I. Azevedo, N. Keyaerts, L. Meeus, and J. Glachant, "Shift, not drift: Towards active demand response and beyond (Draft version “V2” Last update 03/05/2013)," *European University Institute (EUI),* 2013.

[116] L. A. Greening, "Demand response resources: Who is responsible for implementation in a deregulated market?," *Energy,* vol. 35, no. 4, pp. 1518-1525, 2010.

[117] R. J. Bessa and M. A. Matos, "Optimization models for EV aggregator participation in a manual reserve market," *IEEE Transactions on Power Systems,* vol. 28, no. 3, pp. 3085-3095, 2013.

[118] A. Mammoli, H. Barsun, R. Burnett, J. Hawkins, and J. Simmins, "Using high-speed demand response of building HVAC systems to smooth cloud-driven intermittency of distributed solar photovoltaic generation," in *Transmission and Distribution Conference and Exposition (T&D), 2012 IEEE PES*, 2012: IEEE, pp. 1-10.

[119] S. Ghavidel, J. Aghaei, K. M. Muttaqi, and A. Heidari, "Renewable energy management in a remote area using Modified Gravitational Search Algorithm," *Energy,* vol. 97, pp. 391-399, 2016.

[120] M. J. Ghadi, S. H. Gilani, H. Afrakhte, and A. Baghramian, "A novel heuristic method for wind farm power prediction: A case study," *International Journal of Electrical Power & Energy Systems,* vol. 63, pp. 962-970, 2014.

[121] S. H. Gilani, H. Afrakhte, and M. J. Ghadi, "Probabilistic method for optimal placement of wind-based distributed generation with considering reliability

improvement and power loss reduction," in *Thermal Power Plants, 2012 4th Conference on*, 2012: IEEE, pp. 1-6.

[122] J. Aghaei, M. Barani, M. Shafie-Khah, A. A. S. de la Nieta, and J. P. Catalão, "Risk-constrained offering strategy for aggregated hybrid power plant including wind power producer and demand response provider," *IEEE Transactions on Sustainable Energy,* vol. 7, no. 2, pp. 513-525, 2016.

[123] S. Ghavidel, L. Li, J. Aghaei, T. Yu, and J. Zhu, "A review on the virtual power plant: Components and operation systems," in *Power System Technology (POWERCON), 2016 IEEE International Conference on*, 2016: IEEE, pp. 1-6.

[124] A. Rajabi, L. Li, J. Zhang, and J. Zhu, "Aggregation of small loads for demand response programs—Implementation and challenges: A review," in *Environment and Electrical Engineering and 2017 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe), 2017 IEEE International Conference on*, 2017: IEEE, pp. 1-6.

[125] M. H. Abbasi *et al.*, "Risk-constrained offering strategies for a price-maker demand response aggregator," in *Electrical Machines and Systems (ICEMS), 2017 20th International Conference on*, 2017: IEEE, pp. 1-6.

[126] R. Deng, Z. Yang, M.-Y. Chow, and J. Chen, "A survey on demand response in smart grids: Mathematical models and approaches," *IEEE Transactions on Industrial Informatics,* vol. 11, no. 3, pp. 570-582, 2015.

[127] "Assessment of demand-response and advanced metering," ed: Federal Energy Regulatory Commission (FERC), Feb 2011.

[128] J. Aghaei and M.-I. Alizadeh, "Demand response in smart electricity grids equipped with renewable energy sources: A review," *Renewable and Sustainable Energy Reviews,* vol. 18, pp. 64-72, 2// 2013, doi: http://dx.doi.org/10.1016/j.rser.2012.09.019.

[129] S. Mohagheghi, J. Stoupis, Z. Wang, Z. Li, and H. Kazemzadeh, "Demand response architecture: Integration into the distribution management system," in *Smart Grid Communications, 2010 First IEEE International Conference on*, 2010: IEEE, pp. 501-506.

[130] P. Siano, "Demand response and smart grids—A survey," *Renewable and Sustainable Energy Reviews,* vol. 30, pp. 461-478, 2// 2014, doi: http://dx.doi.org/10.1016/j.rser.2013.10.022.

