Tampere University

Dilshan Subasinghe

GENERATING INDIVIDUAL ELECTRICITY LOAD PROFILES WITH THE TOP-DOWN ANALYSIS METHOD

Master of Science Thesis
Faculty of Information Technology and Communication Sciences
Examiners: Prof. Pertti Järventausta and
Dr.Tech. Antti Mutanen
August 2020

ABSTRACT

Dilshan Subasinghe: Generating individual electricity load profiles with the top-down analysis method
Master of Science Thesis, 64 pages, 1 Appendix page
Tampere University
Degree Programme in Electrical Engineering
August 2020

Simulations with realistic network models and electric loads are essential for developing smart grid integration strategies such as integrating distributed generation and electric vehicles into the grids. Accurate simulations require detailed information on the electricity consumption of customers connected to the grid. However, the electricity consumption data from individual customers are challenging to acquire because of data privacy concerns. Especially with the introduction of the new General Data Protection Regulation (GDPR) in the European Union, the electricity distribution system operators are not interested in sharing individual consumption data. Running detailed smart grid simulations requires individual customer load profiles and cannot be based on publicly available average load profiles such as national customer class load profiles. The averaged load profiles do not yield sufficiently accurate results because they do not reflect the temporal load variations present in actual consumption data.

The study material of this thesis will consist of new type consumer load profiles as a replacement for the Finnish customer class load profiles, and their previously calculated statistical properties by Dr.Tech. Antti Mutanen, and some thousands of real smart meter measurements. In this M.Sc. thesis, the goal is to study how those type consumer load profiles in the study material could be reverse-engineered into realistically varying individual synthetic load profiles using the top-town analysis method.

This thesis develops three algorithms for generating individual load profiles based on Markov chain process. The first algorithm uses the traditional Markov chain method to generate synthetic load profiles. Then, the traditional Markov chain method is extended to improve the results, and the new algorithm (i.e. second algorithm) is called the suggested Markov chain algorithm. The third algorithm in this thesis is called the adaptive Markov chain algorithm in the literature and borrows several machine learning concepts to develop it. Finally, an aggregate load profile matching method is described, implemented and applied to realistically adjust and scale the synthetic load profiles generated by the above algorithms. All the algorithms described in this thesis are implemented using MATLAB, and a part of the adaptive Markov chain algorithm is implemented using Python. The suggested Markov chain method, combined with the aggregate load profile matching method, allows generating realistic synthetic load profiles, and meets the goal of this thesis. The results are shown and validated in the final chapters, and they confirm that the suggested Markov chain method works properly for load profile generation and it can better capture the yearly seasonal variations in power consumption. The MATLAB programs are designed and implemented for hourly smart meter measurement input data. These programs can later be flexibly modified for higher-resolution input data and synthetic load profiles. Furthermore, the developed adaptive Markov chain algorithm can be further developed in the future with different deep learning techniques to get more realistic load profiles.

Keywords: Top-Down analysis method, stochastic load profile generation, time-inhomogeneous Markov chain, load profiling

PREFACE

This thesis is funded by the ADAMO project and is being examined by supervisor, Prof. Pertti Järventausta, and co-supervisor, Dr.Tech. Antti Mutanen in Tamper University. First and foremost, my heartfelt and sincere thanks to my supervisor, Prof. Pertti Järventausta, and co-supervisor, Dr.Tech. Antti Mutanen, for giving me the opportunity to complete my thesis. I appreciate their contributions of time and ideas to make my work productive and stimulating. Their valuable suggestions, comments and guidance encouraged me to learn more day by day. Their deep insights helped me at various stages of my research.

Further, I would like to thank Prof. Pertti Järventausta for his patient guidance, nice personality and offering me such a precious chance to work on this interesting and valuable topic that suits my interest in data analysis and mathematical modelling. It technically motivated me, and I enjoyed it a lot during the learning process. Also, I would like to thank Dr.Tech. Antti Mutanen for theoretical guidance, good suggestions and, helpful advice on every regular meeting. Moreover, his work is also a massive part for the fundamental of this thesis.

I would like to express my deep gratitude to my family members, especially parents, for continuous support with motivational thoughts. Finally, I would like to express my sincere appreciation to those who have contributed to this thesis and supported me in one way or another during this incredible journey.

18th August 2020

Dilshan Subasinghe

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

AMR	Automatic Meter Reading
CDF	Cumulative Density Function
CTM	Cumulative Transition Matrices
DSO	Distribution System Operator
FOMC	First-order Markov Chain
GDPR	*European Union* General Data Protection Regulation
GMM	Gaussian Mixture Model
MC	Markov Chain
MCMC	Markov Chain Monte Carlo Simulation
MAPE	Mean Absolute Percentage Error
PDF	Probability Density Function
SM	Smart Meter
SOMC	Second-order Markov Chain
TPM	Transition Probability Matrix

$a(t)$	load temperature dependency parameter
$E[T(t)]$	expected value of the outdoor temperature
F	cumulative transition matrix
h_θ	hypothesis, estimated probabilities of classes
$J(\theta)$	cost function of multinomial logistic regression
K	the total number of classes of multinomial regression model
m	length of the training data set
n_{ij}	the number of transitions from state i to state j
p_{ij}	ij^{th}element of transition probability matrix
$P()$	probability
P	transition probability matrix
s	element of state space
S	state space
t	time in hours
$T_{24}(t)$	the average outdoor temperature from the previous 24 hours
V	the maximum power
x	input features vector
X	the input explanatory variable matrix
X_t	process at time t
y	response variable
φ	softmax function, probability of a class
θ	coefficients of multinomial regression model
$\nabla_{\theta_k}J(\theta)$	gradient for cost function of multinomial regression
β	vector of multiple linear regression coefficients
ϵ	the vector of residuals of multiple linear regression

.

1. INTRODUCTION

Most of the existing traditional power grids around the world were built several decades ago and the power system has well served during that time period. However, recently, the traditional power systems are subjected to regulations by many national governments due to experiencing technical, economic and environmental issues. The modern society evolves this old power system infrastructure to be more reliable, manageable and scalable, while also being secure, cost-effective and interoperable [1]. Such a next-generation power system is called a “smart grid”. Smart grids mean more than energy generation and transmission, and the concepts behind modern electricity grids are also smarter. The flexibility of smart grids has been improved with the use of novel control techniques, ICT technologies, and equipment with two-way communication compatibility between customers and utilities. The power system reliability has been increased significantly by reducing the number of outages and system restoration times have been reduced with fault location, isolation, and service restoration applications. The development of smart grids has increased the research and development of smart metering. A smart meter can be considered as a gateway for two-way communication between customers and energy system’s parties. A next-generation smart meter can measure the energy consumption of the customers in real-time and transmit the data to distribution system operators (DSOs). Therefore, DSOs can manage and coordinate the flexibility of the grid, planning and operation of the network, and promote the energy efficiency with reflective tariff plans. In smart metering, the term "real-time" refers to a time resolution between 5-60 minutes. According to the Finnish energy regulator, over 99 % of premises are equipped with smart meters in Finland [19]. However, the current Finnish smart metering infrastructure is only partially capable of real-time operation. The next generation of smart meters will be updated at a later phase to facilitate real-time data operations, thus shortening the measurement time from 1 hour to 15 - 5 minutes and making the data immediately available. Even now, there is an ongoing project in Finland and testing new generation smart meters in 30 000 households to provide greater flexibility to the electricity grid [24].

There are important requirements for gathering and utilization of these consumption data via smart meters for also different other parties than DSOs. For instance, the power system within the EU electricity market should be able to withstand the intermittent nature of increasing renewable electricity generation. This can be achieved by activating the demand side to enable a more flexible balance between demand and supply. In practice, the smart meters are well-illustrated with initiatives and technology solutions to enable the demand side, thus collecting and communicating information on electricity consumption. As well as, when developing the smart grid operations such as integrating distributed energy resources and new types of loads (e.g. electric vehicles, smart buildings, power electronic equipment etc) into the grids, it is necessary to have more realistic network simulations and load models for smart grid operations. However, there is a difficulty in acquiring customer consumption data from DSO for any of the above mentioned or other purposes for different parties, because of this smart meter measurement data is protected due to privacy and data protection concerns. These privacy and data protection concerns came to effect with the introduction of the European Union General Data Protection Regulation (GDPR) which applies for processing the customer information, collection and utilization of smart meter data [17]. Therefore, DSOs are prevented from sharing individual consumption data of a customer for other parties without the customer's consent and this is the root cause for research question in this thesis.

Under this situation, the requirement for generating more realistically varying synthetic load profiles is raised for different purposes. It is not accurate enough to run a detailed smart grid simulation with average load profiles that are publicly available (e.g. national customer class load profiles), because the average load profiles do not clearly show the dynamic load variations in real-time consumption data of the customers. An average load profile can be obtained by dividing the aggregated load profile with the number of customers of a specific customer class which may lose important features of the load profile such as information on load factors, peak powers of the customers, etc. In this M.Sc. thesis, the goal is to study how derived customer class load profiles (i.e. called "type consumer load profiles") by Mutanen *et al.* could be reverse engineered into realistically varying individual load profiles [5]. The study material includes type consumer load profiles, their previously calculated statistical properties, and some thousands of smart meter measurements from the customers located in a specific area in Finland, 2016. The summary of the research questions and research work can be depicted clearly as in Figure 1.1.

The solution for the above privacy concerns is to generate synthetic load profiles by using load profile generator algorithms. The use of two different algorithms to generate synthetic load profiles for different customer classes can be found in the literature (i.e. bottom-up and top-down approaches). The bottom-up approach begins with each household appliance and models household characteristics, single customer behaviours and activity levels, and then builds up the load profile. For this, in order to study the used household appliances and time of use of electricity by customers, the algorithm requires a considerably high amount of measurement data of appliances as inputs [8]. Therefore, the low availability of data leads to a poor outcome. This is a drawback in the bottom-up algorithm. In contrast, the top-down approach is a quite different load profile generating algorithm which uses existing smart meter measurement data to generate more realistic load profiles for each household using the same statistical properties of available type consumer load profiles. A top-down approach will require less computational effort compared to a bottom-up approach. In this thesis, the aim is to present a top-down model to generate synthetic load profiles using a traditional Markov model for any number of customers based on the data set provided as a study material.

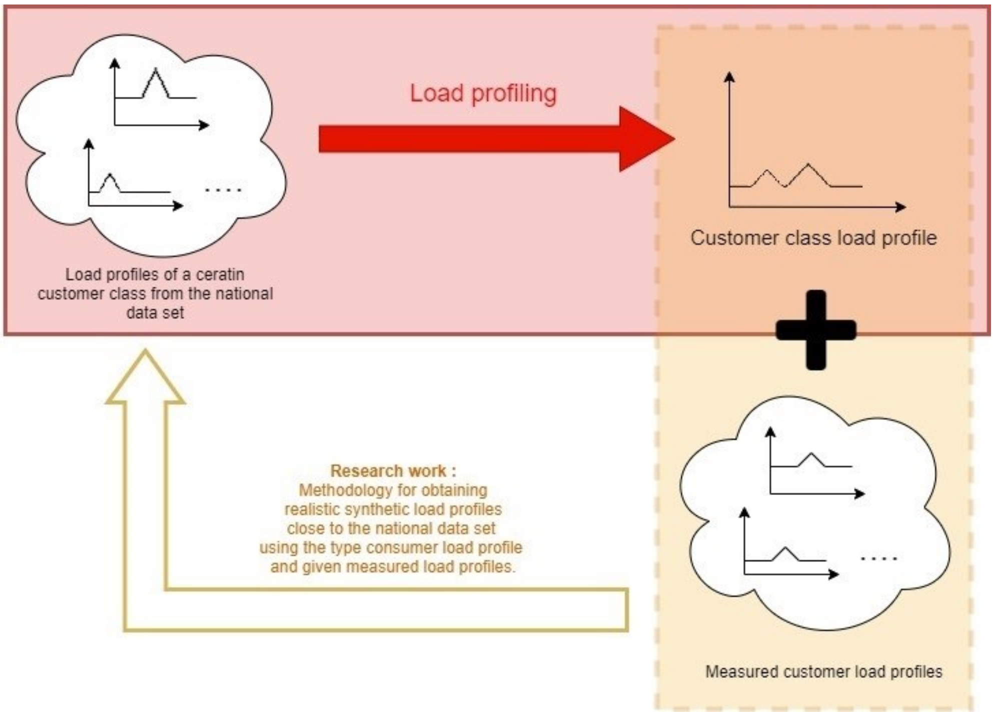

Load profiling
Load profiles of a ceratin customer class from the national data set
Customer class load profile
Research work :
Methodology for obtaining realistic synthetic load profiles close to the national data set using the type consumer load profile and given measured load profiles.
Measured customer load profiles

Figure 1.1 *An overview of the research work. The customer class load profiles and several smart meter measurement data are available from the study material as input. Forming of customer class load profiles is called "load profiling"*

In the literature, some research works can be found which has been done already by using bottom-up [8][16][18][22] and top-down [8][13][23][25] algorithms to generate synthetic load profiles. McLoughlin *et al.* have built a homogeneous Markov chain model with the top-down approach to generate domestic load profiles [13]. The outcomes show satisfactory results for key statistical properties such as mean, standard deviation between measurement and synthetic load profiles. But the Markov chain failed to catch the temporal variations in the input load profiles; it was utterly random. *Bucher et al.* present a combination of both bottom-up and top-down load models [8]. They have built a methodology for generating synthetic load profiles from the top-down approach based on statistical analysis of either measurement data or artificially generated load data from the bottom-up approach. The results of this research show that the top-down synthetic load data exactly corresponded to the statistical properties of the bottom-up synthetic load data. *Labeeuw et al.* present a good approach for this thesis work with inhomogeneous Markov models and the clustering of customer data [25]. They have proposed a Markov chain process for tracking daily behaviour and a Markov decision-making process for spreading the behaviour changes on other days of the week.

The rest of the chapters of this thesis are structured as follows. First, chapter 2 presents the background study for customer load profiles and synthetic load profile generation. Thereafter, chapter 3 presents an overview of the theories and definitions used in the research methodology. chapter 4 describes the data set used for this study. Later, Chapter 5 presents three algorithms for synthetic load profile generation based on Markov chain (MC). The three algorithms include a conventional MC and an adaptive MC described in the literature, as well as a suggested new methodology. In that same chapter, the outputs of the three synthetic load profile generators are compared to each other. This is followed by chapter 6 which evaluates the output of the best load profile generator obtained from the comparison in chapter 5. Finally, the conclusions drawn from the research are presented in chapter 7.

2. BACKGROUND STUDY FOR CUSTOMER LOAD PROFILES AND SYNTHETIC LOAD PROFILE GENERATION

This chapter includes relevant information from background studies on customer load profiles and synthetic load profile generation. Accordingly, some subtopics such as electricity consumption, the existing smart metering, the customer class load profiles in Finland and factors affecting customer load profiles are discussed under this chapter. Furthermore, this chapter highlights the previous research activities at Tampere University and uses them as background material for this thesis.

2.1 Electricity consumption in Finland

Today, power distribution and retail companies are increasingly focusing on collecting customer energy consumption data and analyzing the load profiles regularly to learn how the load demand is varying. This thesis continues the study with Finnish electricity consumption data in a specific area. Therefore, it is meaningful to get an overall idea of the present electricity consumption in Finland before moving on to the next chapters.

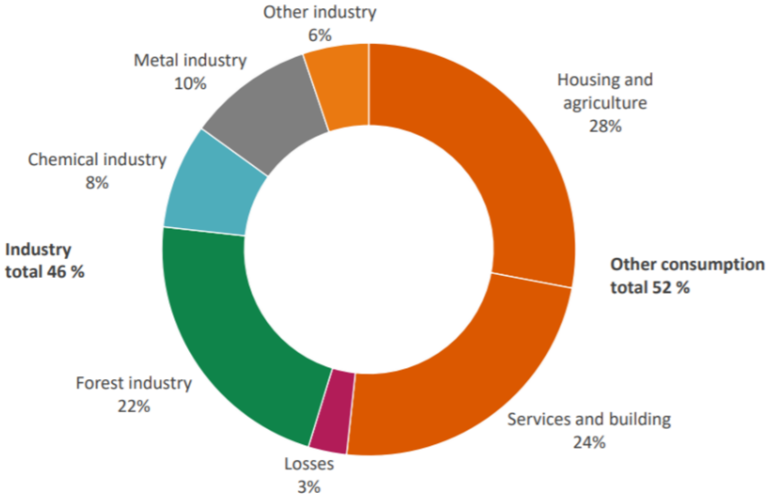


Figure 2.1 *Finnish electricity consumption in different sectors in 2019, 86 TWh [12]*

Figure 2.1 shows the electricity consumption in various consumer sectors in Finland in 2019 as shares of total electricity consumption. The "household and agriculture" sector accounts for a significant proportion of overall electricity consumption (i.e., 28 %), while the "services and building" sector has acquired the second-largest electricity consumption. The household electricity consumption sector takes a significant proportion of overall consumption in most of the other European countries as well [13]. Due to specific geographical latitudes in Finland, its climate is mainly characterized by many cold days. For this reason, heating systems mainly impact on the electricity consumption of consumers. Therefore, heating solutions are playing a vital role in the Finnish electricity consumption shares. For instance, Figure 2.2 illustrates that around 78% of household energy was allocated for heating in 2018. The share of electricity consumption in households accounted for about 33% of the total household energy consumption. From that electricity consumption, the shares of electricity consumption for indoor heating and household appliances are respectively 47% and 36 % [20]. The consumers can choose heating methods freely in the Finnish heating market. The available heating methods are among district heating, electrical heating and other site-specific solutions with heat pumps and different energy sources. Apart from that, consumers can also purchase cooling solutions based on district cooling, heat pumps and other electrical-based equipment. So, the Finnish heating market is quite competitive. Also, the heating market is entirely unregulated [14]. Therefore, there is no legislation regarding the selection or pricing for the heating and cooling methods. Due to these facts, consumers today tend to switch their heating systems from low to high-efficiency systems.

