ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ADAPTRICITY

eeh power systems laboratory

UNIVERSITÀ DEGLI STUDI DI PADOVA

DTU Technical University of Denmark

Dipartimento di Ingegneria Industriale DII
Dipartimento di Ingegneria dell'Informazione DEI

TESI DI LAUREA MAGISTRALE IN
INGEGNERIA DELL'ENERGIA ELETTRICA

Generation of customer load profiles based on smart-metering time series, building-level data and aggregated measurements

RELATORI:

Prof. Andrea Alberto Pietracaprina	Dipartimento di Ingegneria dell'Informazione (DEI) Università degli Studi di Padova
Dr. Andreas Ulbig	Adaptricity AG, Power Systems Laboratory Swiss Federal Institute of Technology (ETH), Svizzera
Prof. Pierre Pinson	Department of Electrical Engineering Technical University of Denmark (DTU), Danimarca

LAUREANDO:

Damiano Toffanin

ANNO ACCADEMICO 2015/2016

Abstract

Many countries are rolling out new devices to measure household electricity consumption, the so-called "smart meters". These devices can be remotely read and are able to measure consumption multiple times per day, usually every 15 or 30 minutes in European countries. Such abundance of data allows to achieve very detailed insight on consumption patterns and power flows in distribution networks.

Nevertheless, smart meters have not been rolled out everywhere yet: some areas of the distribution grid are only partially covered, others are not covered at all. The latter are "blind grid areas" from the perspective of the grid operator. This is likely to be the typical European situation, at least for the next decade. Therefore, it becomes important to provide an estimation on how these "blind grid areas" behave. This can be done by a proper exploitation of the previously unseen abundance of data. The research focus is to generate load profiles estimating a realistic consumption for a specific blind area.

This thesis illustrates a model based on machine learning techniques and Markov chains for generating realistic Synthesized Load Profiles (SLPs) to model consumption of buildings not equipped with smart meters. The project makes use of data of around 40'000 smart meters and 20'000 buildings in the City of Basel, Switzerland.

A missing dataset prevented the testing of the complete machine learning pipeline. Nevertheless, the focus could be successfully shifted towards the improvement a model for probabilistic data-driven generation of SLPs.

A novel notation has been developed, that allows to avoid ambiguity in the definition of features specifically engineered to characterize load profiles for machine learning approaches.

Generation of SLPs has been carried out with a data-driven model based on Markov chains that demonstrated to be very effective in reproducing the most significant properties of the input load profiles utilized to train the model. Annual consumption, distribution of magnitude of load, distribution of magnitude of peak load, distribution of the Time of Use of the peak load and autocorrelation are very well reproduced in most cases.

The model is easily scalable and parallelizable, allowing the generation of large datasets of profiles that remove issues related to privacy and allow for large-scale Monte Carlo simulations.

Contents

List of Figures

Nomenclature

Acronyms

BFset	Dataset of Building Features
BLP	Building Load Profile
CoC	Class of Consumption
HVAC	Heat Ventilation and Air Conditioning
ILP	Individual Load Profile
IWB	Industrielle Werke Basel, power utility of the City of Basel
LF	Load Factor
CF	Coincidence Factor
LPxxx	ID of a sample Load Profile
LKset	Dataset of Links between BFset and LPset
LPset	Dataset of Load Profiles
q-TLP	Quantile Typical Load Profile
RLP	Real Load Profile
SLP	Synthesyzed Load Profile
ToU	Time of Use
TLP	Typical Load Profile