[131] A. Albert and M. Maasoumy, "Predictive segmentation of energy consumers," *Applied Energy,* vol. 177, pp. 435-448, 2016.

[132] N. Mahmoudi-Kohan, M. P. Moghaddam, and M. Sheikh-El-Eslami, "An annual framework for clustering-based pricing for an electricity retailer," *Electric Power Systems Research,* vol. 80, no. 9, pp. 1042-1048, 2010.

[133] E. D. Varga, S. F. Beretka, C. Noce, and G. Sapienza, "Robust real-time load profile encoding and classification framework for efficient power systems operation," *IEEE Transactions on Power Systems,* vol. 30, no. 4, pp. 1897-1904, 2015.

[134] R. Li, Z. Wang, C. Gu, F. Li, and H. Wu, "A novel time-of-use tariff design based on Gaussian Mixture Model," *Applied energy,* vol. 162, pp. 1530-1536, 2016.

[135] J. L. Viegas, P. R. Esteves, R. Melício, V. Mendes, and S. M. Vieira, "Solutions for detection of non-technical losses in the electricity grid: a review," *Renewable and Sustainable Energy Reviews,* vol. 80, pp. 1256-1268, 2017.

[136] E. W. S. Angelos, O. R. Saavedra, O. A. C. Cortés, and A. N. de Souza, "Detection and identification of abnormalities in customer consumptions in power distribution systems," *IEEE Transactions on Power Delivery,* vol. 26, no. 4, pp. 2436-2442, 2011.

[137] R. Jiang, R. Lu, Y. Wang, J. Luo, C. Shen, and X. S. Shen, "Energy-theft detection issues for advanced metering infrastructure in smart grid," *Tsinghua Science and Technology,* vol. 19, no. 2, pp. 105-120, 2014.

[138] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohamad, "Nontechnical loss detection for metered customers in power utility using support vector machines," *IEEE transactions on Power Delivery,* vol. 25, no. 2, pp. 1162-1171, 2010.

[139] S. McLaughlin, B. Holbert, A. Fawaz, R. Berthier, and S. Zonouz, "A multi-sensor energy theft detection framework for advanced metering infrastructures," *IEEE Journal on Selected Areas in Communications,* vol. 31, no. 7, pp. 1319-1330, 2013.

[140] C. León, F. Biscarri, I. Monedero, J. I. Guerrero, J. Biscarri, and R. Millán, "Integrated expert system applied to the analysis of non-technical losses in power utilities," *Expert systems with applications,* vol. 38, no. 8, pp. 10274-10285, 2011.

[141] S.-C. Huang, Y.-L. Lo, and C.-N. Lu, "Non-technical loss detection using state estimation and analysis of variance," *IEEE Transactions on Power Systems,* vol. 28, no. 3, pp. 2959-2966, 2013.

[142] G. M. Messinis and N. D. Hatziargyriou, "Review of non-technical loss detection methods," *Electric Power Systems Research,* vol. 158, pp. 250-266, 2018.

[143] J. E. Cabral, J. O. Pinto, and A. M. Pinto, "Fraud detection system for high and low voltage electricity consumers based on data mining," in *Power & Energy Society General Meeting, 2009. PES'09. IEEE*, 2009: IEEE, pp. 1-5.

[144] S. Y. Han, J. No, J.-H. Shin, and Y. Joo, "Conditional abnormality detection based on AMI data mining," *IET Generation, Transmission & Distribution,* vol. 10, no. 12, pp. 3010-3016, 2016.

[145] V. B. Krishna, G. A. Weaver, and W. H. Sanders, "PCA-based method for detecting integrity attacks on advanced metering infrastructure," in *International Conference on Quantitative Evaluation of Systems*, 2015: Springer, pp. 70-85.

[146] A. H. Nizar, Z. Y. Dong, and J. Zhao, "Load profiling and data mining techniques in electricity deregulated market," in *Power Engineering Society General Meeting, 2006. IEEE*: IEEE, p. 7 pp.

[147] A. Primadianto and C.-N. Lu, "A review on distribution system state estimation," *IEEE Transactions on Power Systems,* vol. 32, no. 5, pp. 3875-3883, 2017.

[148] "Australia's Smart Grid, Smart City Project (SGSC)," *Reports available at: http://www.environment.gov.au/energy/programs/smartgridsmartcity*.