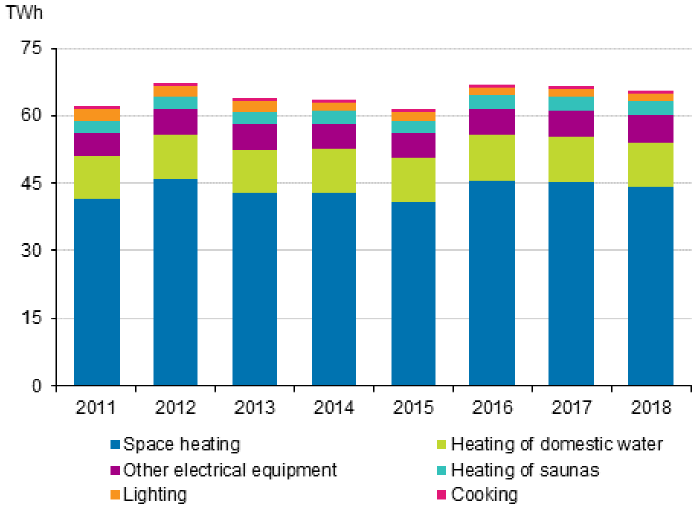
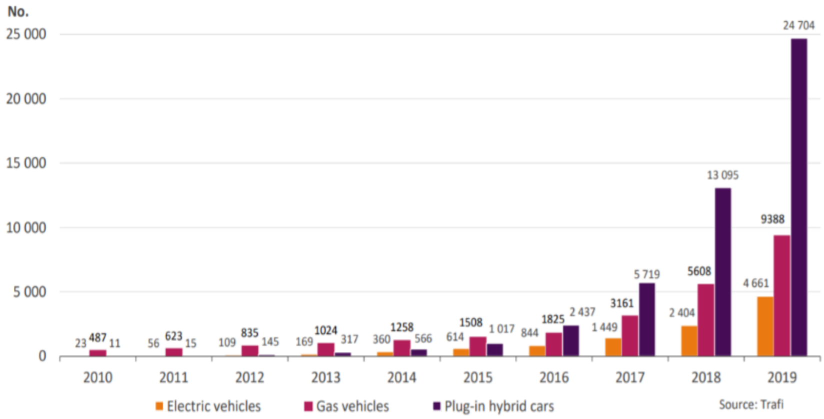


Figure 2.2 *Energy consumption in households 2011-2018 [11]*

The total share of industrial electricity consumption in 2016 is around 46 %, and forest, chemical and metal industries are the primary consumers in that category. The hourly power consumption values of extensive consumers such as industrial customers are comparatively higher than other consumers. Moreover, it appears that new types of loads will continue to be added to the Finnish power system, such as electric vehicles, heat pumps and modern electronic equipment. For instance, Figure 2.3 shows that the number of electric and plug-in vehicles has increased significantly in 2019, which could impact on the annual energy consumption level of the customer classes and load behaviours of customers. Figure 2.4 shows that the number of installed heat pumps is increasing every year, and most of them are air-air type heat pumps that are usually used to supplement the direct electric heating. According to Finnish Heat Pump Association, the number of heat pumps sold in 2019 has increased by 30% from the previous year. Therefore, adding new types of loads to the Finnish power system must be properly modeled, and customer class load profile updating, and customer classification must be done accordingly.



Source: Trafi

Figure 2.3 *The number of electric vehicles, gas vehicles and plug-in hybrid cars in passenger vehicle stock, 2010-2019 [12]*

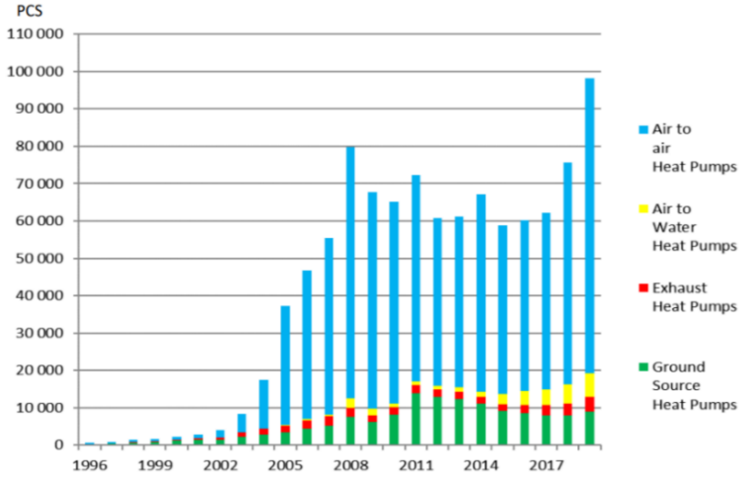


Figure 2.4 *Annual heat pump installations in Finland [15]*

2.2 Electricity metering in Finland

Electricity meter reading is an essential component in energy market-related functions as well for distribution network calculations. In the past, mechanical electric meters were used to measure the electricity usage of customers. The meter readings of those mechanical meters were done by DSO’s meter readers at customer’s premises. In those days, meter readings were obtained infrequently (e.g., like once a year), and balancing bills were prepared for customers once the meter readings were done (e.g. once a year). However, with the improvement of the technology, Automatic Meter Reading (AMR) systems are introduced to collect more important data from customers such as electricity consumption, diagnosis and status via one-way communication mediums. The collected data can be transferred to a central database for further analysis, troubleshooting and billing processes. This AMR system is a digital implementation of a pre-mechanical analogue meter, so it replaces the mechanical induction disk and provides better resolution data readings [10]. One of the main advantages of the AMR system compared to an analogue meter is that it reduces the number of DSO employee site visits. Thus, bills can be adjusted based on actual consumption rather than estimated consumption of energy. AMR systems have been evolved over time, and today it is a trendy research area. With that, different advanced functionalities have been added to the AMR systems, so that naming used by different groups has been changed from AMR to smart meters [10]. In Finland, smart meters have been installed in over 99 % of premises [19]. Therefore, it is possible to get more up to date consumption data and use them for different activities for permitted parties. The next generation metering techniques with bidirectional communications are called advanced metering. Advanced metering systems include all the functionalities of AMR, but AMR may not include all the advanced metering functionalities. Advanced metering systems can be used to collect information with different resolutions [4]. However, in this thesis, hourly measurement data are used in the data set.

2.3 Customer class load profiles in Finland

Customer class load profiles can be used to explain the aggregated behaviour of the customers in different customer classes such as household, commercial and industry etc. The load profiling term is used to form such a customer class load profile for different customer classes. Before introducing AMR systems, load profiling was done by measuring a sample of customers, classifying them by the type of electricity use, and generalizing the generated results to cover the other customers in the same type. The bottom-up approach also has been used as an alternative method for this [4]. In different years, several customer class load profiles were published in Finland. For instance, in the

1980s, large scale cooperation in load research was started by Finnish utilities. During this project, over 1000 customers' hourly measurement data were collected, and after that first measurement period (i.e., 1983 -1985), 18 customer class load profiles were published in 1986 [4]. Then, another set of measurements were collected in the following measurement period (i.e., 1986 - 1988), and one more set of customer class load profiles was published. The total number of customer load profiles that were published at the end of the above mentioned first and second measurement periods was 46. These customer class load profiles mainly belong to classes such as housing, agriculture, industrial, commercial and administration. Each of these classes is further divided into subclasses according to different consumption patterns that describe its features such as the type of heating solution or building type. In Finland, these load profiles are the only publicly available comprehensive set of customer class load profiles [4]. After that, several other new customer class load profiles were published, but those are only used by 15 companies who participated in that project. And also, some companies have built individual customer class load profiles for some of their large power consumers.

2.4 Significant factors affecting customer load profiles

Load analysis is beneficial for finding the factors affecting the power consumption of different customers. The overall general load behaviour of customers in a customer class can be predicted by observing its customer class load profile and customer type of the class. For that, Figure 2.5 shows average load profiles on Mondays of the first week of the four seasons of 2016 for customers living in energy-efficient detached houses with electric heating. The average load profiles shown in the figure are taken from the data set used for this thesis. More details on this data set are covered in Chapter 4. This figure represents only the average load profiles, and individual customer load profiles within the customer class contain more details such as load fluctuations and peaks.

As seen in Figure 2.5, the power consumption varies throughout the day. The power consumption between time 00:00h and 05:00h is significantly lower compared to other hours of the day because residents usually use this time interval to sleep; thus, activity levels are minimum. The initial rise in daily residential power consumption in the average load profile can be observed from around 05:00h to 08:00h, because the residents in the houses wake up in the morning and get ready for work. Then the level of power consumption becomes either slightly stable or reduce until 15:00h with a small slope. The reason for this observation is that the daytime activities are limited in detached houses, and one or several persons of the family may have left for work or school. However, electricity consumption starts to rise again during the afternoon and the highest peak can be observed in the evening due to after work entertainment, cooking and dining activities. This average load profile represents a specific customer class, and its consumption pattern varies depending on the common activity type of the customer class on different days of the week. For instance, the presence of a resident in a detached house could be considered as comparatively high on weekend days during the daytime, so that activity level will also be different on weekend days. Thus, the average power consumption during the daytime on a weekend day can be slightly higher than the typical weekday. In contrast, the evening, the power consumption on a weekday can be higher than a day on the weekend [16]. Hence, customer behaviour and residence characteristics can cause fluctuations in the power consumption of load profiles throughout the day.

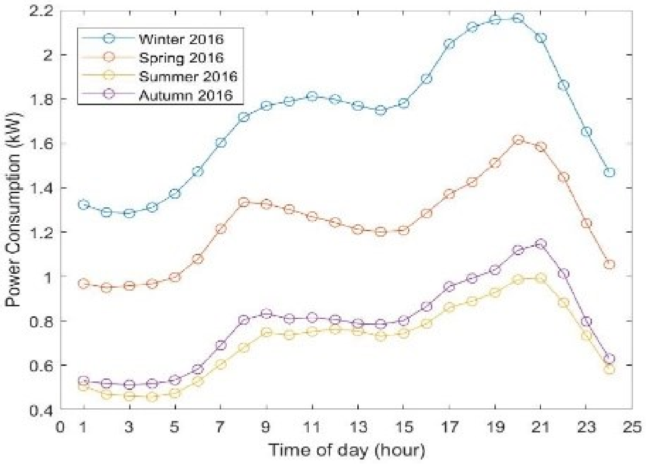


***Figure 2.5** Average load profiles on Mondays of the first week for the 4 seasons in 2016 for a class of customers who live in energy-efficient detached houses with electric heating in a specific area, Finland*

The annual power consumption of a consumer also depends on different other factors such as the number of occupants of a dwelling, geographical location and weather factors etc. Although not applicable to Finland, the geographical location can affect rural and urban areas in other countries. The lifestyles of people from rural areas are quite simple, and necessities are lower than urban living people. Moreover, urban houses are equipped with modern type of electrical equipment such as cookers, heaters and other appliances. Therefore, residents in rural areas consume less electricity during a day compared to residents in urban houses. Furthermore, geographical location may also cause climate factors that typically occur quite identically in successive years such as temperature, humidity and daily light hours. Notably, people in countries close to the equator have to cope with hot climate conditions to make their lives comfortable; in contrast, countries close to poles must cope with cold conditions. So that heating and cooling solutions must be used. These seasonal variations of power consumption due to heating energy consumption can be clearly seen in Figure 2.5. Furthermore, daylight variation throughout the day can cause fluctuations of hourly power consumption due to different uses of domestic appliances (e.g., lighting loads). The electricity consumption of customers also can be affected slightly by wind speed and direction, but the effect is comparatively small [4]. Likewise, above-discussed factors could affect load profiles, and they are essential to predict and exhibit the behaviour of the customers.

2.5 Relevant previous research activities in Tampere University

Since the customer class load profiles described in subchapter 2.3 are more than 28 years old today, those load profiles can be outdated in the current power system. Today, consumption patterns of customers have changed considerably with the competitive heating market and introduction of new types of loads as described in subchapter 2.1 such as electric, plug-in hybrid vehicles, heat pumps. Antti Mutanen has presented some defect fixing methods in existing load profiles in his doctoral thesis [4]. In that PhD thesis, some possible further improvements to increase the accuracy of customer class load profiles have been discussed by using methods such as temperature dependency, customer classification and customer behaviour change detection. The different clustering algorithms like K-Means, ISODATA, GMM can be used for customer classification, and they provide a good basis for analyzing the behaviour of customers further [3][6][7]. It indicates that clustering improves the accuracy of the load profiles, and it can be used to update both customer class and its load profiles simultaneously. Therefore, clustering can be done periodically in order to improve the accuracy of load profiles. Furthermore, Mutanen et al. have defined 14 type consumer classes based on consumer’s activity, fuse size, and average annual energy consumption as an alternative to the Finnish national customer classes described in subchapter 2.3 [5]. The outcomes from these previous research activities are useful for this thesis and are also used as supportive materials for the synthetic load profile generation. Therefore, in this thesis, type consumer classes are used and introduced in chapter 4.

2.6 Synthetic load profile generation and associated theories

Many smart grid simulations such as in distributed power generation simulations and renewable energy simulations require customer electricity load profiles frequently. However, the smart meter data of the customers may not be readily available for other parties due to GDPR, as explained in the introduction of this thesis. Therefore, one solution to this load profile requirement for smart grid simulations is to represent the consumption with its customer class load profile. Such a general load profile might be accurate enough depending on the objective of the simulation. But the customer class load profile does not reflect the load behaviour of each customer in a specific power distribution area. For comparison purposes, Figure 2.6 shows a load profile of a customer and the corresponding customer's customer class load profile from the data set. According to Figure 2.6, the customer class load profile hides a lot of essential details and features of the actual customer load profile. The literature shows that customer load profiles can be generated to overcome this problem by using stochastic processes to represent the consumption at each time step, which is known as synthetic load profile generation [8][13][16][18][22] [23][25]. In this thesis, a traditional, a new approach, and an adaptive methodology for generating more realistic synthetic load profiles by using the top-down analysis method will be presented with observations and analysis. The Markov chain related definitions with different statistics theories can be applied to build a synthetic load profile generator. Furthermore, machine learning concepts can be used to optimize the output of the load profiles to get closer to desired results (e.g. multinomial logistic/multiple linear regression). The background of used theories from the above areas will be discussed in the next chapter (i.e. chapter 3).

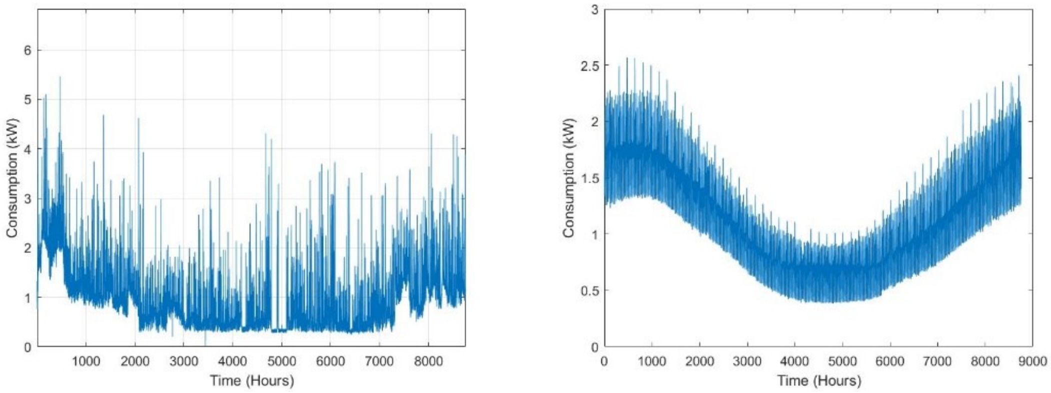


Figure 2.6 *An individual customer load profile (left) and the corresponding customer class load profile (right)*

3. DEFINITIONS AND THEORIES

This chapter presents the definitions and theories used in this research, and they serve as a guide for the methods and analysis presented in the following chapters.

3.1 Markov chain

3.1.1 Definition of Markov chain

Markov chain (MC) is named after the Russian mathematician A. A. Markov (1856 - 1922), who is known for his work on number and probability theories. MC is an important mathematical tool for stochastic processes and gives random outcomes. MCs are often used to study temporal and sequential data. A stochastic process is a mathematical model that evolves in a probabilistic way over time. The underlying idea of MC is to simplify some predictions of the stochastic process. The present state of a stochastic process depends only on the previous states. As explained later in subchapter 3.1.4, the number of previous states in the process considered for the MC depends on the degree of order of the MC. For example, for a first-order MC, the next state of the process depends only on the present state, not the previous states. The following definitions are given for the first-order MC.

Definitions:

Consider a MC with the process $X_0, X_1, X_2, \dots \dots, X_t$ for the following definitions.

1. The state of a MC at time t is the value of X_t.

   e.g. if X_t = 1, the process is at state 1 when time is t

2. The state-space of a MC (i.e. denoted as S) is the set of all existing states. The size of S is a finite value.

   e.g. $S = \{1, 2, 3, 4, 5\}$

3. A trajectory of a MC is a set of specified values for $X_0, X_1, X_2, \dots$

   e.g.: if X_0 = 1, X_1 = 3, and X_2 = 5, then the trajectory from t = 0 to t = 2 is given as 1, 3, 5

As explained earlier, the basic property of a MC is that only the state of the latest time step in a trajectory affects to the next time step (i.e. for first-order MC). This Markov property can be formulated in mathematical notation as in (3.1), where X_{t+1} depends on X_t, and it does not depend on $X_{t-1}, \ldots . . X_1$ or X_0.

$$P(X_{t+1} = s_{t+1} \mid X_t = s_t,\, X_{t-1} = s_{t-1}, \ldots \ldots . X_0 = s_0) = P(X_{t+1} = s_{t+1} \mid X_t = s_t) \qquad (3.1)$$

where $s_0, \ldots . s_t$ represent the states respectively when time is 0 to t [9].

Definition: Let a sequence of discrete random variables be $\{X_0, X_1, X_2, \ldots \ldots, X_t\}$ and this sequence is said to be a MC if is follows the Markov property defined in (3.1).

3.1.2 Determining the states

A state-space can be selected in different ways for a process. In the literature, three approaches to defining a state-space can be found. In the context of synthetic load profile generation, a state represents an interval of power consumption values. One approach to determining state-space is to divide the total range of possible values in the data into equal length-segments. However, due to the lack of data distribution between states, this may lead to states with few transitions and improper modeling.

e.g. Let us consider a dataset D = $\{a_0, a_1 \ldots\}$),

Where the length of each interval = $\frac{\max(D) - \min(D)}{number\ of\ states}$

Alternatively, there is another approach in the literature that can define the limits of states by splitting the cumulative density function. Therefore, states are defined as having the same number of transitions for each state. Figure 3.1 shows the above-mentioned procedure for a state-space with 10 states.

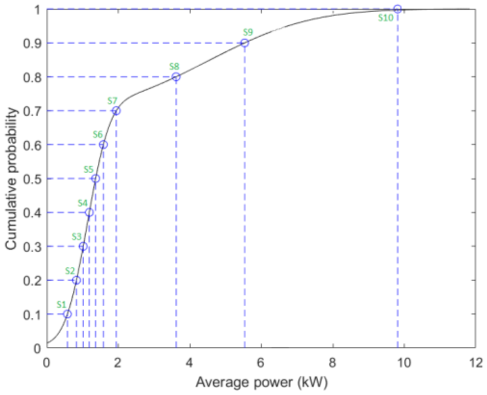


Figure 3.1 *Dividing the cumulative density function into 10 equal divisions in order to define a state-space with 10 states*

Moreover, another method used to define the states of an MC application used in wind speed modeling can be found in the literature [2]. This method is also valid for use in the application for generating synthetic load profiles, because it can be used for any data set independently of the application. First, the mean value (μ) and standard deviation (σ) are determined using the probability distribution of the dataset. Then, the states are defined using the divisions of μ and σ as shown in Figure 3.2.

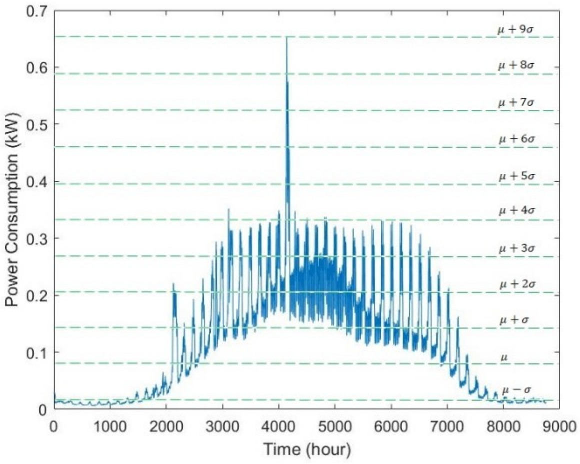


Figure 3.2 *Defining the Markov chain states using* $\boldsymbol{\mu}$ *and* $\boldsymbol{\sigma}$

3.1.3 Transition probability matrix

A transition diagram can be used to show transitions between states in each time step, and the diagram also can be summarized in a matrix. The matrix describing MC is called the Transition Probability Matrix (TPM) and is an important tool in MC analysis.