Symbols used in the Markov Model

T^W	Transition matrix of the weekly model
t^W_{ij}	Entry of T^W, indicating the probability of transition from state i to state j
f^W_{ij}	Relative frequency of occurrence of transitions in the weekly model
T^w_h	Transition matrix for the intra-day model, at ToU $= h$ and week-state $= w$
$t^w_{h,ij}$	Entry of T^w_h, indicating the probability of transition from state i at ToU $= h$ to state j at ToU $= h+1$
$f^w_{h,ij}$	Relative frequency of occurrence of transitions in the intra-day model
w	Week-state. $w \in \{\text{L (Low)}, \text{M (Medium)}, \text{H (High)}\}$
h	Index for the Time of Use. i.e. $h \in \{1, 2, .., 96\}$
l	Index for the ToU-state. i.e. $l \in \{1, 2, .., 5\}$
s	Index for the ToU-sublevel. i.e. $s \in \{01, 02, .., 10\}$
M	Number of ToU-states of the intra-day model
S	Number of sublevels of the intra-day model

Chapter 1

Introduction

1.1 Motivation of the Thesis

From conventional distribution networks to "smart grids" Most of today's power distribution systems still mainly relies on aged technologies, settled more than 50 years ago. The monitoring and control of these grids relies on little or no measurement systems and automation. Distribution grids were designed to be only the final, passive branches of a power system conceived to deliver the energy produced in large, centralized and flexible power plants fueled by fossil energy sources or traditional hydro power. For the time being, the trend is to move towards a system dominated by ensembles of small, distributed and often inflexible renewable generation plants. Small power plants are mostly connected to the distribution grid and are characterized by an inflexible and highly volatile power generation that most distribution systems do not have the capacity to handle in large shares, because of aged protection systems, limited voltage control and thermal constraints on cables and overhead lines. To ensure safety, reliability and efficiency, the power locally produced by small-to-medium photovoltaic system, micro wind generation and small CHP will be needed to be managed locally. The distribution systems will have to be upgraded with a higher capacity of power lines to accommodate a higher flux of power and widespread measurement and communication systems to maintain and control the optimal operation and prevent outages. These new "smart grids" will not be passive infrastructures anymore, but will actively handle and route power, with the capacity of self-monitoring and self-healing.

From smart grids to smart meters To accomplish the ambitious goal of shifting the paradigm in distribution networks and ensure an optimal grid upgrade, it is crucial to understand how the consuming behaviors of different customers impact the states of the system. Until a decade ago, because of missing widespread measurement systems at the customer end of the grid, it was impossible to accurately estimate the power flows at the final branches of the grid. To date, many countries are rolling out new devices for measuring of household consumption, the so called "smart meters". These devices are able to measure and record the electricity consumption multiple times per day, usually every 15 or 30 minutes in European countries. The measurements are remotely transmitted to the utilities or the operators, where they can be exploited to provide a very detailed insight on

how the energy is consumed and how the power flows through the network. A previously unseen amount of data is available to improve the understanding of the grid and enhance its efficiency and robustness, by the identification of bottlenecks and the estimation of the amount of renewable generation that can be safely be accommodated. Eventually, a better understanding of the network is translated into more targeted and effective investment in grid upgrades.

Form smart meters to probabilistic Monte Carlo simulations The fluctuating nature of renewable generation increases the complexity of the tasks needed to guarantee reliability and safety [4]. The always increasing computational power available today allows system planners and operators to model and simulate the behavior of the system under many different scenarios, by tackling the probabilistic nature of phenomenons affecting the power system, e.g. weather, human behavior, random failures, etc. These probabilistic techniques, relying on repeated experiments with random samples, are labeled as Monte Carlo simulations. Because of their ability to model random behavior of residential loads and generators at a high time resolution, they are a powerful tool to assess reliability and risks [5].

From Monte Carlo simulations to the generation of synthesized load profiles Monte Carlo simulations are based on a large number of experiments, each one representing a possible and realistic state of the system. In these simulations, the wealth of data provided by smart metering systems can be fruitfully exploited. Data allow to "learn" individual residential consumption patterns and to "teach" a probabilistic model to reproduce them. These synthesized load profiles (SLPs) mimic the behavior of the original real load profile (RLP) and can be produced in large quantities and adapted to many different flavors. This solves issues related to privacy, since the original data is used to train the model only. Big datasets of synthetic load profiles can therefore be generated and released for public use.