[149] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*. Springer-Verlag New York, Inc., 2005.

[150] C. C. Aggarwal and C. K. Reddy, *Data clustering: algorithms and applications*. Chapman and Hall/CRC, 2013.

[151] I. Prahastono, D. King, and C. Ozveren, "A review of electricity load profile classification methods," in *Universities Power Engineering Conference, 2007. UPEC 2007. 42nd International*, 2007: IEEE, pp. 1187-1191.

[152] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences,* vol. 10, no. 2-3, pp. 191-203, 1984.

[153] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Transactions on neural networks,* vol. 11, no. 3, pp. 586-600, 2000.

[154] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, no. 3, pp. 381-396, 2002, doi: 10.1109/34.990138.

[155] C. C. Aggarwal and C. K. Reddy, "Data clustering," 2014.

[156] G. Gan, C. Ma, and J. Wu, *Data clustering: theory, algorithms, and applications*. Siam, 2007.

[157] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," *Expert systems with applications,* vol. 40, no. 1, pp. 200-210, 2013.

[158] K. Zhou, C. Fu, and S. Yang, "Fuzziness parameter selection in fuzzy c-means: the perspective of cluster validation," *Science China Information Sciences,* vol. 57, no. 11, pp. 1-8, 2014.

[159] I. Ozkan and I. Turksen, "Upper and lower values for the level of fuzziness in FCM," in *Fuzzy Logic*: Springer, 2007, pp. 99-112.

[160] K.-L. Wu, "Analysis of parameter selections for fuzzy c-means," *Pattern Recognition,* vol. 45, no. 1, pp. 407-415, 2012.

[161] N. R. Pal and J. C. Bezdek, "On cluster validity for the fuzzy c-means model," *IEEE Transactions on Fuzzy systems,* vol. 3, no. 3, pp. 370-379, 1995.

[162] L. O. Hall, A. M. Bensaid, L. P. Clarke, R. P. Velthuizen, M. S. Silbiger, and J. C. Bezdek, "A comparison of neural network and fuzzy clustering techniques in segmenting magnetic resonance images of the brain," *IEEE transactions on neural networks,* vol. 3, no. 5, pp. 672-682, 1992.

[163] F. Murtagh and P. Legendre, "Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?," *Journal of classification,* vol. 31, no. 3, pp. 274-295, 2014.

[164] G. McLachlan and D. Peel, *Finite mixture models*. John Wiley & Sons, 2004.

[165] "Matlab help; Online: https://au.mathworks.com/help/stats/clustering-using-gaussian-mixture-models.html."

[166] C. Fraley and A. E. Raftery, "MCLUST version 3: an R package for normal mixture modeling and model-based clustering," WASHINGTON UNIV SEATTLE DEPT OF STATISTICS, 2006.

[167] G. McLachlan and D. Peel, "Finite mixture models, willey series in probability and statistics," ed: John Wiley & Sons, New York, 2000.

[168] A. Notaristefano, G. Chicco, and F. Piglione, "Data size reduction with symbolic aggregate approximation for electrical load pattern grouping," *IET Generation, Transmission & Distribution,* vol. 7, no. 2, pp. 108-117, 2013.

[169] T.-S. Xu, H.-D. Chiang, G.-Y. Liu, and C.-W. Tan, "Hierarchical K-means method for clustering large-scale advanced metering infrastructure data," *IEEE Transactions on Power Delivery,* vol. 32, no. 2, pp. 609-616, 2017.

[170] I. P. Panapakidis, M. C. Alexiadis, and G. K. Papagiannis, "Deriving the optimal number of clusters in the electricity consumer segmentation procedure," in *European Energy Market (EEM), 2013 10th International Conference on the*, 2013: IEEE, pp. 1-8.

[171] E. Carpaneto, G. Chicco, R. Napoli, and M. Scutariu, "Electricity customer classification using frequency–domain load pattern data," *International Journal of Electrical Power & Energy Systems,* vol. 28, no. 1, pp. 13-20, 2006.

[172] B. Desgraupes, "Clustering indices," 2013.

[173] D. J. Hand, H. Mannila, and P. Smyth, *Principles of data mining*. 2001.