Let P be a transition matrix of a MC process, and p_{ij} represents the element in i^{th} row and j^{th} column of P. Each element of P satisfies the following features at time t.

1. **Rows** of P represent **now**, or **from** (X_t) state;
2. **Columns** of P represent **next**, or **to** (X_{t+1}) state;
3. The conditional probability for **next** = j, when **now** = i, which also means the probability of moving **from** state i **to** state j is given by the element p_{ij} of P.

$$p_{ij} = P(X_{t+1} = j \mid X_t = i) \tag{3.2}$$

The transition probability matrix (P) must represent all states in the state-space S. Let the size of S be N, therefore, P becomes a square matrix with a dimension of $N\ x\ N$ for first-order MC. The sum of all the probabilities in each row of P is equal to 1. For example, for the i^{th} row,

$$\sum_{j=1}^{N} p_{ij} = \sum_{j=1}^{N} P(X_{t+1} = j \mid X_t = i) = 1 \tag{3.3}$$

A MC is called a **homogeneous** if its transition probabilities p_{ij} are independent of time. That means, the transitions follow the same pattern without matter of when it started. In contrast, a **non-homogeneous** MC has transition probabilities with functions of time. In this thesis, the power consumption values of different hours is predicted by using previous power states. Thus, non-homogeneous MCs will be used.

Definition If a state s_t of a MC cannot leave from that state, it is called as an absorbing state (i.e. $p_{ii} = 1\,, p_{ij} = 0 \;\; where\; i,j \in S\; and\; i \neq j$). Therefore, once the outcome is reached to an absorbing state, it is impossible to make a transition to another state [9].

3.1.4 Constructing the transition probability matrix

A MC can be characterized based on its degree of orders. In a first order MC, the probability of a transition to a state at time t depends only on the immediately preceding state at $t-1$ as mentioned earlier. Similarly, second or higher orders MCs are processes that the current state depends on two or more preceding states. With the symbols used in subchapter 3.1.3, the transition probability matrix for a first-order MC can be presented as below.

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{bmatrix} \tag{3.4}$$

If n_{ij} is the number of transitions from state i to state j in the sequences in the data set, the transition probabilities can be estimated using the expression in (3.5), because the summation of probabilities of a row in transition matrix is equal to 1 as shown in (3.3).

$$p_{ij} = \frac{n_{ij}}{\sum_{k=1}^{N} n_{ik}} \tag{3.5}$$

By using (3.5), the homogeneous transition matrix can be constructed from the relative frequencies (i.e. from state i to state j) in the sequences. In contrast, the non-homogeneous transition matrix can be estimated for each time t by considering the relative frequencies at time t in the sequences. A second-order MC can be illustrated symbolically as below.

$$P = \begin{bmatrix} p_{111} & p_{112} & \cdots & p_{11N} \\ p_{121} & p_{122} & \cdots & p_{12N} \\ \vdots & \vdots & \vdots & \vdots \\ p_{1N1} & p_{1N2} & \cdots & p_{1NN} \\ p_{211} & p_{212} & \cdots & p_{21N} \\ p_{221} & p_{222} & \cdots & p_{22N} \\ \vdots & \vdots & \vdots & \vdots \\ p_{NN1} & p_{NN2} & \cdots & p_{NNN} \end{bmatrix} \tag{3.6}$$

In a second-order transition matrix, the transition probability p_{ijk} represents the probability of the next state k if the preceding states were i and j respectively. It can be seen that the sum of the probabilities of each row is also equal to 1 as in (3.3) for a higher-order MC transition matrix.

A high number of states is better for a detailed MC model because it can capture more precise variations of the random process. When the number of states is increased, the size of the transition matrix is also increased because the size of the matrix depends on the number of states as explained earlier in the same subchapter. Moreover, the available data will be distributed across the states when the number of states is increased. Therefore, a higher number of states might lead to an over-fitting model because there is less data available to compute the probabilities for transitions using (3.5). Furthermore, a higher-order MC contains more transitions, as clearly seen from the sizes of the transition matrices illustrated above for the first and second-order MCs in the same subchapter. Therefore, a higher-order MC significantly reduces the amount of data available to calculate the probability of each transition in (3.5), because the available data are distributed among the transitions [25].

3.2 Multinomial logistic regression

A logistic regression model can be used to classify observations into one of two classes. In case of more than two classes, multinomial logistic regression is used, and it is also called softmax regression. In this thesis, the adaptive MC in the literature is developed with multinomial logistic regression. In other words, the transition matrix of the adaptive MC is built based on time-related inputs and classification of power states. An explanation of how to use this section in a real MC application is described in subchapter 5.6. A multinomial logistic regression has a target y which ranges more than two classes. Therefore, the training data set used for multinomial logistic regression forms

$\{(x^{(1)},y^{(1)}),\ldots\ldots,(x^{(m)},y^{(m)})\}$ for m observations, where $x^{(i)} \in \mathcal{R}^{n+1}$ is the number of input features, $y^{(k)} \in \{1,\ldots,K\}$, and K is the number of classes.

In multinomial logistic regression, the probability of y being each class k (i.e. $P(y = k|x)$) need to be estimated for a given input features set. For that, the generalization of sigmoid, which is called the softmax function, can be used as in (3.7).

$$\varphi_i = \frac{e^{\eta_i}}{\sum_{j=1}^{K} e^{\eta_j}} \tag{3.7}$$

$$\eta_i = \theta_i^T x \tag{3.8}$$

where $\mathrm{i} \in \{1,\ldots,\mathrm{K}\}$, x is the input features vector (i.e. $x = x^{(i)} = (1, x_1, x_2 \ldots\ldots.x_n), x \in \mathcal{R}^{n+1}$) and θ is coefficients of the model (i.e. $\theta = (\theta_1, \theta_2 \ldots\ldots.\theta_K)$, $\theta_i \in \mathcal{R}^{n+1}$). Therefore, hypothesis (i.e. $h_\theta(x)$) gives the estimated probabilities of K number of classes for a given input features set (i.e. $x = x^{(i)}$) as below.

$$h_\theta(x) = \begin{bmatrix} P(y = 1|x) \\ P(y = 2|x) \\ \vdots \\ P(y = K|x) \end{bmatrix} = \frac{1}{\sum_{j=1}^{K} e^{\theta_j^T x}} \begin{bmatrix} e^{\theta_1^T x} \\ e^{\theta_2^T x} \\ \vdots \\ e^{\theta_K^T x} \end{bmatrix} \tag{3.9}$$

The $h_\theta(x)$ is $Kx1$ dimensioned vector and θ is a $Kx(n+1)$ dimensioned matrix. Note that $\frac{1}{\sum_{j=1}^{K} e^{\theta_j^T x}}$ in (3.9) normalizes the distribution and therefore, sum of the elements in $h_\theta(x)$ equals to 1. The multinomial logistic regression has the following cost function.

$$J(\theta) = -[\sum_{i=1}^{m}\sum_{k=1}^{K} 1\{y^{(i)} = k\} \log \frac{e^{(\theta_k^T x^{(i)})}}{\sum_{j=1}^{K} e^{\theta_j^T x^{(i)}}}] \tag{3.10}$$

where m is the length of the training data set. The function 1{} which is called “indicator function”, evaluates 1 or 0 if the condition in the brackets is true or false respectively (i.e. $1\{a\ true\ statement\} = 1, 1\{a\ false\ statement\} = 0$).

The objective is to obtain the minimum value of the cost function to find the coefficients of the model. But analytically the minimum cost function cannot be solved. Thus, an iterative optimization algorithm can be used to find the coefficient. For that, the formula for gradient is obtained as in (3.11) by taking the derivatives of (3.10).

$$\nabla_{\theta_k} J(\theta) = -\sum_{i=1}^{m} x^{(i)}[1\{y^{(i)} = k\} - \log \frac{e^{(\theta_k^T x^{(i)})}}{\sum_{j=1}^{K} e^{\theta_j^T x^{(i)}}}] \tag{3.11}$$

$\nabla_{\theta_k}\mathrm{J}(\theta)$ is a vector and its l^{th}element is the partial derivative of $\mathrm{J}(\theta)$ with respect to the l^{th} element of θ_k. Likewise, the minimum of $\mathrm{J}(\theta)$ can be calculated by using a standard optimization package and the gradient function [21].

4. AVAILABLE DATA FOR THE STUDY

This chapter describes the content of the study material used in this study. *Mutanen et al.* have defined 14 type consumer classes based on consumer's activity, fuse size, and average annual energy consumption as a replacement for the Finnish customer classes described in subchapter 2.3 [5]. This study material uses those defined type consumer classes and they are presented in Table 4.1. The study material contains a smart meter measurement data set (i.e. referred as "measured data set" in the next subchapters) from customers located in a specific area of Finland, and previously calculated statistical properties for a measured large data set collected from different areas of Finland. However, the study material does not contain the smart meter measurements of the large data set. Therefore, the measured data set is used as the input for the synthetic load profile generator as explained in the following chapters, because it is the only consumption data set available in the study material. The large data set was used to analyze and calculate different parameters in previous research activities. The content of each data set is described below in detail.

Table 4.1 *Definition of the type consumer classes used in this thesis (source: [5])*

Type consumer class	Activity description	Energy consumption (MWh/a)
1	Summer cabin	1.0
2	Apartment, 1 - phase connection	1.5
3	Apartment, 3 - phase connection	2.5
4	Detached house, no electric heating	5.0
5	Detached house, energy efficient, electric heating	10
6	Detached house, direct electric heating and timed domestic water heater	16
7	Detached house, electric storage heater	19
8	Outdoor lighting, pecu switch	34
9	Farm, cattle farming	42
10	Business, short opening hours	50
11	Industry, small-scale, 1 - shift	180
12	Business, long opening hours	600
13	Industry, connected to medium voltage network, 1 - shift	1000
14	Industry, connected to medium voltage network, 3 - shift	6000

The previously computed data in the study material for the large data set consist of different calculated type consumer load profiles in the year 2018 calendar, hourly energy distributions for four different distribution models (i,e. normal, log-normal (Logn), Gaussian mixture model (GMM), and Logn + GMM), hourly energy histogram and the 10 highest average peak powers for different calculation methods etc. Table 4.2 shows the number of customers in the large data set used to calculate the above data.

The measured data set consists of hourly smart meter measurement data from 1682 customers in 2016, and the customers are grouped according to the type consumer classes. Describing the content of this measured data set further, it includes sub data sets such as active power (kW) data for type consumer classes 1 to 14, imported reactive power measurement data for type consumer classes 10 to 14, and exported reactive power data for type consumer classes 11 and 13. The analysis in this thesis is carried out only for active power measurement data. Table 4.2 shows the number of customers for each type consumer class in the measured data set, and it will be useful to understand the differences in the results between type consumer classes described in the next chapters. The size of the measured data set is 8784 x 1682 (i.e. 2016 is a leap year and therefore there are 8784 hours).

Table 4.2 *The number of customers available in the measured data set for each type consumer class*

Type consumer class	**Number of customers**	
	Small data set (referred to "measured data set")	**Large data set**
1	247	10246
2	172	15375
3	456	57734
4	213	11453
5	165	1759
6	113	1600
7	80	616
8	36	1209
9	35	207
10	69	604
11	24	77
12	44	187
13	21	115
14	7	47

As well as, there is a data matrix called "info2016" that includes time-related data (e.g. timestamp, season, month, day, hour etc.), long-term average temperature and hourly temperature measurement data for the geographical area of the measured customers. The size of this information matrix is 8784 x 12 and, it contains the data mentioned above for every hour of the measured data set.

5. METHODOLOGIES FOR GENERATING SYNTHETIC LOAD PROFILES AND COMPARING THEM

Generation of synthetic load profiles with traditional MC methodology can be found in several research activities in the literature. First, this chapter explains the algorithm of traditional MC for the application of synthetic load profile generation. Later, a slightly improved version of the traditional MC methodology for improving outputs will be presented as a new approach to synthetic load profile generation. In literature, another approach of MC using multinomial logistic regression models called “adaptive MC” can be found. The adaptive MC gives better seasonal variations in annual synthetic load profiles compared to the traditional MC. All the above three methodologies will be clearly explained with details in this chapter. The algorithms suggested in this chapter will be a good source for future research activities. In the end, a comparison between the traditional MC and suggested approach of MC will be provided.

5.1 Traditional first-order Markov chain methodology

A Markov model requires a finite number of Markov states to proceed with the steps of the chain algorithm. Therefore, it is crucial to choose an appropriate state-space system with an appropriate number of states to minimize the issues in the resultant output data (i.e. synthetic load profiles). In this thesis, the appropriate number of states is selected by running the MC algorithm several times with a different number of states, as explained in subchapter 6.1. A Transition Probability Matrix (TPM) of a Markov model is constructed based on the state space and the input data set. There is no precise way to define a state space for a Markov model. In the literature, different approaches to defining a state space are proposed according to the characteristics of the input data set. Those approaches are explained in subchapter 3.1.2 in details. The algorithm of a traditional MC to generate a synthetic load profile is explained in this subchapter step by step. The smart meter (SM) measurement data set described in chapter 4 is used as the input data matrix for this MC. The algorithm generates load profiles only for one type consumer class at a time. Therefore, the input data matrix only consists of the filtered measurement data for the relevant type consumer class. Assuming one hour time resolution with N number of states for a particular type consumer class, this matches to a TPM with a dimension of N x N. If the input data set has c number of customers for the chosen type consumer class, the matrix dimension of the input data set will be $8784\ x\ c$ (i.e. the input data set contains measured data in 2016 and this year has 366 days). The load distributions of type consumer class (e.g. hourly distributions) are observed, and it can be seen that those load distributions differed over time. Therefore, the synthetic time series data must be time-independent, so that non-homogeneous TPMs are implemented in the model. Which means every hour, a different TPM will be constructed in the algorithm. Only precise subsets from the input data matrix will be used to calculate probability entries of a TPM in each hour. For instance, only states recognized at hour 1000 are used to calculate the probabilities of transitions from hour 1000 to 1001 for a First-order MC (FOMC). The constructing of TPM is a straightforward process and, has been explained in subchapter 3.1.4. A FOMC is implemented in this thesis since higher-order MC reduces the amount of data available for constructing the TPM. The algorithm is described in the following text by outlining the major steps in order.

1. Choose a consumer type, select the input data matrix and a suitable number of states.

2. Define the state space for the input data matrix (i.e., smart meter measurements). Each state means an interval of the input data series. Since input data is used to determine the states, it is better to define states according to a statistical method so that an equal amount of data will be distributed among the states. Therefore, this will prevent as much as possible the absorbing states (i.e. also known as closed states) that cause breaking of MC (i.e., see subchapter 3.1.3).

3. Convert the power values of the input data matrix into states and obtain the input states matrix. The state can be determined by comparing each time step's power value of the input data matrix with power intervals of each state in the state space. The state with the power interval that belongs to the input power value is chosen as the corresponding state for the time step and so on.

4. Define the initial state, set it as the state at hour 1 in the output matrix. (i.e., because, for a FOMC, the first state needs to be given as an input to the algorithm). This initial state is randomly obtained based on the probability of each state in the first hour of the input state matrix ($c\ x\ 1$ array). The probability of each state can be calculated based on the frequency of occurrence of each state in the first-hour row of the input states matrix.

5. The following loop can be executed until the required hourly output data is covered. Since the initial state is selected randomly in step 4, the sequence gives hourly data from second hour onward.

   Let the current hour be n,

   5.1 Obtain the state at $(n-1)^{th}$ hour (say i) from the output data matrix (if, the current hour is 2, the previous state (i.e., 1st hour) is equal to the initial state).

   5.2 Find all the transitions at $(n-1)^{th}$ hour from input states matrix and generate non-homogeneous transition probability matrix (TPM) for all the found transitions (say P)

   5.3 Generate cumulative transition matrix for TPM from step 5.2 (say F)

   5.4 Generate a uniform random number between 0 and 1 (say U_i).

   5.5 Find the j^{th} column in the i^{th} row (i.e. state at $(n-1)^{th}$ hour) of the cumulative matrix such that,

   $$F_{i(j-1)} < U_i \leq F_{ij}$$

   5.6 This column number j is the state of the current hour and set it in the output matrix.

   5.7 State of the current hour (output of step 5.6) must be transformed into power values as below to represent the output as a power value at hour n.

   - Generate a uniform random number between 0 and 1 (z)
   - Fit a Probability Density Function (PDF) (e.g. GMM) to the power values of the input data matrix for hour n. Derive Cumulative Density Function (CDF) from PDF and sample the corresponding power value from this CDF in the range defined by the state j using z

At the end of this algorithm, a synthetic load profile will be created for the selected type consumer class. A random walk is performed to generate synthetic load profiles using a MC. In the literature, two different methods can be found for analyzing a time series using random walks. In this thesis, the Markov chain Monte Carlo simulation (MCMC) method is used (i.e. steps 5.4 - 5 .6 in the loop). Above explained steps of the traditional algorithm can be clearly understood by using a visually presented flow chart as in Figure 5.1.

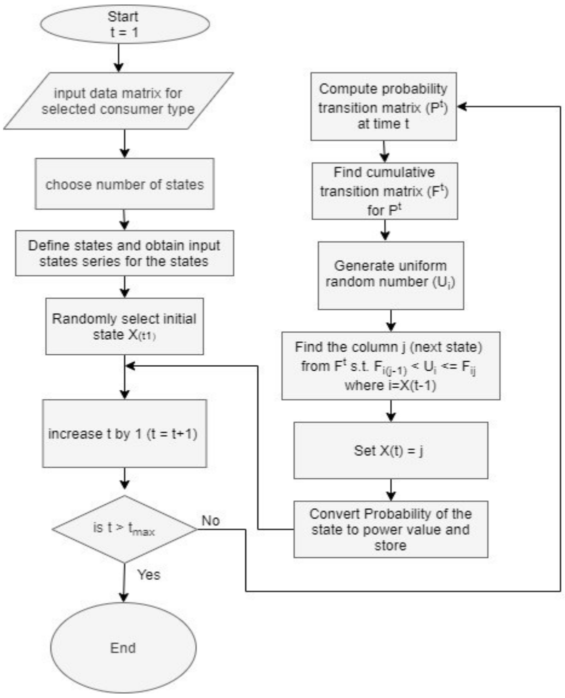

Start
t = 1
input data matrix for selected consumer type
choose number of states
Define states and obtain input states series for the states
Randomly select initial state X(t1)
increase t by 1 (t = t+1)
is t > t_max
No
Yes
End
Compute probability transition matrix (P^t) at time t
Find cumulative transition matrix (F^t) for P^t
Generate uniform random number (U_i)
Find the column j (next state) from F^t s.t. F_i(j-1) < U_i <= F_ij where i=X(t-1)
Set X(t) = j
Convert Probability of the state to power value and store

Figure 5.1 *Flow chart representation for traditional MC in synthetic load profile generation application*

5.2 Implementation of traditional Markov chain and its output

The traditional Markov chain presented in subchapter 5.1 was implemented in MATLAB software and used to generate 100 load profiles for the type consumer class 7 by using SM data in the data set. At very first glance to the load profiles in the sample, it seems that each load profile has unexpected constant high amplitude spikes. Two such instantaneous load profiles from the sample are shown in Figure 5.2 below. The investigated reason for high and continuous spikes is the limited amount of data in the data set for the type consumer class. Moreover, another issue that was observed, the hourly peaks of the generated load profile are failed to track by the algorithm. Therefore, generated synthetic load profiles slightly deformed visually than the available measured load profiles. At least, none of the output load profile gives a visually satisfying result with compared to the measured load profiles in the data set.

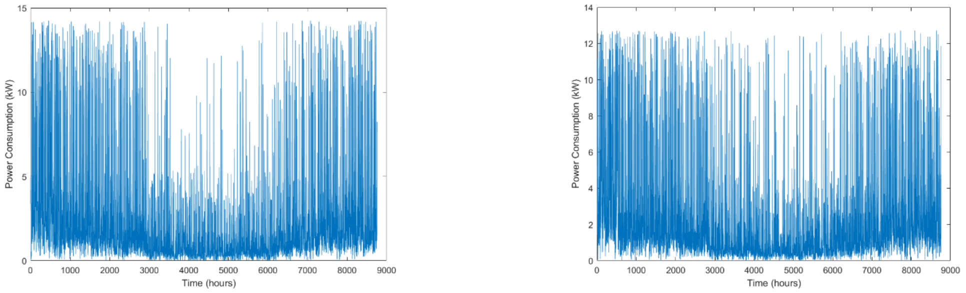


Figure 5.2 *Two synthetic load profiles that were generated from traditional Markov chain methodology for type consumer class 7*

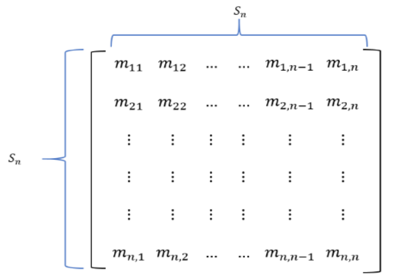
5.3 A suggested new approach for generating synthetic load profiles

Due to poor results in generated synthetic load profiles by the traditional MC as discussed in the subchapter 5.2, the traditional MC algorithm is modified and introduced as another version of the traditional MC in this thesis in order to minimize the issues and generate more realistic synthetic load profiles. Though the input data is limited, another additional set of accessible data can be found in the study material. This suggested methodology uses data from that study material. The algorithm in the suggested methodology works better than the traditional methodology, which will be analyzed later and, can be used when the amount of input data for the TPM is slightly small. However, it is always recommended to collect a considerably large data set for the type consumer class to provide as an input for the MC algorithm. In the design phase of this new algorithm, two alternative ideas were proposed related to the way of considering the previous hour's state to construct the TPM. Both ideas were implemented as two separate methods in this thesis work to find the effect of these methods for the synthetic load profiles and their electrical parameters. Then most viable method will be chosen through the observations as explained later with decision makings. After initial tests and observations from the algorithm, some improvements were added to optimize the outputs to get closer to the desired outcomes of the load profiles (i.e. annual average energy, peak powers). Those improvements are included through both scientific and trial and error methods. The idea behind the suggested new approach for MC can be reviewed as below.

The main issue in the traditional MC methodology is the continuous and high spikes through the time series as shown in Figure 5.2. Therefore, a method should be developed to control the peak power of each hour. In the study material, hourly power distributions, peak power distributions, corresponding histograms and results from different other analyses for each type consumer class can be found. They were calculated and recorded for the large data set by *Mutanen et al.* [5]. The idea is, when the hourly peak power is known, the possible output power variations can be controlled between 0 and the peak power of that hour. In the traditional method, the power range is always constant. So, a single state set is used with a constant maximum power for every hour. Therefore, a dynamic state-space generating scheme is suggested to that algorithm which generates a dedicated set of states according to the power limitation based on the hourly power histogram of a particular hour instead of a fixed state-space like in the traditional method. This can be achieved by slightly modifying the traditional MC algorithm in order to redefine a dedicated state space for each hour. More precisely, step 2 of the traditional algorithm in subchapter 5.1 should be added into the loop in step 5 at the beginning of the loop. When the state space of an hour is changed, the input state matrix should also be redefined according to the dedicated new state space for the particular hour. When it comes to constructing the TPM, two different methods can be suggested to generate non-homogeneous TPM for each hour based on the previous hour's state. Those two methods will be explained in subchapters 5.3.1 and 5.3.2.

5.3.1 Method of constructing the TPM by converting the previous hour's state to the current hour's state

This method suggests converting the previous hour's power in the input matrix into current state space's state matrix in order to construct the TPM. In the traditional MC methodology, TPM represents the transitions that always uses one state-space. Therefore, this method was proposed as an attempt to construct TPMs similar to the traditional MC methodology. When constructing the TPM for a FOMC, first, row vectors which represent the hour n and $n-1$ will be obtained from the input data matrix, and both vectors will be translated into current hour's (i.e. n^{th} hour) state-space states. The rest of the steps to construct the TPM is straightforward and, the probability of each element of TPM will be calculated based on the number of transitions from the previous state to the current state as in subchapter 5.5. The TPM representation for this method is shown in Figure 5.3.


S_n
S_n
m_{11} m_{12} ⋯ ⋯ $m_{1,n-1}$ $m_{1,n}$
m_{21} m_{22} ⋯ ⋯ $m_{2,n-1}$ $m_{2,n}$
$m_{n,1}$ $m_{n,2}$ ⋯ ⋯ $m_{n,n-1}$ $m_{n,n}$

$m_{ij} = ij$ *th element of the TPM* $S_n =$ *State space at hour n*

***Figure 5.3** The TPM representation for the method with converting the previous hour's state into the current hour's state*

When the TPM is implemented by using "from" and "to" states of transitions with a single state space set (i.e., state space in the hour n) as shown in Figure 5.3, the previous hour's output state also must be converted into current state space in order to apply the MCMC correctly because the current and previous state sets might be in 2 different scales so that the same state number may have two different power intervals. In other words, the previous state can be a different state with respect to the current hour's state space when comparing the power values, and this can be further explained below.

Let S_n be state system at hour n, Then, the states set at hours $n-1$ and n are given as,

$$S_{n-1} = \{s_{(n-1),r}\} \qquad r\, \forall 1 \ldots . N \tag{5.1}$$

$$S_n = \{s_{n,r}\} \qquad r \; \forall \; 1 \dots . N \tag{5.2}$$

where N is the number of states.

If state spaces in 2 consecutive hours are in 2 different power scales (let V_{n-1}, V_n where $V_{n-1} \ll V_n$), and assuming states have been determined by dividing the maximum power equally into the number of states, the power of the state k is given as,

$$s_{(n-1),k} = k.\frac{V_{n-1}}{N} \tag{5.3}$$

$$s_{n,k} = k.\frac{V_n}{N} \tag{5.4}$$

$$\text{where} \qquad s_{(n-1),k} \neq s_{n,k} \qquad \because V_{n-1} \ll V_n$$

Therefore, a state in the previous hour's state space set (i.e., $s_{(n-1),k}$) is not equal to the state in the current hour's state space set (i.e., $s_{n,k}$).

In order to find the previous hour's output state, several mathematical operations can be performed. In this thesis, this conversion was implemented by converting the previous hour's generated output power to a state of the current hour's state space with several computational logics in order to avoid absorbing states. Let previous power value which is generated by the algorithm be v_{n-1} and the previous state can be found by finding r such that,

$$s_{n,r-1} < v_{n-1} \leq s_{n,r} \tag{5.5}$$

where r is the state number.

The above-mentioned computational logics shift one state forward and backwards from the found output state to find a non-absorbing condition if the output state is already giving absorbing states.

5.3.2 Method of constructing the TPM without converting the previous hour's state to the current hour's state

The second method is to construct the TPM by just keeping the previous state in the state space of the previous hour without any change. That means TPM's rows will represent the states in the previous hour's state space, and columns will represent the states in the current hour's states space. When constructing the TPM for a FOMC, transitions of states from hour n-1 to n will be collected from the row vectors in the input data matrix same as in 5.3.1, but states will remain in corresponding hour's state space set. Compared to the first method in 5.3.1, this method reduces the computational tasks in the algorithm, so that it increases the efficiency. The rest of the steps to construct the TPM is straightforward and, the probability of each element of TPM will be calculated based on the number of transitions from previous states to current states as described in subchapter 5.5. The TPM representation for this method is shown in Figure 5.4.

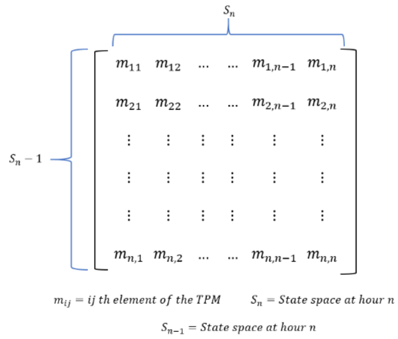

S_n
$S_n - 1$
m_{11} m_{12} … … $m_{1,n-1}$ $m_{1,n}$
m_{21} m_{22} … … $m_{2,n-1}$ $m_{2,n}$
$m_{n,1}$ $m_{n,2}$ … … $m_{n,n-1}$ $m_{n,n}$
m_{ij} = ij th element of the TPM
S_n = State space at hour n
S_{n-1} = State space at hour n

Figure 5.4 *The TPM representation for the method without converting the previous hour's state into the current hour's state*

5.3.3 The suggested MC algorithm

The suggested modified version of traditional MC algorithm for the synthetic load profile generator is presented in this subchapter. The primary modifications to the traditional MC to obtain the suggested MC are briefly explained in subchapter 5.3. Step 5.3 in the traditional MC can be performed with either method described in subchapters 5.3.1 or 5.3.2. This algorithm is also using the same dimensional input matrix as in traditional MC and non-homogeneous TPM will be constructed in each hour. First, A FOMC is implemented for the suggested methodology and later it is developed for a second-order MC to compare the variations between synthetic load profiles and, to show the effect of limited data availability for higher-order MCs. The algorithm is described in the following text by outlining the major steps in order.

1. Choose a consumer type, select the input data matrix and define the number of states.
2. Define state space for hour 1 (S_1). The power range of the state space is selected from 0 to the maximum power of the particular hour. The peak power of an hour can be determined from the power histogram of hour n in the study material.
3. Convert the power values of the input data matrix and obtain the input state matrix using the state space for hour 1 (i.e. S_1).
4. Define the initial state of the FOMC as described in step 4 of subchapter 5.1 for the traditional MC algorithm.

5. The following loop can be repeated until the required number of hourly data is covered in the synthetic load profile. Since the initial state is selected randomly in step 4, the loop gives hourly output power values from 2nd hour onward.

   Let the current hour be n,

   5.1 Define the state space for hour n (S_n).

   5.2 Convert the power values of the input data matrix into state matrix using the state space S_n for hour n.

   5.3 Generate non-homogeneous TPM (say P) for all the found transitions from hour $n-1$ to hour n by using either method in subchapters 5.3.1 or 5.3.2.

   5.4 if the method in subchapter 5.3.1 was used in the previous step (i.e. step 5.3), convert previous hour's generated output power value into a state (say i) of current state space S_n as explained at the end of the same subchapter. Otherwise, keep the previous hour's output state (say i) same as it is in previous hour's state space (i.e., no conversion is required)

   5.5 Generate cumulative transition matrix for TPM from step 5.4 (say F).

   5.6 Generate a uniform random number between 0 and 1 (say U_i)

   5.7 Find the j^{th} column in the i^{th} row (i.e. state at $(n-1)^{th}$ hour) of the cumulative matrix such that,

$$F_{i(j-1)} < U_i \leq F_{ij}$$

   5.8 Set column number j as the state of the current hour and set it in the output matrix.

   5.9 State of the current hour must be transformed into power values as below to represent the outcome as a power value for hour n.

   - Generate another uniform random number between 0 and 1 (z).
   - Obtain CDF from the hourly GMM PDF (i.e. from the study material) for hour n from the study material. Sample the corresponding power value from this CDF in the range defined by the state j using z

Above described steps can be presented in a flow chart as in Figure 5.5, and it only represents the process of generating a single synthetic load profile. Multiple synthetic load profiles can be generated by running the process repetitively. The execution time of the algorithm can make faster when generating more than one synthetic load profile especially for large type consumer classes (e.g., 12, 13 and 14) by determining hourly state spaces, TPMs and Cumulative Transition Matrices (CTMs) and taking those steps off from the loop (i.e. reducing execution time by running the same repetitive steps only once).

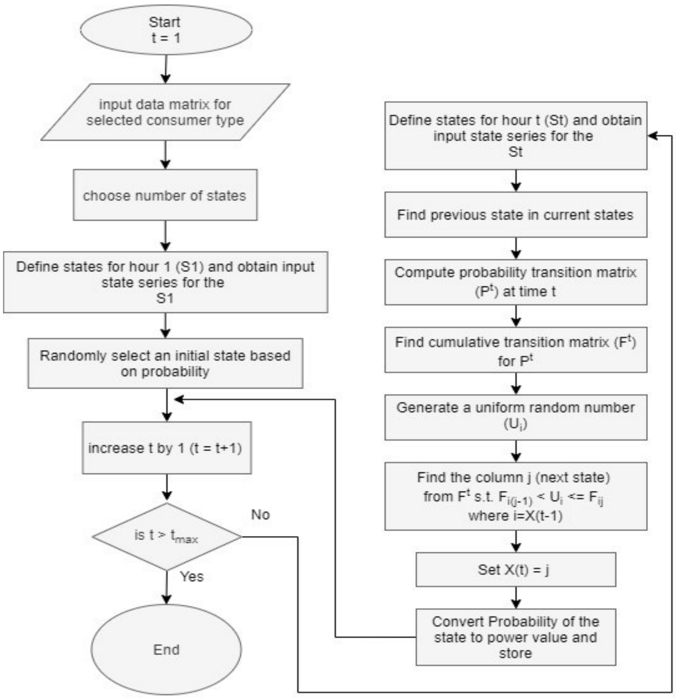

Start
t = 1
input data matrix for selected consumer type
choose number of states
Define states for hour 1 (S1) and obtain input state series for the S1
Randomly select an initial state based on probability
increase t by 1 (t = t+1)
is t > t_{max}
No
Yes
End
Define states for hour t (St) and obtain input state series for the St
Find previous state in current states
Compute probability transition matrix (P^t) at time t
Find cumulative transition matrix (F^t) for P^t
Generate a uniform random number (U_i)
Find the column j (next state) from F^t s.t. $F_{i(j-1)} < U_i <= F_{ij}$ where i=X(t-1)
Set X(t) = j
Convert Probability of the state to power value and store

Figure 5.5 *Flow chart representation for suggested MC in synthetic load profile generation application*

5.3.4 Comparison of the results between 5.3.1 and 5.3.2

To find the effect of the two methods described in subchapters 5.3.1 and 5.3.2 for output synthetic load profiles, samples of 100 synthetic load profiles were generated for type consumer classes 1-13 by using two methods respectively. From the results, no significant difference could be observed in the synthetic load profiles visually at first glance. As the next chapter explains, three measures will be used to find the accuracy of the MC algorithm in this thesis. Those are average annual energy, highest peak power and load duration curves. In this subchapter, a quick comparison was made between the two methods for the generated synthetic load profiles samples before the analysis in chapter 5 using two of the measures. First, the annual average energy was calculated for the samples of each type consumer class and each method, and values are presented in

Table 5.1. According to Table 5.1, the annual energies are almost close to each other between the two methods. Moreover, the highest peak power was also calculated using the generated samples for each type consumer class and method, and any significant difference between the values of the two methods could not be observed. However, several computational logics have been used in the implementation of state conversion method, and for this reason, this method takes comparatively large execution time compared to the other method when generating more synthetic load profiles. The execution time of the algorithm is an important measure in load profile generation and analysis. Also, as subchapter 5.3.5 describes, the state conversion method could be conceptually ineffective and may give disproportional transitions in different cases. Therefore, considering the execution time and the circumstances in subchapter 5.3.5, the rest of the comparisons in this thesis will be continued with the method described in subchapter 5.3.2 (i.e., without the state conversion method).

Table 5.1 *Evaluation of annual average energies for the generated samples of synthetic load profiles with two methods as described in subchapter 5.3.1 and 5.3.2*

Consmer Class	**Sample size of the generated synthetic load profiles**	**Annual average energy (without temp. norm) (MWh)**		**Difference (%)**
		Markov chain with converting previous hour's state into the current hour's state space method	**Markov chain without converting previous hour's state into the current hour's state space method**	
1	100	0.864	0.867	0.35
2	100	1.44	1.44	0.00
3	100	2.33	2.34	0.43
4	100	4.75	4.73	0.42
5	100	9.50	9.50	0.00
6	100	15.40	15.30	0.65
7	100	17.46	17.64	1.02
8	100	30.91	30.83	0.26
9	100	43.38	43.50	0.28
10	100	45.47	43.88	3.62
11	100	162	160.53	0.92
12	100	540	551	2.00
13	100	840	821.58	2.24
14	50	4625.48	4521.05	2.31

5.3.5 Potential issues arising from the method outlined in 5.3.1; With examples

The hypothesis used in the state conversion method is to keep the previous hour's input state vector in the current state space instead of the previous hour's state space, so that format of the TPM will be identical to the form of traditional MC methodology's TPM (i.e. see Figure 5.3). However, this method creates a conceptually high number of hits in the last rows of the TPM when the previous hour's maximum power is larger than the current hour's maximum power due to different scales in the state spaces. Also, when the previous hour's maximum power value is smaller than the current hour's maximum power, the process will be less effective due to sub optimal use of TPM. These two scenarios can be further understood by using the numerical example presented below. Figure 5.6 shows TPM representation for the case when the previous hour's state space's scale is larger than the current hour's state space’s scale.

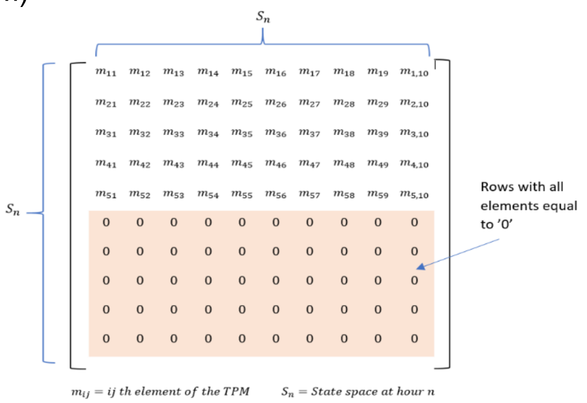
Let the number of states for the type consumer class be 10 and states are defined using equal division method, maximum power at hour n, V_n = 20 kW. Then, the state space at hour n is given as,

$$S_n = \{2, 4, 6, 8, 10, 12, 14, 16, 18,20\} \ kW$$

Next, consider the 2 cases described in the beginning of this subchapter.

Case 1: when maximum power at hour n-1 is less than maximum power at hour n ($\boldsymbol{V_{n-1} < V_n}$, assume $\boldsymbol{V_{n-1} = 10\ kW}$)

If above state-space S_n is used to translate the power values at hour $n-1$, the maximum possible state that can be found in the input states vector becomes 5, because of the maximum power at the hour $n-1$ is 10 kW. Therefore, rows 6 to 10 in the TPM will be 0 (sub optimal use of TPM)



$$\underbrace{\begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} & m_{15} & m_{16} & m_{17} & m_{18} & m_{19} & m_{1,10} \\ m_{21} & m_{22} & m_{23} & m_{24} & m_{25} & m_{26} & m_{27} & m_{28} & m_{29} & m_{2,10} \\ m_{31} & m_{32} & m_{33} & m_{34} & m_{35} & m_{36} & m_{37} & m_{38} & m_{39} & m_{3,10} \\ m_{41} & m_{42} & m_{43} & m_{44} & m_{45} & m_{46} & m_{47} & m_{48} & m_{49} & m_{4,10} \\ m_{51} & m_{52} & m_{53} & m_{54} & m_{55} & m_{56} & m_{57} & m_{58} & m_{59} & m_{5,10} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}}$$

S_n
S_n
Rows with all elements equal to ’0’
$m_{ij} = ij\ th\ element\ of\ the\ TPM$
$S_n = State\ space\ at\ hour\ n$

***Figure 5.6** Matrix representation when maximum power at the previous hour is higher than the current hour*

Case 2: when maximum power at hour n-1 is greater than maximum power at hour n ($\boldsymbol{V_{n-1} > V_n}$ assume $\boldsymbol{V_{n-1} = 30\ kW}$)

If state space S_n is used to translate the power values at hour $n-1$, all the power values above the maximum power at hour n (i.e. 20 kW) do not anymore get a room for a state higher than 10. Therefore, all the power values above 20 kW at hour n will be counted for the last state (i.e. 10) of S_n. So, there will be disproportionally a high number of hits for the last row in the TPM.