[174] I. Dent, T. Craig, U. Aickelin, and T. Rodden, "An approach for assessing clustering of households by electricity usage," 2012.

[175] S. Zhong and K.-S. Tam, "Hierarchical classification of load profiles based on their characteristic attributes in frequency domain," *IEEE Transactions on Power Systems,* vol. 30, no. 5, pp. 2434-2441, 2015.

[176] Y. Xiao, J. Yang, H. Que, M. J. Li, and Q. Gao, "Application of wavelet-based clustering approach to load profiling on AMI measurements," in *Electricity Distribution (CICED), 2014 China International Conference on*, 2014: IEEE, pp. 1537-1540.

[177] M. Misiti, Y. Misiti, G. Oppenheim, and J.-M. Poggi, "Clustering signals using wavelets," in *Proceedings of the 9th international work conference on Artificial neural networks*, 2007: Springer-Verlag, pp. 514-521.

[178] A. Antoniadis, X. Brossat, J. Cugliari, and J.-M. Poggi, "Clustering functional data using wavelets," *International Journal of Wavelets, Multiresolution and Information Processing,* vol. 11, no. 01, p. 1350003, 2013.

[179] I. T. Jolliffe, "Principal Component Analysis and Factor Analysis," in *Principal component analysis*: Springer, 1986, pp. 115-128.

[180] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics,* vol. 2, no. 4, pp. 433-459, 2010.

[181] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, 2003: ACM, pp. 2-11.

[182] S. Aghabozorgi and T. Y. Wah, "Clustering of large time series datasets," *Intelligent Data Analysis,* vol. 18, no. 5, pp. 793-817, 2014.

[183] G. Huebner, D. Shipworth, I. Hamilton, Z. Chalabi, and T. Oreszczyn, "Understanding electricity consumption: A comparative contribution of building factors, socio-demographics, appliances, behaviours and attitudes," *Applied energy,* vol. 177, pp. 692-702, 2016.

[184] A. Kavousian, R. Rajagopal, and M. Fischer, "Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior," *Energy,* vol. 55, pp. 184-194, 2013.

[185] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.,* vol. 27, pp. 623-656, 1948.

[186] "Met Éireann Data from Irish Meteorological Service, online: https://www.met.ie/climate/available-data/historical-data."

[187] S.-J. Kim and G. B. Giannakis, "Scalable and robust demand response with mixed-integer constraints," *IEEE Transactions on Smart Grid,* vol. 4, no. 4, pp. 2089-2099, 2013.

[188] M. Parvania, M. Fotuhi-Firuzabad, and M. Shahidehpour, "Optimal demand response aggregation in wholesale electricity markets," *IEEE Transactions on Smart Grid,* vol. 4, no. 4, pp. 1957-1965, 2013.

[189] D. T. Nguyen, H. T. Nguyen, and L. B. Le, "Dynamic Pricing Design for Demand Response Integration in Power Distribution Networks," *IEEE Transactions on Power Systems,* vol. 31, no. 5, pp. 3457-3472, 2016.

[190] L. Gkatzikis, I. Koutsopoulos, and T. Salonidis, "The role of aggregators in smart grid demand response markets," *IEEE Journal on Selected Areas in Communications,* vol. 31, no. 7, pp. 1247-1257, 2013.

[191] P. Carreira, R. Nunes, and V. Amaral, "Smartlink: A hierarchical approach for connecting smart buildings to smart grids," in *Electrical Power Quality and Utilisation (EPQU), 2011 11th International Conference on*, 2011: IEEE, pp. 1-6.

[192] A. C. Chapman, G. Verbič, and D. J. Hill, "Algorithmic and strategic aspects to integrating demand-side aggregation and energy management methods," *IEEE Transactions on Smart Grid,* vol. 7, no. 6, pp. 2748-2760, 2016.

[193] G. Di Bella, L. Giarré, M. Ippolito, A. Jean-Marie, G. Neglia, and I. Tinnirello, "Modeling energy demand aggregators for residential consumers," in *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, 2013: IEEE, pp. 6280-6285.