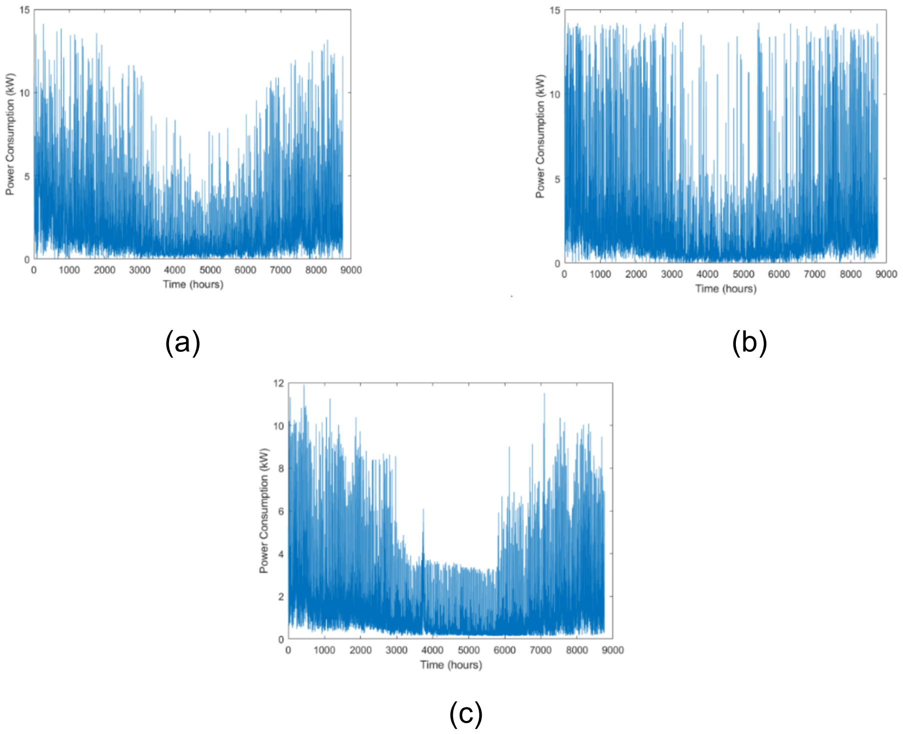
5.4 Comparison between the suggested and traditional methodologies

As a summary, the main differences between the two methodologies can be listed as below.

- Dedicated state space for every hour
- The input state vector of each hour is defined according to the corresponding state space of the hour
- The output state of an hour is translated into power by using the cumulative probability function from the hourly distributions of the large data set. In the traditional method, the cumulative probability function is determined by using the hourly distributions of the measured SM data set.

These differences come along as several additional or modified steps in different locations of the traditional MC algorithm (i.e. modified version of traditional MC) as explained in subchapter 5.3. However, the same functionalities of these steps can be combined into a one and included into a single location (i.e., inside step 5.7) of the traditional MC as an alternative algorithm to the suggested methodology. Nevertheless, it is easier to understand the process with the algorithm presented in this thesis, rather than adding all the steps into a single step directly.

To compare the outputs of the suggested and traditional methodologies, two samples with 100 synthetic load profiles for type consumer class 7 was generated using two methodologies. Figure 5.7 shows a load profile for the customer number 25 in the measured data set with the closest two synthetic load profiles that were generated from traditional and suggested MC methodologies. The index used to find most approaching synthetic load profile is the minimum MAPE and RMSE combination of synthetic load profile in the samples compared to the measured load profile. It should be noted that every synthetic load profile generated by traditional methodology has constant high peaks with a value close to the maximum power of the state space, and reasons for this is discussed in the subchapter 5.2. For instance, in the presented synthetic load profile from traditional methodology has a constant band of spikes from hours around 0 to 3000 and 6000 to 8760. Moreover, the same high spikes can be seen hours from around 3000 to 6000 as well, but less frequently. These variations have slightly deviated when compared to the measured customer load profile in Figure 5.7. However, the synthetic load profile from suggested methodology shows spikes with varying amplitudes, and also the upper and lower bands have followed the measured customer's load profile quite similarly. Still, slight spikes can be seen throughout the time series in the synthetic load profile from suggested methodology also, but the values of the spikes are limited, and they show similar characteristics as in measured customer load profile. (e.g. hours between 3000 and 6000). The presence of spikes in a synthetic load profile must be acknowledged because a synthetic load profile generator gives a probabilistic load profile based on the provided input data set and random uniform numbers as for MCMC in the algorithm. The synthetic load profiles from the traditional methodology have RMSEs in the range of 2.12 kW – 2.63 kW compared to the data of customer number 25 in the measured data set, while the suggested methodology gives only RMSEs in the range of 1.71 kW - 1.86 kW. Moreover, RMSEs were calculated for all the synthetic load profiles in the samples against the measured data set and results were in the range of 1.78 kW – 3.43 kW and 1.39 kW– 2.91 kW for traditional and suggested methodologies respectively. The selected load profiles of Figure 5.7 have RMSE of 2.31 kW and 1.73 kW respectively for traditional and suggested methodologies. Based on these observations, it can be concluded that the suggested MC methodology is giving better output load profile compared to the traditional MC methodology.



(a)

(b)

(c)

***Figure 5.7** Generated (a) suggested vs (b) traditional synthetic load profiles vs (c) a most approaching load profile in SM data set (consumer type 7 customer 25)*

5.5 The suggested approach for second-order Markov chain

In this thesis, the effects of FOMC vs Second-order MC (SOMC) for the output synthetic load profiles were observed before carrying out the main analysis. The suggested new methodology is used to implement the algorithm for SOMC over traditional MC due to the conclusion from subchapter 5.4. The steps related to a SOMC should be added or modified in the methodology explained in subchapter 5.3.3 for the implementation of SOMC, because steps given in subchapter 5.3.3 are developed for a FOMC. More about FOMC and SOMC have been clearly defined in subchapter 3.1.4.

The SOMC requires two initial states for first 2-time steps at the beginning of the MC. These two initial states will be randomly selected based on the probability of each state's occurrence. The state at the 1st time step is found similarly as in FOMC algorithm. Once the first state is selected, transitions from that state at hour 1 are chosen. Thus, the probability of getting each state in 2nd time step from the previously found state can be calculated by using the hit counts and the second state can be initialized randomly by using these state probabilities. After selecting initial states, in order to find the next state in each hour, transitions from 2 previous output states are used to construct the TPM in SOMC. Constructing the TPM is a straightforward process and similar to the steps explained in subchapters 3.1.4 and 5.3.3. The dimension of the TPM will be (N x N) x N, where N is the number of states in the state space. If the output state at $(n-1)^{th}$ hour is i and $(n-2)^{th}$ hour is j, row ij of TPM represents transitions from states ij to others, and next state k can be found with the random walk procedure as described in the suggested algorithm for FOMC. This SOMC also implemented in MATLAB and used to generate 100 synthetic load profiles samples for each type consumer class. For these samples, the average energies and peak powers were calculated and compared with the corresponding values from the generated samples for FOMC in subchapter 5.4. It can be observed that the average energies and peak powers are almost close to each other. However, as explained in subchapter 3.1.4, higher-order MC reduces the amount of the data available for constructing TPMs. Moreover, the measured data set used in the thesis also has a limited number of customers per each type consumer class. Due to these reasons, SOMC gave synthetic load profiles with limited combinations of input data set compared to FOMC. Therefore, in this thesis, the final analysis will be carried out with FOMC rather than SOMC.

5.6 The adaptive Markov chain in the literature

In the literature, a synthetic load profile generator has been developed by using binomial logistic regression and the MC model (i.e. also known as an ‘adaptive MC’) [23]. According to the analysis of that research, adaptive MC has minimized error between aggregated SM data and synthetically generated data, as well as, it has successfully captured seasonality as compared to traditional MC. In this thesis, the same adaptive MC is explained step by step clearly, which cannot be found in the literature.

The idea behind adaptive MC is to generalize the concept of time-inhomogeneity without loss of accuracy. For that purpose, each element in a row of TPM is represented by a multinomial logistic regression $h_{i\theta}(x)$ that learns the corresponding transition probability of the element.

$$h_{i\theta}(x) = [\varphi_{i1} \quad \varphi_{i2} \quad \dots\dots\dots \quad \varphi_{i(n-1)} \quad \varphi_{in}] \tag{5.6}$$

$$\text{Where} \qquad \varphi_{ij} = \frac{e^{\eta_{ij}}}{\sum_{k=1}^{n} e^{\eta_{ik}}} \qquad , \quad \varphi_{ij}(x) \in [0, 1]\ , \eta_{ij} = \theta_{ij}^{T}x$$

Where i, j represents an arbitrary row, column of TPM and n is the total number of power states (i.e. also equal to the length of a row/column in TPM) respectively. In this application, x represents the time related features (i.e. $x = (1, hour, day, month)$). θ denotes the vector of coefficients for the features. (i.e., $\theta = (\theta_0, \theta_1, \theta_2, \theta_3)$). The coefficients should be calculated using a learning process that aimed at minimizing a cost function. The theoretical background of multinomial logistic regression has been discussed in subchapter 3.2. In this methodology, the hour feature is defined by using the values from 1 to 24, where 1 = 0001 h and 24 = 2400 h and so on (i.e. $hour = \{1, 2, \dots\dots, 24\}$). The day of the week is defined by using the values from 1 to 7, where 1 stands for Monday and 7 for Sunday etc. (i.e. $day = \{1, \dots. 7\}$). Also, the month feature can be defined as $month = \{1, \dots 12\}$, where 1 = January and 12 = December.

For the sake of simplicity of explanation, one SM customer in type consumer class is considered from the data set in this subchapter. But the same methodology can be expanded simply for a group of customers of the same type consumer class. First, a state-space should be defined for the input data matrix and the input data matrix should be converted into states in order to obtain input state matrix as described in steps 2 and 3 of the traditional MC algorithm. This input state matrix can be used as the overall training data set for the logistic regression, where each time step of the data set represented by the three time-related features. Table 5.2 shows a sample training data set to demonstrate how the training data set looks like.

$$\begin{bmatrix} \varphi_{11} & \varphi_{12} & \cdots & \cdots & \varphi_{1(n-1)} & \varphi_{1n} \\ \varphi_{21} & \varphi_{22} & \cdots & \cdots & \varphi_{2(n-1)} & \varphi_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \varphi_{n1} & \varphi_{n2} & \cdots & \cdots & \varphi_{n(n-1)} & \varphi_{nn} \end{bmatrix}$$

***Figure 5.8** The TPM representation with multinomial logistic regression*

***Table 5.2** Demonstration of an overall sample training data set for a single customer*

Customer 1	**State at time t**	**Feature “hour”**	**Feature “weekday”**	**Feature “month”**
hour1, Fri, Jan	1	1	5	1
hour2, Fri, Jan	2	2	5	1
hour3, Fri, Jan	1	3	5	1
hour4, Fri, Jan	3	4	5	1
…	…	…	…	…
hour11, Wed, Jun	1	11	3	6
hour12, Wed, Jun	6	12	3	6
…	…	…	…	…
hour22, Sat, Dec	3	22	6	12
hour23, Sat, Dec	1	23	6	12
hour24, Sat, Dec	4	24	6	12

Based on this training data set, the TPM can be constructed and Figure 5.8 shows how the TPM looks like after applying multinomial logistic regressions. Each element of the TPM is a function of the three input features defined previously which outputs a value between 0 and 1. Each row of the TPM can be thought of as a multinomial logistic regression model. For clarification purposes, let’s take the example of “Hour t - state 1” which means that at current time t, the power state is 1 in order to train the functions φ_{11} to φ_{1n}. For that, only the rows in the training data set containing “state 1” at time t must be considered (i.e. highlighted in orange in Table 5.2). Then, in order to train the multinomial logistic regression φ_{1j} (i.e. $= \{1, .., n\}$), the state at time t+1 is set as the target. The target represents a class in a multinomial logistic regression. Therefore, it is also called multiclass logistic regression. Based on the previous sample training data set given in Table 5.2, the specific training data set for calculating the coefficients of the functions in φ_{1j} would be:

Table 5.3 *Selected data set from overall training data set in order to calculate functions of* $\boldsymbol{\varphi_{1j}}$

Training data set to determine φ_{1j}	Feature "hour"	Feature "weekday"	Feature "month"	Target (Class)
hour 1, Fri, Jan	1	5	1	2
hour 3, Fri, Jan	3	5	1	3
…	…	…	…	…
hour 11, Wed, Jun	11	3	6	6
…	…	…	…	…
hour 23, Sat, Dec	23	6	12	4

After applying multiple linear regression to the selected dataset in Table 5.3, the coefficient matrix for the number of n functions is obtained. The coefficient matrix has a dimension of $4\ x\ n$ (i.e. coefficients for intercept and 3 features (total is 4)). Since each row of the TPM is a multinomial logistic regression model, the above example steps should be applied to all transitions from each state in the overall training data set.

When the training for each case was done, each element of the TPM can be derived for a certain time in terms of features x. For instance, when the features for a certain time t (e.g., hour 4, Thu, Apr (4, 4, 4)) is fixed, the derived functions can output the probability for transitions to each of the states at time t+1 (i.e.., hour 5, Thu, Apr). The input for the logistic regression functions (φ_{ij}) is features of time t (e.g. 4, 4, 4). Since each row is a multinomial logistic regression, the sum of the output of the functions in the same row is equal to one as in traditional MC. In this methodology, 24 × 7 × 12 = 2016 combinations of hours, weekdays and months exist. Therefore, the adaptive MC can also be considered as a traditional MC whose TPM has a dimension of 2016 × n × n. Note that the regression models can be tuned by choosing different other time-related features to capture the seasonality (e.g. hourly temperatures). And also, above defined values for features can be further adjusted to improve accuracy (e.g. weekdays, weekends can be grouped separately and used the values 1 and 2 instead of values 1 to 7, Using values 1 to 4 for the months according to the four seasons of the year instead of 1 to 12 etc.). Once the TPM is constructed for time t, the synthetic power value for time t can be obtained using the random walk process explained in the traditional MC section.

This algorithm was implemented in MATLAB and using Python language. However, the synthetic load profiles from the program were not visually satisfied as expected because this algorithm relies significantly on the accuracy of the multinomial logistic regression models. The outputs were generated for the type consumer class 7 and that training data set was an extremely imbalanced one. Therefore, a proper resampling technique should be used (e.g. near-miss, over-sampling, under-sampling etc). Due to limited timeframe, no further improvements to this algorithm have been made, and these developed steps can be used in the future research work with proper deep learning techniques for further tuning the accuracy of the models. However, as discussed later in chapter 6, the suggested MC methodology (see subchapter 5.3) in this thesis is also showing better results (i.e. low MAPE and capturing seasonal variations accurately).

5.7 Temperature normalization of consumption data

In subchapter 2.4, several factors affecting electricity consumption are discussed in detail. It is not fair to compare consumption data from different years because of these dependencies on data. In order to compare original consumption data from different years, those data must be normalized to a common environment to treat them equally. According to subchapter 2.4, weather factors such as temperature, daylight, as well as wind and humidity affect electrical demand. In this thesis, only the temperature dependency has been considered because the outdoor temperature is the major weather dependent factor for electric load. It can be assumed that the temperature sensitive part of the load depends on the temperature by normalizing the temperature of the consumption data. The used temperature dependency model in this thesis is shown in (5.7) [4].

$$\Delta P(t) = a(t)\ x\ (T_{24}(t) - E[T(t)]) \qquad (5.7)$$

where the symbols are denoted as,

$\Delta P(t)$: the outdoor temperature dependent part of electric load at time t

$a(t)$: the customer class specific load temperature dependency parameter (W/°C)

$T_{24}(t)$: the average outdoor temperature from the previous 24 hours, and

$E[T(t)]$: the expected value of the outdoor temperature.

The $E[T(t)]$ contains long term monthly average temperatures of the data recorded location. The $T_{24}(t)$ can be calculated by taking the average of previous 24 hours outdoor temperatures as in (5.8):

$$T_{24}(t) = \frac{\sum_{i=t-24}^{t-1} T(i)}{24} \qquad (5.8)$$

The customer class specific load temperature dependency parameter contains the 6 values for the year which each value represents two consecutive months starting from January. Calculation of temperature dependency parameters and more details can be found in the literature [4]. Once, the temperature dependent part of the load in each time is found, temperature normalization can be done by performing algebraic operations.

5.8 Approximation to the type consumer load profiles by optimally matching the aggregate load profile

If the synthetic load profiles generated from any of the methods described in the previous subchapters are realistic, their average load profile should also reach toward the corresponding type consumer load profile. In other words, these synthetic load profiles must fill in the aggregate load profile obtained by multiplying the type consumer load profile by the number of synthetic customers. This thesis suggests a multiple linear regression for scaling the synthetic load profiles in a realistic range in order to best fit to the aggregate load profile. This method is best suited for a large synthetic load profile sample, because a small sample may mathematically but not realistically fit well to the aggregate load profile. The results of this method are shown later in subchapter 6.3. In multiple linear regression, the synthetic load profiles and aggregate load profile are taken as explanatory variables and response variable respectively. When there are N number of time steps in each load profile, the variables can be represented as follows.

$$\text{Explanatory variable: } x_j = \left(x_{1j}, x_{2j}, \dots \dots, x_{Nj}\right)^T \tag{5.9}$$

$$\text{Response variable: } y = (y_1, y_2, \dots \dots, y_N)^T \tag{5.10}$$

When there are m number of explanatory variables, the corresponding linear regression model can be introduced as below,

$$y = X\beta + \epsilon \tag{5.11}$$

where,

$X = (x_1, x_2, \dots \dots . x_m)$ is the input explanatory variable matrix
$\beta = (\beta_1, \beta_2 \dots \dots . \beta_m)^T$ is the vector of regression coefficients
$\epsilon = (\epsilon_1, \epsilon_2 \dots \dots . \epsilon_N)^T$ is the vector of residuals

Each regression coefficient represents the corresponding scaling factor of the synthetic load profile. The regression coefficients can be estimated by solving the regression in (5.11). This regression problem is solved by minimizing the squared distance between the linear combination of explanatory variables and the response variable y as below.

$$\hat{\beta} = \underset{\beta}{\arg min} \sum_{i=1}^{N} \epsilon_i^2 = \underset{\beta}{\arg min} \epsilon^T \epsilon = \underset{\beta}{\arg min} (y - X\beta)^T (y - X\beta) \tag{5.12}$$

The equation in (5.12) can be simplified and, the estimator of β can be derived as in (5.13).

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{5.13}$$

In this load profile generation application, N becomes 8760, and m differs according to the number of customers of each type consumer class in the data set. But this chosen time window for N (i.e. 8760) is a bit long. Therefore, the accuracy of the linear regression model can be lower, as the model attempts to fit a larger input data sample. Therefore, a piecewise regression can be used to further improve the accuracy of the model. In piecewise regression, the above explanatory variable is partitioned into small intervals. After that, the linear regression is performed on each interval at breakpoints independently. By doing so, the piecewise regression can model the data in each interval and thus the entire data set. Once the regression coefficients are found using the above procedure, the scaled synthetic load profiles can be derived by multiplying each individual load profile with corresponding regression coefficient (i.e. $x_j\beta_j$). The sum of these scaled synthetic load profiles are closer to the aggregate load profile as shown in the subchapter 6.3 later.

6. RESULTS AND VALIDATION

This chapter presents the results from the synthetic load profile generator and analysis of generated synthetic load profiles. The suggested FOMC methodology was used to generate synthetic load profiles based on the decisions taken in comparison to the other methodologies in chapter 5. Mainly, three measures are used to analyze the results (i.e., annual average energy, peak powers and load duration curves). Later, validation of synthetic load profiles is presented at the end of this chapter.

6.1 Selecting the number of states for each type consumer class

As described in chapter 5, the number of states should be carefully selected in order to get better outputs from the MC algorithm. As the number of states of MC increases, the output synthetic load profile attempts to over fit the input data. When the number of states of the MC is lower, hourly power consumption in a synthetic load profile is varying rapidly and randomly. Therefore, many spikes can be seen in the load profile and this results in lower details in synthetic load profiles as discussed in subchapter 3.1.4. These two scenarios are clearly depicted in Figure 6.1 with two random load profiles from the traditional MC methodology when the number of states is equal to 5 and 25. The selection of an optimal number of states for each type consumer class is described in this subchapter.

The state selection process of this thesis was performed using the traditional MC methodology. First, samples with 100 synthetic load profiles were generated for a different number of states such as 5, 10, 15, 20 and 25 for all type consumer classes. There was a total of 5 x 14 (i.e. 5 different number of states, 14 type consumer classes) samples. Then the average load profile was generated for each sample, and MAPE was calculated for each synthetic average load profile with respect to the corresponding measured customer average load profile. Figure 6.2 shows the effect of the number of states on MAPE for type consumer class 7. According to Figure 6.2, when the number of states is increased, MAPE is decreased. A lower MAPE is theoretically better and has a less error in synthetic load profiles. However, a large number of states may provide less amount of data for calculating probabilities of TPMs, thereby increasing the risk of overfitting of output data. Therefore, a reasonable high number of states was chosen for each type consumer class considering the visual inspection and the MAPE between average load profiles. Thus, the number of states for each type consumer class is selected using the method described above, and they are recorded in Table 6.1.

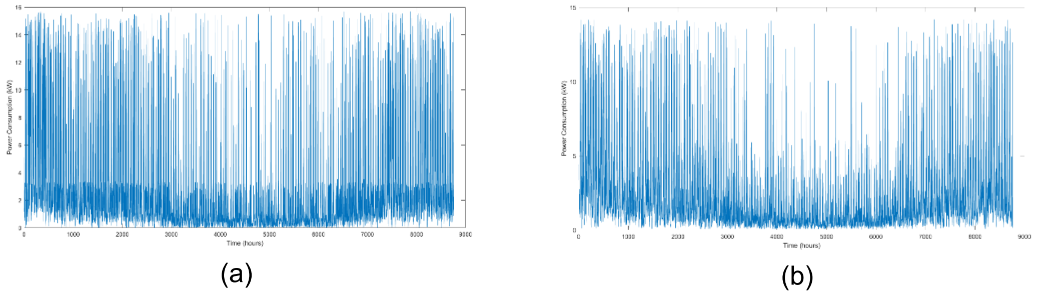


Figure 6.1 *Two instantaneous synthetic load profiles when the number of states is (a) =5 (b) =25 for type consumer class 7*

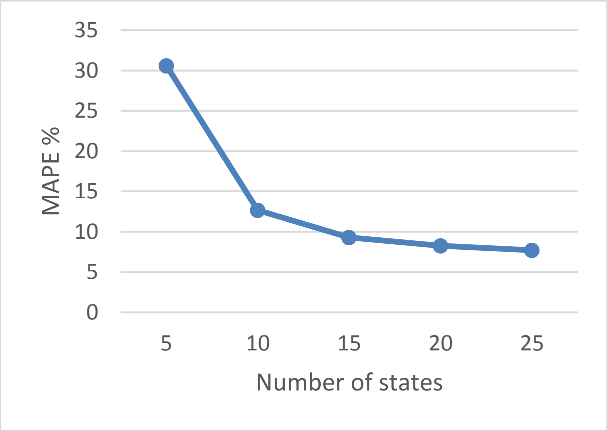


Figure 6.2 *Evolution of MAPE between average load profiles from measured load profiles and synthetic load profiles for the type consumer class 7 with the traditional MC*

Table 6.1 *Number of states selected for each type consumer class based on MAPE and visual inspection*

Type consumer	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Size of the state space	20	15	15	15	15	15	15	15	15	15	10	10	10	10

6.2 Analyzing the synthetic load profiles from suggested FOMC

This analysis is carried out with the suggested FOMC methodology described in subchapter 5.3 with the method in subchapter 5.3.2 based on decisions taken with the initial observations from the two methodologies explained in subchapters 5.3.4 and 5.4. The generated load profiles by the suggested MC methodology can be seen to be more realistic compared to the traditional MC methodology, especially when the input data used to calculate the probabilities of MC's TPM are less. In order to analyze the synthetic load profiles, samples of 100 synthetic load profiles for type consumer classes 1-13 and a sample of 50 synthetic load profiles for type consumer class 14 were generated and stored them in MATLAB variables to analyze the output data with three measures. The measures are annual average energy, peak powers and load duration curves of synthetic load profiles. The next three subchapters compare the synthetic data in terms of the above measures with the SM and type consumer data.

For type consumer classes 11 to 14 with large power consumers, they take a long time for generating a one synthetic load profile with compared to the small type consumer classes, because their consumption values are higher, the acceptable resolution for defining a CDF without distorting the output is small (i.e. more data needs to be stored). Therefore, the computer needs more memory and processing power for computing. Therefore, the sample size of type consumer class 14 has been limited to 50. Type consumer classes 1 and 13 took around 15 minutes and 18 hours respectively for generating 100 load profiles, while type consumer class 14 took about 32 hours for generating only 50 load profiles. Different techniques can be used to optimize the MC algorithm further and improve the execution time without distorting the output of the algorithm, and this can be done as a separate research task in future.

6.2.1 Annual average energy of load profiles

In this analysis, the average load profiles of the synthetic and measured customers are compared with type consumer load profiles. In this thesis, synthetic load profiles are generated using the input data set measured in 2016. That data set is fed directly into the input of the synthetic load profile generator without any composing of data such as temperature normalization. The output of the synthetic load profile generator is based on the measured data. Therefore, the output of the synthetic load profile generator also can be thought of as a load profile without temperature normalization. However, the type consumer load profiles provided in the study material are temperature normalized and applied to 2018 calendar. Therefore, in order to compare average measured customer or synthetic average load profiles with the type consumer load profiles, first, those average load profiles must be temperature normalized and then projected from 2016 to 2018. Figure 6.3 shows the flow chart of the average profile forming procedure developed to compare with the type consumer load profiles.

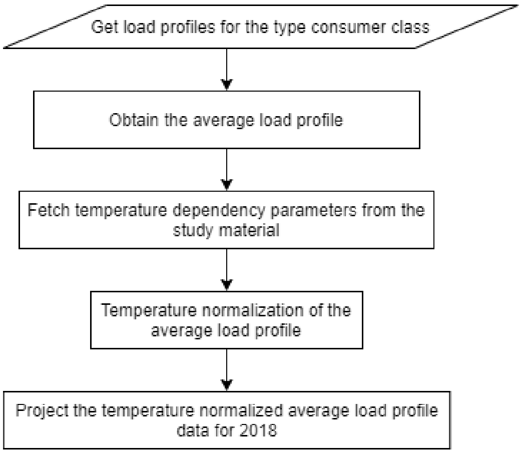

Get load profiles for the type consumer class
Obtain the average load profile
Fetch temperature dependency parameters from the study material
Temperature normalization of the average load profile
Project the temperature normalized average load profile data for 2018

***Figure 6.3** Flow chart of the average profile forming procedure in order to compare the average load profile with given type consumer class load profiles in the study material*

The flow chart shown in Figure 6.3 is applied to a chosen type consumer class and this average profile forming procedure was implemented in MATLAB. The function starts with loading the load profiles (i.e. measured or synthetic load profiles in 2016) to the program, then generating the average load profile for the given input load profiles. After that, the temperature dependency parameters are fetched for the type consumer class of the input load profiles from the study material, and the average load profile is temperature normalized using the method explained in subchapter 5.7. The temperature normalization allows consumption data to be treated equally to other years. Later, the temperature normalized average load profile should be projected to 2018 in order to change the weekdays and special days of 2016 to corresponding days of 2018. Antti Mutanen has developed a function in his research for projecting load profiles between years. In this thesis, the same function has been used to project the temperature normalized load profiles from 2016 to 2018. After applying this process, this temperature normalized and projected average load profile can be used to compare with the corresponding type consumer load profile in the study material. Rest of the content in this subchapter uses the term average load profile for temperature normalized and projected average load profile. There are 14 type consumer classes in the study material, and synthetic, measured average load profiles for type consumer classes 1, 4, 7 and 10 are shown in Figure 6.4. The first and second columns represent the synthetic and measured average load profiles respectively, while the third column represents the corresponding type consumer load profile from the study material.

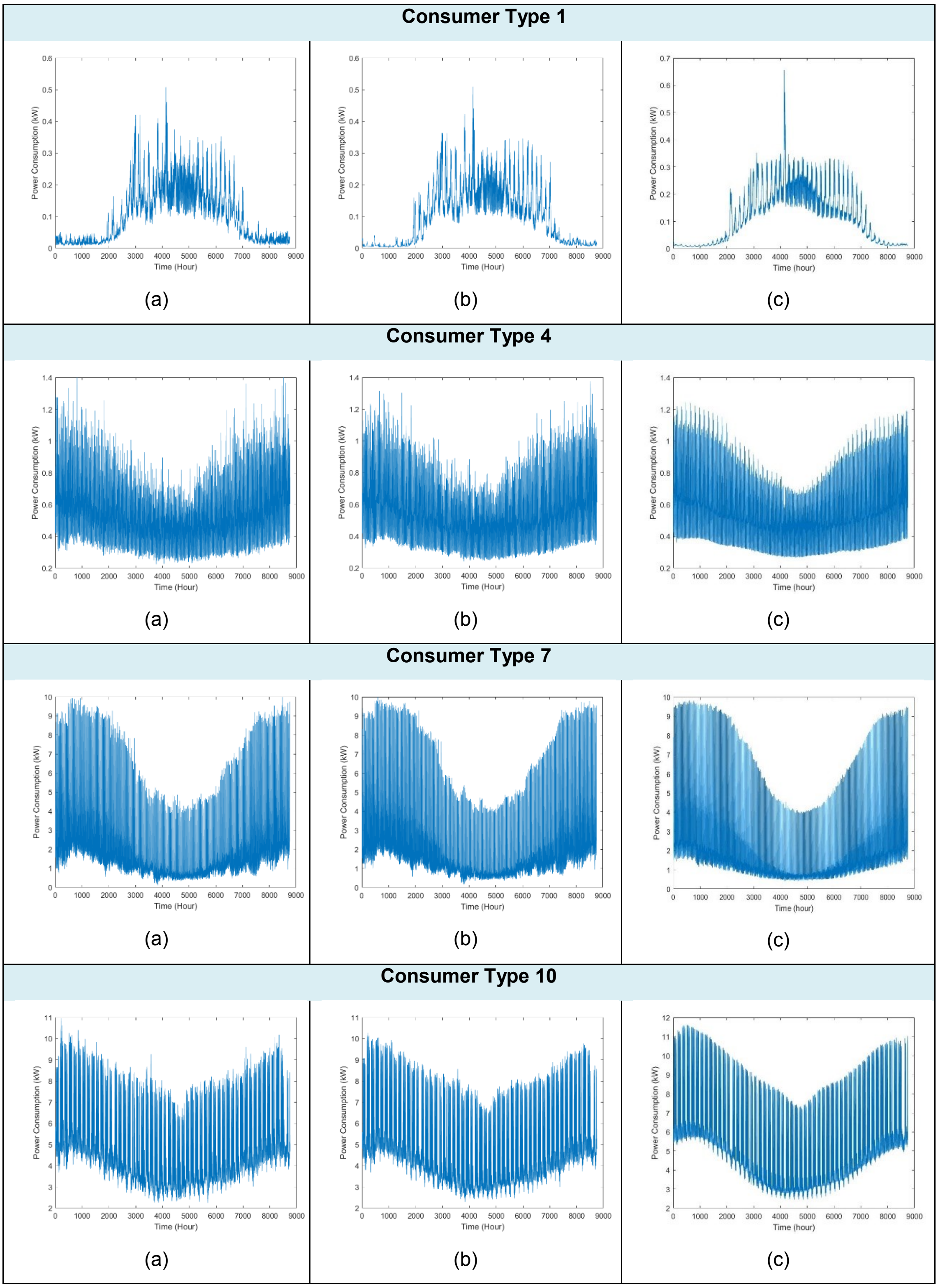


Figure 6.4 *Average load profiles for type consumer classes 1,4,7 and 10 (rows 1,2,3,4 respectively); Columns represents (a) average profile from 100 generated synthetic load profiles (b) from measured customer load profiles (c) type consumer load profile.*

According to Figure 6.4, the synthetic average load profile of each type consumer class has followed the shape of its corresponding measured customer average load profile almost identically. At first glance, synthetic and measured average load profiles look very similar, because the highlighted spikes and variations look also similar in both load profiles. This is quite natural because the measured data set is the input for synthetic load profile generator. Both these average load profiles are less smooth with compared to its corresponding type consumer load profile. The type consumer load profile is derived for comparatively a large number of customers (i.e. from the large data set), therefore, the spikes have been eliminated. However, the measured data set is a small data set compared to the large data set, and this might be the reason for having less smoothness in the measured average load profile. Therefore, the synthetic load profiles appear to be trying to follow the measured data set. However, the sample size is only 100, so that it is too early to guess without testing this for a large sample of synthetic load profiles. This will be analyzed for a large sample of synthetic load profiles in subchapter 6.3.

The annual average energy for each type consumer class was calculated from the derived synthetic and measured average load profiles. Average annual energies for type consumer classes are already available in the study material. All these values are illustrated in Table 6.2. Table 6.2 also includes the calculated average energies of average load profiles with and without temperature normalization. The absolute percentage errors of the annual average energy values between the measured and synthetic average load profiles as well as type consumer load profile and synthetic average load profile were also calculated and tabulated in Table 6.3 for each type consumer class. As can be seen from the tables, there are slight differences between the annual average energy values. Table 6.3 shows that the error between measured and synthetic average load profiles ranges from 0.20 % to 4.08 % for type consumer classes 1-12. But in the same column, type consumer classes 13 and 14 have comparatively bit higher errors (i.e. 9.29 % and 6.78 % respectively) than others. However, these high errors are still less than 10% and moderately acceptable. The annual average energy error between type consumer and synthetic load profiles are comparatively higher than the errors between measured and synthetic average load profiles.

Table 6.2 *Average annual energies calculated for type consumer load profiles and, synthetic and measured load profiles with and without temperature normalization*

Type consumer Class	Annual average energy (MW)				
	Before temperature normalization		After temperature normalization		Type consumer load profiles
	Synthetic average load profile	Measured average load profile	Synthetic average load profile	Measured average load profile	
1	0.867	0.858	0.879	0.867	1
2	1.44	1.45	1.43	1.45	1.5
3	2.34	2.44	2.35	2.45	2.5
4	4.73	4.85	4.79	4.9	5
5	9.50	9.60	9.87	9.95	10
6	15.30	15.42	16.23	16.30	16
7	17.94	18.34	18.94	19.28	19
8	32.21	32.23	32.31	32.17	34
9	43.50	43.60	43.84	43.96	42
10	43.88	44.0	44.55	44.64	50
11	160.53	167.87	164.19	170.74	180
12	551	566	548.65	561.68	600
13	821.58	907.7	825.22	909.78	1000
14	4521.05	4859.2	4519.5	4848.3	6000

Table 6.3 *Percentage errors between annual average energy values of synthetic and measured average load profiles/ synthetic average and type consumer load profiles*

Type consumer	Annual average energy error (%)	
	between synthetic and measured average load profiles	between synthetic average and type consumer load profiles
1	1.38	12.10
2	1.38	4.67
3	4.08	6.00
4	2.24	4.20
5	0.80	1.30
6	0.43	1.44
7	1.76	0.32
8	0.44	4.97
9	0.27	4.38
10	0.20	10.90
11	3.84	8.78
12	2.32	8.56
13	9.29	17.48
14	6.78	24.68

As shown in Table 6.2, the average annual energies of synthetic and measured customers in type consumer class 13 and 14 are significantly lower than their corresponding type consumer load profile's average annual energy compared to other type consumer classes. The reason for that is, the type consumer classes 13 and 14 consist only 21 and 7 customers respectively in the measured data set, and their power consumption values are also typically high. The type consumer classes 13 and 14 refer industrial customers connected to medium voltage network with 1 shift and 3 shifts respectively. The load behaviour of different customers in these classes can be numerous. Therefore, such a small number of customers in a type consumer class with high power consumers may not reflect the large data set. Therefore, repeating the fact that the annual average energy errors in type consumer classes 13 and 14 are relatively high in both columns in Table 6.3 compared to the others. Furthermore, it is known that the measured data set used to derive the average load profiles is small data set compared to the large data set. So, the average energy values from the measured data set are more sensitive and can fluctuate for load changes. Due to these reasons and according to the conclusion from the comparison of the tables, average annual energies of measured and synthetic average load profiles tend to each other because MC follows the input data set, while the errors between synthetic and type consumer load profiles are comparatively high always. Another significant big error that could be observed in the 2nd column is 12.10 % for type consumer class 1. The customers in type consumer class 1 consume less hourly power throughout the year (e.g. most of the hourly powers are less than 1 kW in the time series), so the percentage error equation yields a high error due to its ratio.

Table 6.4 represents the calculated MAPEs for the same above mentioned synthetic average load profiles against measured average and type consumer load profiles to measure the accuracy of the synthetic average load profiles data. As seen from Table 6.4, the MAPEs between the type consumer and synthetic average load profiles plus the type consumer and measured average load profiles are both considerably higher and approximately close each other. But the MAPEs between measured and synthetic average load profiles are comparatively very low. Therefore, there is a relatively *balanced* data *flow* between measured and synthetic average load profile data sets. Also, for the type consumer class 1, a significant higher MAPE can be observed in all three columns of Table 6.4. Based on the MAPE equation, if actual values are too small (e.g. less than 1), the ratio becomes a more substantial value, and eventually, the outcome returns a large error percentage. In the type consumer class 1, the power values are comparatively low and there is a large percentage of data with less than 1kW power consumption values. Due to this reason, MAPEs of the type consumer class 1 are considerably higher.