[194] S. I. Vagropoulos and A. G. Bakirtzis, "Optimal bidding strategy for electric vehicle aggregators in electricity markets," *IEEE Transactions on power systems,* vol. 28, no. 4, pp. 4031-4041, 2013.

[195] F. A. Rahiman, H. H. Zeineldin, V. Khadkikar, S. W. Kennedy, and V. R. Pandi, "Demand Response Mismatch (DRM): Concept, Impact Analysis, and Solution," *IEEE Transactions on Smart Grid,* vol. 5, no. 4, pp. 1734-1743, 2014.

[196] J. Medina, N. Muller, and I. Roytelman, "Demand response and distribution grid operations: Opportunities and challenges," *IEEE Transactions on Smart Grid,* vol. 1, no. 2, pp. 193-198, 2010.

[197] P. Cappers, J. MacDonald, J. Page, and J. Potter, "Future Opportunities and Challenges with Using Demand Response as a Resource in Distribution System Operation and Planning Activities," 2016.

[198] J. A. Schachter and P. Mancarella, "Demand response contracts as real options: a probabilistic evaluation framework under short-term and long-term uncertainties," *IEEE Transactions on Smart Grid,* vol. 7, no. 2, pp. 868-878, 2016.

[199] S. Vandael, B. Claessens, M. Hommelberg, T. Holvoet, and G. Deconinck, "A scalable three-step approach for demand side management of plug-in hybrid vehicles," *IEEE Transactions on Smart Grid,* vol. 4, no. 2, pp. 720-728, 2013.

[200] S. H. Madaeni and R. Sioshansi, "Measuring the benefits of delayed price-responsive demand in reducing wind-uncertainty costs," *IEEE Transactions on Power Systems,* vol. 28, no. 4, pp. 4118-4126, 2013.

[201] D. S. Callaway and I. A. Hiskens, "Achieving controllability of electric loads," *Proceedings of the IEEE,* vol. 99, no. 1, pp. 184-199, 2011.

[202] J. Hansen, J. Knudsen, and A. M. Annaswamy, "Demand response in smart grids: Participants, challenges, and a taxonomy," in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, 2014: IEEE, pp. 4045-4052.

[203] M. Bedir, E. Hasselaar, and L. Itard, "Determinants of electricity consumption in Dutch dwellings," *Energy and buildings,* vol. 58, pp. 194-207, 2013.

[204] B. Anderson, S. Lin, A. Newing, A. Bahaj, and P. James, "Electricity consumption and household characteristics: Implications for census-taking in a smart metered future," *Computers, Environment and Urban Systems,* vol. 63, pp. 58-67, 2017.

[205] F. McLoughlin, A. Duffy, and M. Conlon, "Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study," *Energy and buildings,* vol. 48, pp. 240-248, 2012.

[206] "UK Office for National Statistics," *Online: https://www.ons.gov.uk/*. [Online]. Available: https://www.ons.gov.uk/.

[207] R. Pearson, "The GoodmanKruskal package: Measuring association between categorical variables. Online: https://cran.r-project.org/web/packages/GoodmanKruskal/vignettes/GoodmanKruskal.html," ed, 2016.

[208] A. Dutt, "To eat or not to eat! That's the question? Measuring the association between categorical variables, online: https://www.r-bloggers.com/to-eat-or-not-to-eat-thats-the-question-measuring-the-association-between-categorical-variables/," 2017.

[209] R. Pearson, "Package 'GoodmanKruskal', Online: https://cran.r-project.org/web/packages/GoodmanKruskal/GoodmanKruskal.pdf," 2016.

[210] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013.

[211] "Multinomial logistic regression," *UCLA Institute for Digital Research & Education, Online: https://stats.idre.ucla.edu/spss/output/multinomial-logistic-regression/*.

[212] B. Desgraupes, "Clustering Indices, ," 2017, Online: https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf. [Online]. Available: https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf.

[213] J. Yang, J. Zhao, F. Wen, and Z. Y. Dong, "A framework of customizing electricity retail prices," *IEEE Transactions on Power Systems,* vol. 33, no. 3, pp. 2415-2428, 2017.

[214] A. J. Conejo, M. Carrión, and J. M. Morales, *Decision making under uncertainty in electricity markets*. Springer.