Table 6.4 *Calculated MAPEs for the synthetic average load profiles against measured average and type consumer class load profiles*

Type consumer class	**Mean Absolute Percentage Error (MAPE %)**		
	synthetic data vs measured data	**synthetic data vs type consumer data**	**measured data vs type consumer data**
1	32.91	43.18	29.27
2	6.61	9.27	6.66
3	7.85	10.39	6.58
4	5.82	7.41	5.00
5	4.67	6.57	4.96
6	4.40	7.74	6.70
7	5.82	14.19	12.77
8	11.84	41.81	40.89
9	3.91	8.39	8.01
10	3.34	11.13	10.35
11	8.50	12.37	9.25
12	3.80	9.26	7.00
13	8.83	14.47	8.86
14	7.26	24.61	19.85

6.2.2 Peak powers of load profiles

The second measure used in this thesis to compare the synthetic load profiles with the referenced load profiles is the average peak powers of the load profiles. As described in chapter 4, the study material contains the 10 highest average annual peak powers and monthly peak powers for the large data set with different calculation methods. In this analysis, the highest 10 average annual peak powers for the measured and synthetic data sets are also calculated separately. Unlike the annual average energy calculations, the average annual peak powers of the study material have been calculated for the load profile data without temperature normalization. Therefore, synthetic and measured load profiles can be used directly to calculate annual peak powers without temperature normalization. In this comparison, the peak powers obtained from the study material and the measured data set are used to compare with the average annual peak powers calculated for the synthetic load profiles data set. The corresponding peak powers of each data set are tabulated in Appendix A1. Below, Table 6.5 shows the highest average annual peak powers (i.e. first) extracted from Appendix A1 for each type consumer class. The percentage of errors between the data sets for the highest average annual peak powers in Table 6.5 are shown in Table 6.6. According to Table 6.6, the peak power error between the synthetic and measured data sets ranges from 1.17 % to 8.12 % and the errors in most of the classes are less than 5 %. The errors between the synthetic and large data sets other than classes 13 and 14 are also acceptable. The reason for this relatively large deviation in classes 13 and 14 is the presence of high power consumers in the large data set and the fact that not all of that data is included in the measured data set (i.e. size of the measured data set is small). Therefore, the synthetic load profile set is limited to the given measured data set, and peak power errors are comparatively small for the measured data set. Based on the error values in Table 6.6, the percentage errors of average annual peak powers between synthetic and measured data are also almost acceptable, especially when considering the type consumer classes with large power consumers (e.g. 13 and 14).

Table 6.5 *Highest average annual peak powers calculated for synthetic, measured and type consumer load profile data*

Type consumer Class	Highest annual peak power (kW)		
	Synthetic load profiles	Measured customer load profiles	Type consumer load profiles
1	3.295	3.217	3.473
2	1.961	1.890	1.890
3	4.123	4.437	3.937
4	5.762	5.956	5.710
5	7.551	7.366	7.372
6	9.567	9.680	9.215
7	12.771	13.051	13.091
8	10.143	9.845	10.064
9	18.460	19.196	18.885
10	19.164	17.724	19.287
11	87.465	81.277	82.288
12	133.184	126.702	128.061
13	353.975	334.735	419.397
14	1050.548	1094.971	1323.079

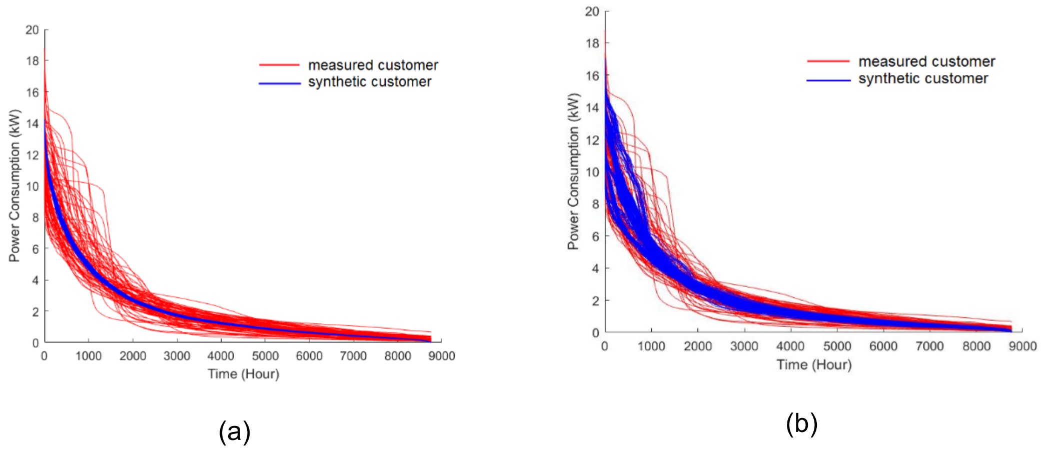
Table 6.6 *Percentage error between the highest average annual peak power values of synthetic and measured load profile, and synthetic and type consumer load profile data*

Type consumer Class	Highest peak power error (%)	
	between synthetic and measured load profile data	between synthetic and type consumer load profile data
1	2.42	5.13
2	3.76	3.76
3	7.08	4.72
4	3.26	0.91
5	2.51	2.43
6	1.17	3.82
7	2.15	2.44
8	3.02	0.78
9	3.83	2.25
10	8.12	0.63
11	7.61	6.29
12	5.12	4.00
13	5.74	15.56
14	4.06	20.60

6.2.3 Load duration curves of load profiles

The third measure of the generated synthetic load profiles is the load duration curve. For that purpose, all the power values in the annual synthetic time series for each measured customer and synthetic customer are sorted in descending order and stored in MATLAB. Figure 6.5a shows the load duration curves for type consumer class 7, with red colour the measured customers as references and blue coloured curves for synthetic customers. According to Figure 6.5a, the synthetic load duration curves do not clearly show the variations as in the measured load duration curves. It is challenging to get the same load duration curves for synthetic customers as in measured customers because MC provides only an instant load profile of all probabilistic combinations of input data set at a time, as explained in subchapter 6.4 later (see Figure 6.12). Likewise, a particular input data set may give hundreds of thousands of output combinations. Therefore, a very large synthetic load profile sample might be required to choose an appropriate set of load duration curves to represent the objectives of this subchapter.

As an alternative for this, synthetic load profiles can be generated by grouping similar characteristics of the input customers (e.g. grouping by highest peak) and running the synthetic load profile generator by taking each group as input. This new synthetic load profile set is independent of the previous samples and will not cause any conflict with the findings in the previous subchapters, because the new set also follows the measured data set. Figure 6.5b shows load duration curves of the new synthetic load profile set with the same legends used in Figure 6.5a. According to Figure 6.5b, the synthetic customer load duration curves appear to be successfully following the measured customer load duration curves. Hence, the load duration curves of the output of the synthetic load profile generator follow the load duration curves of the input data set.

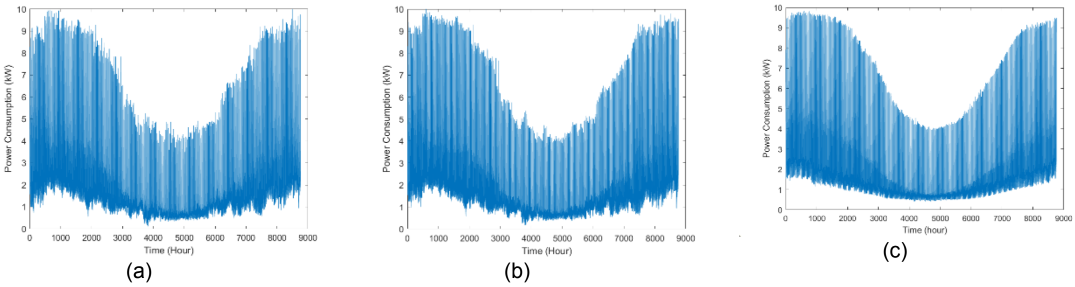


(a) (b)

Figure 6.5 *Comparison of load duration curves of synthetic customer data with measured customer data for type consumer class 7.*

6.3 Effect on measures when synthetic load profiles sample is large

All the results in the previous subchapters were obtained for samples with a small number of synthetic load profiles (i.e. 100), indicating that synthetic load profile data are close to the measured data. However, it remains to be seen how this differs for a larger sample. This requires large samples of synthetic load profiles, and it takes a long time to calculate load profiles, so this test was only performed for type consumer class 7. Accordingly, a sample of 4960 synthetic load profiles was generated. For this sample, first, the temperature normalized and projected average load profile was derived and it is shown in *Figure 6.6*a.



(a) (b) (c)

Figure 6.6 Average load profiles for type consumer class 7 (a) average profile from 4960 synthetic load profiles (b) from measured customer load profiles (c) type consumer load profile.

The calculated average annual energy and highest peak power for the synthetic average load profile are 19.12 MWh and 12.73 kW. From the previous results, the corresponding values for the small sample were 18.94 MWh and 12.771 kW. Accordingly, the annual average energy has slightly changed and still appears to be close to both type consumer and measured average annual energy. The MAPE of the synthetic average load profile with respect to the measured average load profile and type consumer load profile is 1.28 % and 12.69 % respectively. The corresponding values for the previous small sample were 5.82 % and 14.19 %. It is clear that MAPEs have been decreased and MAPE between measured and synthetic average load profiles has reduced by around 80 % for the large sample and the MAPE value is comparatively low. This means the synthetic and measured average load profile data have a good balance of data throughout the time series and they are getting closer. Anyhow, MAPE between synthetic and type consumer average load profiles has also been reduced slightly and can be thought as they are slowly converging. Overall, as described in the literature and confirmed by this and previous subchapters, the synthetic load profile generator follows the input data set and increases the accuracy, especially when the sample is large.

After that, the optimal aggregate load profile matching method described in subchapter 5.8 is applied to the above large sample to compare its results with the previous results.

As explained in the subchapter 5.8, the aggregate load profile for this method is generated using the type consumer class load profile. The input synthetic load profiles for this method are not temperature normalized, so the type consumer load profile should be temperature denormalized to treat all the load profiles equally. Subsequently, the aggregate load profile can be generated by multiplying the temperature denormalized type consumer load profile and the number of input customers for the regression model (i.e. equals to the synthetic load profile sample size (= 4960)). The piecewise regression model described in subchapter 5.8 has been implemented using MATLAB. The regression coefficients of the model for this synthetic load profile sample can be found using this script. Then, the scaled synthetic load profiles can be directly obtained using the calculated regression coefficients. Figure 6.7a shows the temperature normalized and projected average load profile for the scaled synthetic load profile set. The average load profile for measured customers and type consumer load profile in Figure 6.7b and Figure 6.7c are same as in *Figure 6.6*b and *Figure 6.6*c because they are not changed.

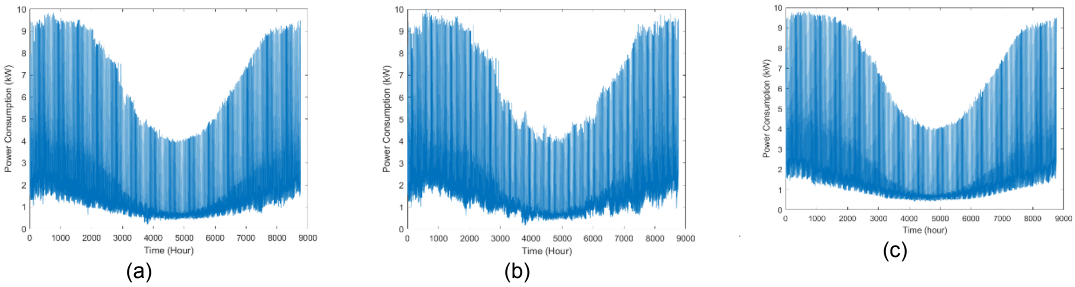


Figure 6.7 Average load profiles for type consumer class 7 average profile (a) from scaled synthetic (b) from measured customer (c) type consumer load profiles.

As shown in Figure 6.7, the average synthetic load profile appears to be closer to the type consumer load profile. Table 6.7 shows the values of the measures for the scaled synthetic load profile set and compared with the previously calculated values.

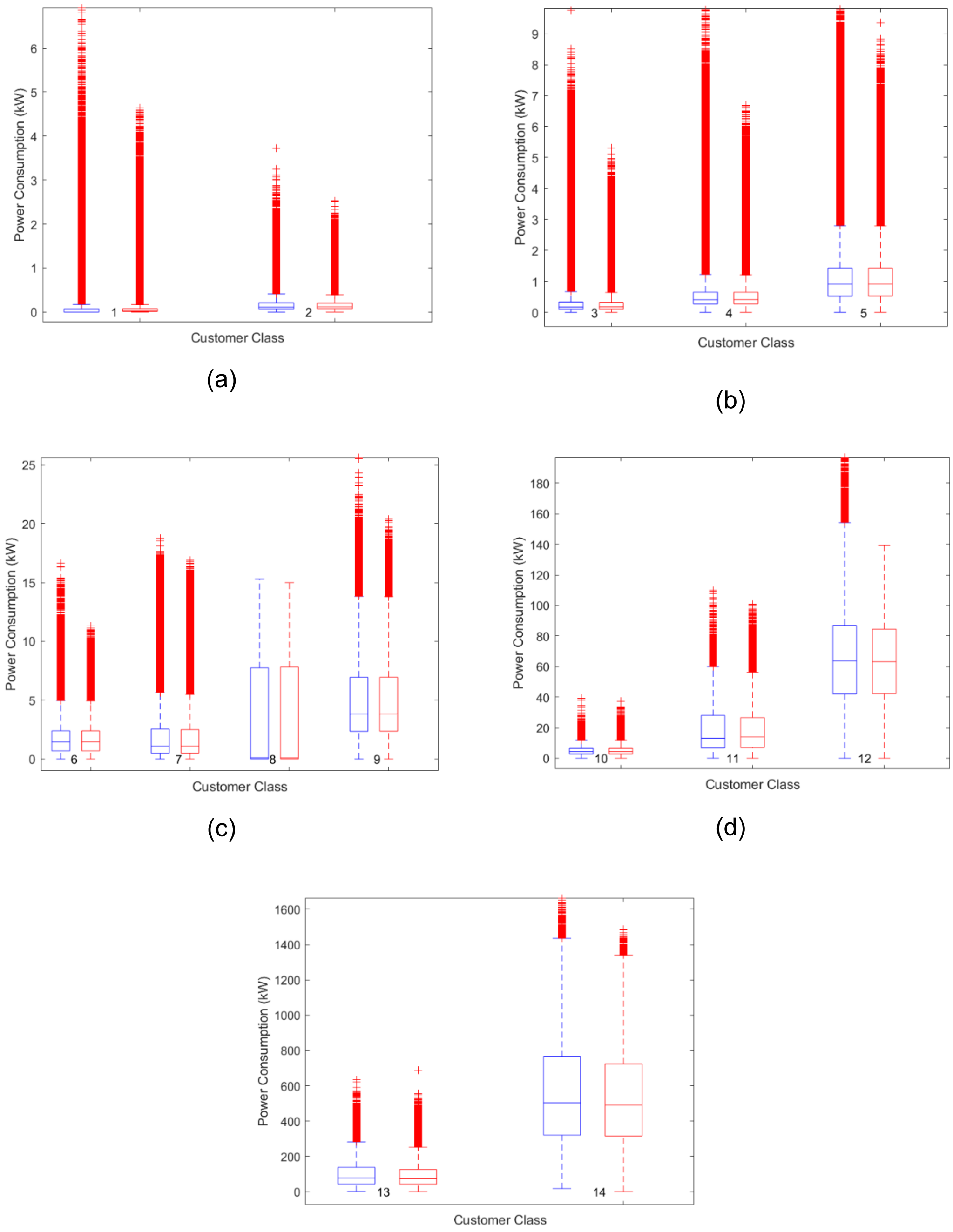
Table 6.7 *Percentage error between the highest average annual peak power values of synthetic and measured load profile, and synthetic and type consumer load profile data*

Measure	Value (With optimal load profile matching)	Error compared to type consumer (%)	
		With optimal load profile matching	Without optimal load profile matching
Average annual energy	18.62 MWh	2.00	0.63
MAPE of average load profile	4.25 %	4.25	12.69
The highest peak power	13.34 kW	2.61	1.76

Table 6.7 confirms that the optimal load profile matching method can be used to adjust the generated synthetic load profile sample to bring the aggregate load profile closer to the type consumer load profile. The results are also compared between the piecewise and linear regressions used inside this method. The piecewise regression provided a smoother and better aggregate load profile than the linear regression. The statistical properties of each individual scaled synthetic load profiles were not analyzed in this study. However, they were visually satisfying. This method also maintains errors of less than 5% for all the measures compared to without it. Therefore, this latter method can be used to further realistically adjust and scale the generated synthetic load profiles from the suggested MC methodology.

6.4 Validation of load profiles

A box plot can be used to get an idea of the dispersion of the data sets. An overview of power distribution of measurement data (left) and synthetically generated data (right) for type consumer classes 1-14 is represented by using box plots in Figure 6.8. The median and quartiles are used to construct the box of the plots. The lower quartile of the box represents the first 25 % of data, and the higher quartile represents the first 75% of data of the data set. So, the box indicates 50% of the data corresponding to the middle symmetrically distributed data in the data set. The median of the data set is indicated by the horizontal line located inside the box. According to Figure 6.8, the box length of both distributions in all the type consumer classes are identical except class 13, 14. As well as, medians of data sets of each type consumer class (i.e. as shown with the horizontal line inside the boxes) are in the same level and the skewness of data sets also appear similarly. The outliers of the data set can be found outside the whiskers. Outliers can be seen in both measured and synthetically generated data sets except for class 12. The maximum potential outlier of the synthetically generated data sets in every type consumer class is always a bit lower with compared to the measured data set. This is due to the generated data set is small and it is required to generate more load profiles to reflect peak values. The box plots confirm that the power distributions of each class are matched well. Therefore, both distributions of measured and generated load profile data have similar statistical characteristics because box plots are very similar for classes 1-12. The type consumer class 1 has a significantly minimal box length because a large percentage of data in the synthetic and measured data sets has small power values, and this is clearly reflected with a low maximum power value in the boxes of the box plots for class 1. The box plots of type consumer class 8 confirms that there are no outliers. The box plot for type consumer class 13 and 14 are different because the measured data set contains high power values and also a smaller number of customers so that the data set is highly volatile. Thus, the measured data may not represent the characteristics of a diverse data set, and the synthetically generated data set contains more load profiles than the measured data set so that it includes different possible combinations of load behaviours from the measured customers. According to this comparison, it can be concluded that the power distributions of synthetic data set manage to carry similar statistical details as in measured data set.



(a)

(b)

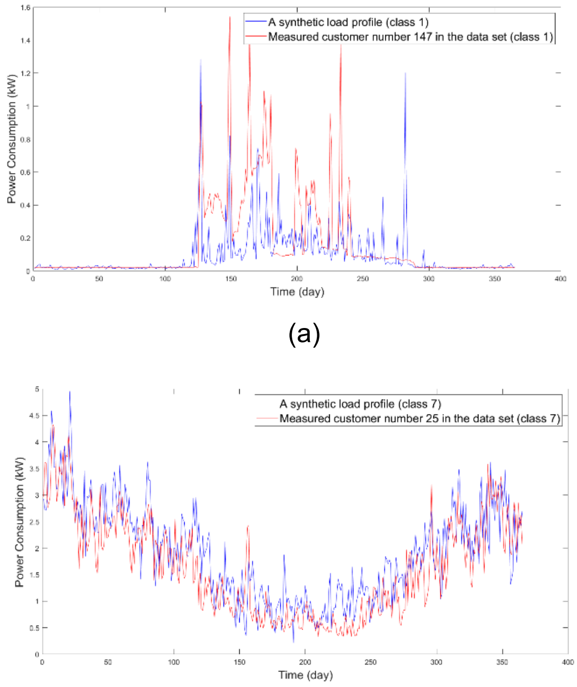
(c)

(d)

(e)

Figure 6.8 *Box plots representing the power distributions for type consumer classes (a) 1 and 2 (b) 3 to 5 (c) 6 to 9 (c) 10 to 12 (c) 13 and 14 for measured data (left) and synthetically generated data (right)*

In the literature, several MC methodologies can be found for synthetic load profile generation. Only a few of them could be able to generate synthetic load profiles that follow the seasonal variations successfully throughout the year. One of these MCs (i.e. called adaptive MC as described in chapter 5.6) was also implemented in this thesis with machine learning algorithms to compare with the suggested MC methodology in this thesis. However, the output of the adaptive MC was not satisfactory due to a highly imbalanced data set. This built adaptive MC can be continued using the steps developed in this thesis as another research task by using suitable resampling and deep learning techniques. However, the suggested methodology could be tested for seasonal variation by plotting daily power consumption from yearly load profiles. Figure 6.9a and Figure 6.9b represent the daily power consumptions of a measured and a synthetic customer for classes 1 and 7. According to Figure 6.9, The power fluctuations during summer vs other seasons for the customers in summer cottages and detached houses can be clearly observed. Therefore, Figure 6.9 confirms that the suggested methodology can nicely reflect the yearly seasonal variations in power consumption of customers. As the suggested MC model correctly takes into account the seasonal variations, it can generate synthetic load profiles that well fit to the input data.