[215] J. Dhaene, S. Vanduffel, M. J. Goovaerts, R. Kaas, Q. Tang, and D. Vyncke, "Risk measures and comonotonicity: a review," *Stochastic models,* vol. 22, no. 4, pp. 573-606, 2006.

[216] S. Ghavidel, A. Rajabi, M. J. Ghadi, A. Azizivahed, L. Li, and J. Zhang, "Risk-constrained demand response and wind energy systems integration to handle stochastic nature and wind power outage," *IET Energy Systems Integration,* 2019.

[217] M. H. Abbasi, M. Taki, A. Rajabi, L. Li, and J. Zhang, "Coordinated operation of electric vehicle charging and wind power generation as a virtual power plant: A multi-stage risk constrained approach," *Applied Energy,* vol. 239, pp. 1294-1307, 2019.

[218] M. Carrion, A. J. Conejo, and J. M. Arroyo, "Forward contracting and selling price determination for a retailer," *IEEE Transactions on Power Systems,* vol. 22, no. 4, pp. 2105-2114, 2007.

[219] S. Nojavan, K. Zare, and B. Mohammadi-Ivatloo, "Optimal stochastic energy management of retailer based on selling price determination under smart grid environment in the presence of demand response program," *Applied energy,* vol. 187, pp. 449-464, 2017.

[220] W. Wei, F. Liu, and S. Mei, "Energy pricing and dispatch for smart grid retailers under demand response and market price uncertainty," *IEEE transactions on smart grid,* vol. 6, no. 3, pp. 1364-1374, 2014.

[221] M. Zugno, J. M. Morales, P. Pinson, and H. Madsen, "A bilevel model for electricity retailers' participation in a demand response market environment," *Energy Economics,* vol. 36, pp. 182-197, 2013.

[222] S. A. Gabriel, A. J. Conejo, M. A. Plazas, and S. Balakrishnan, "Optimal price and quantity determination for retail electric power contracts," *IEEE Transactions on power systems,* vol. 21, no. 1, pp. 180-187, 2006.

[223] M. Carrión and J. M. Arroyo, "A computationally efficient mixed-integer linear formulation for the thermal unit commitment problem," *IEEE Transactions on power systems,* vol. 21, no. 3, pp. 1371-1378, 2006.

[224] L. Wu, "A tighter piecewise linear approximation of quadratic cost curves for unit commitment problems," *IEEE Transactions on Power Systems,* vol. 26, no. 4, pp. 2581-2583, 2011.

[225] T. P. Mian Khuram Ahsan, Zhengming Li, "The Linear Formulation of Thermal Unit Commitment Problem with Uncertainties through a Computational Mixed Integer," *Journal of Power and Energy Engineering,* vol. 6, pp. 1-15, 2018, doi: 10.4236/jpee.2018.66001

[226] "General Algebraic Modelling System (GAMS)." https://www.gams.com/ (accessed.

[227] "NEOS Server: State-of-the-Art Solvers for Numerical Optimization." https://neos-server.org/neos/ (accessed.

[228] J. Czyzyk, M. P. Mesnier, and J. J. Moré, "The NEOS server," *IEEE Computational Science and Engineering,* vol. 5, no. 3, pp. 68-75, 1998.

[229] "Spanish Electricity Market Data." https://www.esios.ree.es/en (accessed.

[230] "Pecan Street Dataport." https://www.pecanstreet.org/dataport/ (accessed.

[231] "Smart Grid Smart City (SGSC) Project Data." https://data.gov.au/data/dataset/smart-grid-smart-city-customer-trial-data (accessed July 2020.

[232] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, and J. Albrecht, "Smart*: An open data set and tools for enabling research in sustainable homes," 2012.

[233] "Ausgrid Solar Home Electricity Data." https://www.ausgrid.com.au/Industry/Our-Research/Data-to-share/Solar-home-electricity-data (accessed July 2020.

[234] M. Ghorbanian, S. H. Dolatabadi, and P. Siano, "Big data issues in smart grids: A survey," *IEEE Systems Journal,* vol. 13, no. 4, pp. 4158-4168, 2019.

[235] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Transactions on Smart Grid,* vol. 10, no. 3, pp. 3125-3148, 2018.