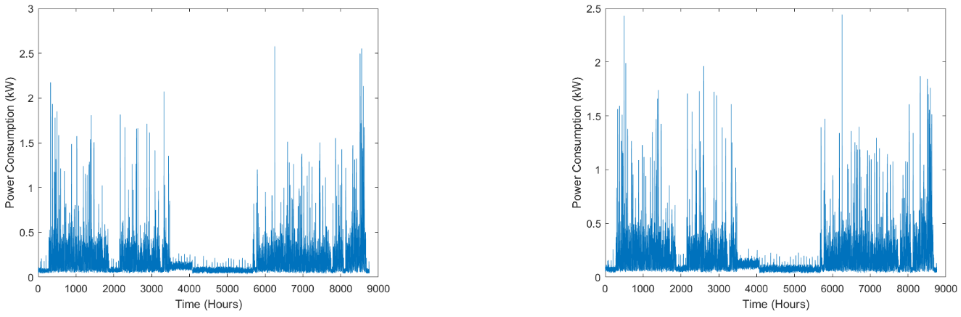
(a)

(b)

Figure 6.9 *Daily power consumptions of a measured customer and a synthetic customer using suggested MC methodology for type consumer class (a) 1 and (b) 7*

A MC provides a randomly selected output of all possible output combinations available from the input data set in one execution round. The validity of this statement and the suggested MC's ability to track the load behaviour of an input customer is verified below.

For these purposes, an input data set of customers with similar load behaviours was created. This new input data set is referred as artificial data set below. The artificial data set consists of 150 customers and is created by choosing a single customer (i.e. customer number 25) in type consumer class 2 of the measured data set. Each artificial customer load profile was created by generating a random constant load and adding that load to the previously selected real consumer load profile. This implies that each customer's hourly energy consumption values in the artificial data set may vary slightly, but the load behaviour patterns of the customers are similar to those of the actual consumer selected from the measured data set (i.e. customer number 25). Next, this artificial data set was fed as input to the suggested MC and generated a sample of synthetic load profiles. Figure 6.10 shows a random customer load profile from the aforementioned artificial data set and the synthetic load profile sample. According to Figure 6.10, it can be seen that synthetic load profile generator has successfully tracked the intra-year load behaviour patterns (e.g., at hours 1800-2200 and 3500-5800) similar to the customers with same load behaviour in the data set (i.e. load behaviour for all customers in the artificial data set is similar) though the consumption values are slightly different. An enlarged load profile of Figure 6.10 can be used further to analyze the load variations of the synthetic load profiles. For instance, let's consider day 1 (i.e. hours from 1 to 24) of the two load profiles in Figure 6.10, as shown in Figure 6.11. According to Figure 6.11, it can be seen that the synthetic profile has followed the load fluctuations of the input load profile approximately. The outputs from MC give all the possible combinations of the input load profiles. For instance, Figure 6.12 shows all the input load profiles in the artificial data set and the synthetic load profile in Figure 6.10 in one plot. It can be seen that the synthetic load profile represents one possible combination of the input load profiles. Therefore, from a large generated sample, it would be able to find a closer synthetic load profile with similar load behaviour changes of an input load profile.



(a) (b)

***Figure 6.10** A randomly selected customer load profile from the (a) artificial data set, and (b) synthetic data set generated by providing the artificial data set as input to the suggested MC*

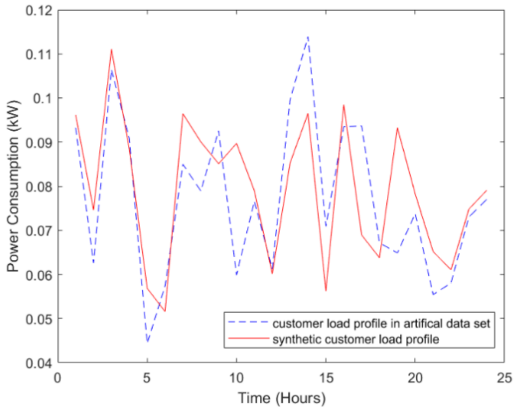


***Figure 6.11** Daily load profile of day 1 for the load profiles in Figure 6.10*

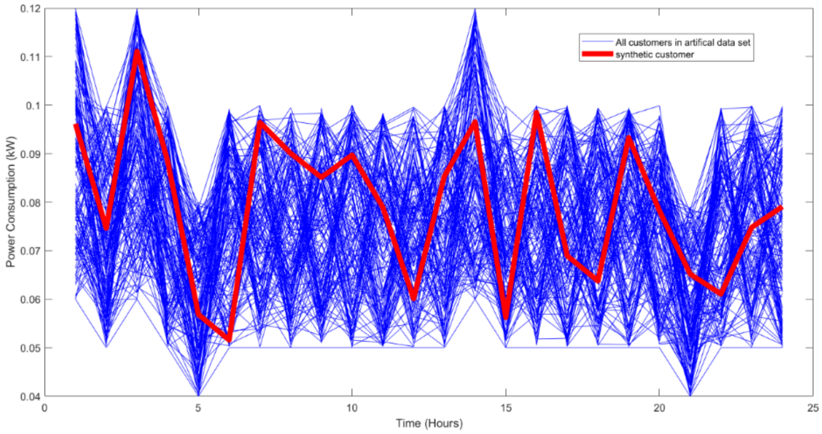


***Figure 6.12** All the input load profiles in the artificial data set and shown synthetic load profile in Figure 6.10*

These test cases in subchapter 6.4 confirm that the suggested synthetic load profile generator in thesis generates well-fitting load profiles for a given input load profile data set.

7. CONCLUSIONS

This thesis presented three methods of generating synthetic load profiles using Markov chain models. They are the traditional MC method, the suggested MC method and the adaptive MC method. The traditional MC method is a basic algorithm of MC applications and can be found in research on synthetic load profile generation in the literature. This traditional MC methodology can be slightly improved depending on the availability of additional data and application requirements. In addition to some thousands of smart meter measurement data from a specific area in Finland, the hourly power distributions, type consumer load profiles for each type consumer class derived from a large smart meter measurement data set from different areas in Finland are available (i.e. the large data set is not available) in this thesis. The suggested MC method of this thesis is a slightly improved version of the traditional MC method using additional data from the study material. After that, an attempt is made to build an adaptive MC with clear steps. Finally, a comprehensive aggregate load profile matching method is described to adjust and scale the generated load profiles into more realistic load profiles.

The thesis works begin with the implementation of the traditional MC algorithm. The algorithm is tested for several type consumer classes. The synthetic load profiles obtained from the traditional MC are greatly distorted. The synthetic load profiles have high and continuous power consumption spikes throughout the time series. Therefore, the suggested MC is developed to minimize these unsatisfactory effects from the traditional MC. The suggested MC is tested for two samples of different sizes (i.e. small - 100 and large - 4960). The results are analyzed using three measures (i.e. average annual energy, average peak power and load duration curve). The values of measures are calculated for each sample, and they are compared with the corresponding values of measures for input measured customer data set and type consumer data. At first glance, the small sample seems to attempt to follow the measured data set for all three measures when compared, because the errors between synthetic and measured data are relatively low compared to the synthetic and type consumer data. However, the large sample also shows that it is further getting closer to the measured data set than its type consumer data. For the large sample, the corresponding errors for each measure are further reduced relative to the small sample's errors. As described in the literature and from the observations of this thesis, it is confirmed that the synthetic load profile generator follows its input data set (i.e. measured data) and increases the accuracy, especially when the sample is large. But the measured data set used to generate synthetic load profiles is quite small compared to the large data set, so either the measured or synthetic data sets do not appear to reach toward the type consumer data. The results shows that the generated individual synthetic load profiles follows the input measured data set closely. But, the aggregate load profile of them is bit deviated compared to the type consumer load profile. To minimize this deviation in the aggregate load profile, an optimally matching aggregate load profile method is described. By using this method, the previously generated large sample is adjusted and scaled realistically in order to reach toward the corresponding type consumer load profile. Therefore, finally, the combination of these suggested MC and load profile matching methods achieves the goal of generating more realistic synthetic load profiles in this thesis. Anyone who needs to generate synthetic load profiles for different purposes can follow the methods in this thesis.

The suggested MC works properly for load profile generation. The load profile generator can capture the yearly seasonality successfully. Also, the power distributions of generated synthetic data confirm that synthetic data have similar statistical properties as in measured data. The performance and operation of the suggested MC are further analyzed under validation section. This study is only carried out for active power load profiles. The reactive power load profiles can also be generated in a similar way depending on the availability of reactive power smart meter measurement data. The developed adaptive MC methodology in this thesis can be further developed in the future with different deep learning techniques to get more realistic load profiles.

REFERENCES

[1] A. Bari, J. Jiang, W. Saad, A. Jaekel, Challenges in the Smart Grid Applications: An Overview, International Journal of Distributed Sensor Networks, 2014, 1-11 p, Available: https://doi.org/10.1155/2014/974682

[2] A. D. Sahin, Z. Sen, First-order Markov Chain Approach to Wind Speed Modelling, Journal of Wind Engineering and Industrial Aerodynamics, 2001 Vol. 89, pp. 263–269, Available: http://dx.doi.org/10.1016/S0167-6105(00)00081-7

[3] A. Mutanen S. Repo, P. Järventausta, Customer Classification and Load Profiling Based on AMR Measurements, 21st International Conference and Exhibition on Electricity Distribution, 2011, pp. 1-4, Available: http://sgemfinalreport.fi/files/CIRED2011-Mutanen.pdf

[4] A. Mutanen, Improving Electricity Distribution System State Estimation with AMR-Based Load Profiles, dissertation, Tampere University of Technology, 2018, Available: http://urn.fi/URN:ISBN:978-952-15-4105-6

[5] A. Mutanen, K. Lummi, P. Järventausta, Load Models for Electricity Distribution Price Regulation, 25th International Conference on Electricity Distribution, 2019, Available: http://dx.doi.org/10.34890/820

[6] A. Mutanen, Using AMR Measurements in Load Profiling and Network Calculation, Load and response modeling workshop in project SGEM, 2011, pp. 17-21, Available: http://sgemfinalreport.fi/files/W188.pdf

[7] B. Stephen, A. Mutanen, S. Galloway, G. Burt, P. Järventausta, Enhanced Load Profiling for Residential Network Customers, IEEE Transactions on Power Delivery , 2014, Vol. 29, Iss 1, pp. 88-96, Available: https://doi.org/10.1109/TPWRD.2013.2287032

[8] C. Bucher, G. Andersson, Generation of Domestic Load Profiles - an Adaptive Top-Down Approach, 12th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), 2012, pp. 436-441

[9] Course Notes: STATS 325 Stochastic Processes, Department of Statistics, University of Auckland, ch8, pp. 149-170, Available: https://www.stat.auckland.ac.nz/~fewster/325/notes/325book.pdf

[10] E. Kabalci, Y. Kabalci , Smart Grids and Their Communication Systems, Springer, Singapore, 2019

[11] Energy consumption in households fell further in 2018, website, Available (accessed on 24.05.2020): www.stat.fi/til/asen/2018/asen_2018_2019-11-21_tie_001_en.html

[12] Energy Year 2019 - Electricity, website, Available (accessed on 24.05.2020): https://energia.fi/en/news_and_publications/publications/energy_year_2019_-_electricity.html

[13] F. McLoughlin, A. Duffy, and M. Conlon, The Generation of Domestic Electricity Load Profiles through Markov Chain Modelling, 3rd International Scientific Conference on Energy and Climate Change, 2010, pp. 18–27

[14] Freedom of choice in the heating market, Finnish Energy, website. Available (accessed on 24.05.2020): https://energia.fi/en/energy_sector_in_finland/energy_market/heating_markets

[15] Heat Pump Investments up to Half a Billion a Year in Finland, website, Available (accessed on 24.05.2020): https://www.sulpu.fi/in-english

[16] Jukka V. Paatero , Peter D. Lund, A model for generating household electricity load profiles, International Journal of Energy Research, 2006, Vol. 30, Iss 5, pp. 273-290, Available: https://doi.org/10.1002/er.1136

[17] K. Huhta, Smartening up while keeping safe? advances in smart metering and data protection under EU law, Journal of Energy and Natural Resources Law, 2019, Vol. 38, Iss 1, pp. 6-7, Available: https://doi.org/10.1080/02646811.2019.1622244

[18] K. Tazi, F.M. Abbou, F. Abdi, Load Analysis and Consumption Profiling: An Overview, Proceedings of the 1st International Conference on Electronic Engineering and Renewable Energy, 2018, pp. 697–705, Available: https://doi.org/10.1007/978-981-13-1405-6_80

[19] National Report 2018 to the Agency for the Cooperation of Energy Regulators and to the European Commission, Finland, 10 p, Available (accessed on 24.05.2020): https://energiavirasto.fi/documents/11120570/13026619/National+Report+2018+Finland.pdf/beeaec3e-3fdf-d93c-fec9-9ee21a395fc9/National+Report+2018+Finland.pdf

[20] Official Statistics of Finland (OSF): Energy consumption in households [e-publication] ISSN=2323-329X. 2018. Helsinki: Statistics Finland, website. Available (accessed on 04.04.2020): http://www.stat.fi/til/asen/2018/asen_2018_2019-11-21_tie_001_en.html

[21] P. Deng, H. Wang, S. Horng, D. Wang, J. Zhang, H. Zhou, Softmax Regression by Using Unsupervised Ensemble Learning, *9th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP)*, 2018, pp. 196-201, Available: https://doi.org/10.1109/PAAP.2018.00041

[22] R.M. Ward, R. Choudhary, Y. Heo, & J.A.D. Aston, A data-centric bottom-up model for generation of stochastic internal load profiles based on space-use type. Journal of Building Performance Simulation, 2019, Vo. 12, Iss. 5, pp. 620-636. Available: https://doi.org/10.1080/19401493.2019.1583287

[23] T. Zufferey, D. Toffanin, D. Toprak, A. Ulbig and G. Hug, Generating Stochastic Residential Load Profiles from Smart Meter Data for an Optimal Power Matching at an Aggregate Level, *2018 Power Systems Computation Conference (PSCC)*, 2018, pp. 1-7, Available: https://doi.org/10.23919/PSCC.2018.8442470

[24] Video: Uuden sukupolven älykäs mittausjärjestelmä, website. Available (accessed on 24.05.2020): https://www.elenia.fi/uutiset/video-uuden-sukupolven-älykäs-mittausjärjestelmä

[25] W. Labeeuw and G. Deconinck, Residential Electrical Load Model based on Mixture Model Clustering and Markov Models, IEEE Transactions on Industrial Informatics, 2013, Vo. 9, Iss 3, pp. 1561–1569, Available: https://doi.org/10.1109/TII.2013.2240309

APPENDIX A1: PEAK POWER TABLE FOR THE GENERATED SMALL SAMPLE OF SYNTHETIC LOAD PROFILES IN CHAPTER 6 AND MEASURED DATA SET

*** N is the highest Nth peak power of the load profile data set**

Order number (N)	Peak Power (kW) of Consumer Type									
	1		2		3		4		5	
	Measured	Synthetic	Measured	Synthetic	Measured	Synthetic	Measured	Synthetic	Measured	Synthetic
1	3.22	3.30	1.89	1.96	4.44	4.12	5.96	5.76	7.37	7.55
2	2.91	3.01	1.72	1.79	4.11	3.76	5.53	5.39	6.85	7.10
3	2.73	2.81	1.63	1.69	3.90	3.55	5.32	5.18	6.59	6.80
4	2.61	2.68	1.57	1.63	3.77	3.41	5.14	5.03	6.39	6.60
5	2.52	2.56	1.51	1.58	3.65	3.29	5.02	4.94	6.24	6.43
6	2.45	2.46	1.47	1.54	3.55	3.17	4.90	4.82	6.13	6.30
7	2.38	2.38	1.44	1.50	3.47	3.08	4.80	4.72	6.03	6.13
8	2.32	2.31	1.41	1.46	3.39	3.00	4.71	4.63	5.94	6.04
9	2.27	2.26	1.38	1.43	3.32	2.92	4.63	4.55	5.86	5.94
10	2.23	2.19	1.36	1.40	3.25	2.83	4.56	4.46	5.80	5.85

Order number (N)	Peak Power (kW) of Consumer Type									
	6		7		8		9		10	
	Measured	Synthetic	Measured	Synthetic	Measured	Synthetic	Measured	Synthetic	Measured	Synthetic
1	9.68	9.57	13.05	12.77	9.85	10.14	19.20	18.46	17.72	19.16
2	9.14	9.02	12.66	12.53	9.81	10.10	18.47	17.76	16.90	18.46
3	8.78	8.71	12.45	12.41	9.80	10.08	18.02	17.43	16.46	18.04
4	8.56	8.47	12.32	12.31	9.79	10.05	17.71	17.22	16.22	17.66
5	8.40	8.31	12.22	12.24	9.77	10.03	17.52	17.02	16.02	17.38
6	8.26	8.17	12.09	12.17	9.76	10.02	17.34	16.89	15.83	17.19
7	8.14	8.08	12.02	12.11	9.75	10.00	17.11	16.77	15.61	16.98
8	8.05	7.96	11.96	12.06	9.74	9.99	16.95	16.66	15.40	16.83
9	7.97	7.87	11.89	12.01	9.73	9.98	16.81	16.56	15.32	16.64
10	7.88	7.79	11.83	11.96	9.72	9.97	16.69	16.48	15.20	16.48

Order number (N)	Peak Power (kW) of Consumer Type							
	11		12		13		14	
	Measured	Synthetic	Measured	Synthetic	Measured	Synthetic	Measured	Synthetic
1	81.28	87.46	126.70	133.18	334.74	353.98	1094.97	1050.55
2	79.41	84.84	124.40	131.41	328.33	331.93	1083.47	1034.26
3	78.09	83.07	123.22	130.49	324.16	325.32	1078.49	1024.56
4	76.95	81.38	122.31	129.73	319.66	321.92	1074.13	1018.31
5	75.90	80.41	121.39	129.17	318.11	319.20	1068.87	1011.08
6	75.32	79.45	120.85	128.68	316.48	316.91	1066.80	1005.21
7	74.66	78.64	120.31	128.27	314.47	314.88	1063.06	1001.28
8	74.36	77.80	119.93	127.92	312.80	313.29	1060.63	997.12
9	73.98	77.17	119.41	127.58	311.88	311.78	1058.13	993.47
10	73.43	76.64	119.08	127.23	310.53	310.35	1057.16	990.65