[236] "ESSnet Big Data." https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data (accessed July 2020).

[237] "Bits to Energy Lab Initiative." https://www.bitstoenergy.com/ (accessed July 2020).

[238] "CITIES Innovation Center." https://www.citiesinnovation.org/ (accessed July 2020).

[239] M. Molina-Solana, M. Ros, M. D. Ruiz, J. Gómez-Romero, and M. Martin-Bautista, "Data science for building energy management: A review," *Renewable and Sustainable Energy Reviews,* vol. 70, pp. 598-609, 2017.

[240] "Teradata: Siemens and Teradata form global strategic partnership for big data in the utility sector.," *Online: https://www.teradata.com/Press-Releases/2013/Siemens-and-Teradata-form-global-strategic-pa*.

[241] D. Laney, "3-D data management: Controlling data volume, velocity and variety.," *Technical report, Gartner,* 2001.

[242] J. Gantz and D. Reinsel, "Extracting value from chaos," 2011.

[243] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Networks and Applications,* vol. 19, no. 2, pp. 171-209, 2014.

[244] S. Bera, S. Misra, and J. J. Rodrigues, "Cloud computing applications for smart grid: A survey," *IEEE Transactions on Parallel and Distributed Systems,* vol. 26, no. 5, pp. 1477-1494, 2015.

[245] M. Yigit, V. C. Gungor, and S. Baktir, "Cloud computing for smart grid applications," *Computer Networks,* vol. 70, pp. 312-329, 2014.

[246] T.-c. Fu, "A review on time series data mining," *Engineering Applications of Artificial Intelligence,* vol. 24, no. 1, pp. 164-181, 2011.

[247] T. W. Liao, "Clustering of time series data—a survey," *Pattern recognition,* vol. 38, no. 11, pp. 1857-1874, 2005.

[248] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering–A decade review," *Information Systems,* vol. 53, pp. 16-38, 2015.

[249] A. Sardá-Espinosa, "Comparing time-series clustering algorithms in R using the dtwclust package," *Vienna: R Development Core Team (2017)*.

[250] T. Giorgino, "Computing and visualizing dynamic time warping alignments in R: the dtw package," *Journal of Statistical Software,* vol. 31.7, no. 7, pp. 1-24, 2009.

[251] M. Müller, *Information retrieval for music and motion*. Springer.

[252] P. Montero and J. A. Vilar, "TSclust: An R package for time series clustering," *Journal of Statistical Software,* vol. 62.1, pp. 1-43, 2014.

[253] I. Benítez, J.-L. Díez, A. Quijano, and I. Delgado, "Dynamic clustering of residential electricity consumption time series data based on Hausdorff distance," *Electric Power Systems Research,* vol. 140, pp. 517-526, 2016.

[254] I. Benítez, A. Quijano, J.-L. Díez, and I. Delgado, "Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers," *International Journal of Electrical Power & Energy Systems,* vol. 55, pp. 437-448, 2014.

[255] M. Sun, Y. Wang, F. Teng, Y. Ye, G. Strbac, and C. Kang, "Clustering-Based Residential Baseline Estimation: A Probabilistic Perspective," *IEEE Transactions on Smart Grid,* 2019.

[256] T. Jarábek, P. Laurinec, and M. Lucká, "Energy load forecast using S2S deep neural networks with k-Shape clustering," in *2017 IEEE 14th International Scientific*

Conference on Informatics, 14-16 Nov. 2017 2017, pp. 140-145, doi: 10.1109/INFORMATICS.2017.8327236.

[257] S. Ryu, H. Choi, H. Lee, H. Kim, and V. W. Wong, "Residential Load Profile Clustering via Deep Convolutional Autoencoder," in *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, 2018: IEEE, pp. 1-6.

[258] Y. Wang, Q. Chen, D. Gan, J. Yang, D. S. Kirschen, and C. Kang, "Deep learning-based socio-demographic information identification from smart meter data," *IEEE Transactions on Smart Grid,* vol. 10, no. 3, pp. 2593-2602, 2018.

[259] R. Lu and S. H. Hong, "Incentive-based demand response for smart grid with reinforcement learning and deep neural network," *Applied energy,* vol. 236, pp. 937-949, 2019.