From Synthetic Load Profiles to the correlation with buildings To date, smart metering systems have not been rolled out everywhere, in many distribution grids some areas are only partially covered, others are not covered at all. This is likely to be the typical European situation for at least the next decade. Therefore, it is valuable to model and estimate how these "blind grid areas" behave. The challenge is to generate load curves that are realistic for the specific area. Electricity consumers differ among each others. Residential consumption is determined by daily routines, type of activity and of appliances, size of the household and size of a family. There is a proven correlation between the features of a building (size, number of dwellings, heating system, etc.) and the energy consumption. The knowledge of the features of the buildings located in a "blind grid area" can provide a certain amount of insight and allow a more accurate estimation of consumption patterns. This knowledge can also be exploited to estimate and predict the consumption in a new residential area.

1.2 Goal of the Thesis

The aim of this project was to develop and validate a data-driven model to generate realistic synthesized load profiles. A further target is to exploit correlations between consumption patterns and buildings' characteristics to improve the effectiveness of the model.

This master thesis has been carried out at Adaptricity, a spin-off of ETH Zürich developing simulation and optimization software tools for adapting electric distribution grids to the transition towards renewable energies. The project is based on a case study for the City of Basel: smart metering data is provided by the local utility IWB (Industrielle Werke Basel) and buildings' data is supplied by the Swiss Federal Statistical Office.

- The Load Profiles dataset (LPset) includes time series of electric energy consumption of around 40,000 consumers in the City of Basel, spanning 13 months, from April 2014 to May 2015, with a time resolution of 15 min.
- The Building Features dataset (BFset) consists of features of residential or partial residential buildings, including:
  - Geographical coordinates, district, postal code.
  - Building category, i.e. family houses, apartment buildings, residential building with ancillary use, buildings with partial residential use.
  - Construction period.
  - Heating system, i.e. typology and fuels for space heating and hot water.
  - Structure of the building, i.e. number of floors, number of dwellings.
  - Number of residents in the building.

Disclaimer on missing data The full methodology relies on machine learning techniques (clustering, classification) and has been designed with the assumption of having a working dataset in which IDs of load profiles are linked to the IDs of buildings. The matches are contained in a dataset owned by IWB that was supposed to be provided in the second half of the project. Unfortunately, this dataset was not delivered within the programmed completion date of the thesis. For this reason, it was not possible to run and validate the model entirely. Nevertheless, all algorithms have been implemented and interlinked, even though it was not possible to test them. The algorithm for the generation of load profiles was designed to run in stand-alone mode, i.e. it only requires an input RLP to generate a set of SLPs resembling the original. The generation model could be validated independently from the rest. For these reasons, the generation model based on Markov chains is the main focus of this thesis.

Summary of the proposed methodology The following paragraph briefly summarizes the methodology, under the assumption of having all data at disposal.

The task requires the adoption of machine learning techniques and probabilistic models and can be divided into two sub-tasks:

1. **Classification of buildings:** by clustering the associated load profiles, buildings can be grouped into different Classes of Consumption (CoC). A classifier can then be trained to assign a new building to a certain Class of Consumption.

2. **Generation of load profiles:** starting from the knowledge of the Class of Consumption and the features of a blind building, a probabilistic model can be built to generate realistic load profiles. The model tackling this task has also to be able to work in stand-alone mode, i.e. it has to be able to generate realistic synthesized load profiles to mimic the properties of a real profile utilized as input.

1.3 Project outline

As aforementioned, because of a missing dataset linking the load profiles to the buildings, it was not possible to test and validate the complete methodology. Nevertheless, the thesis focused on the characterization of residential load profiles and the development and validation of a probabilistic methodology for profile generation.

A model based on Markov chains has been developed. Since the load profiles are sampled in discrete time, the model has proven to be well suited to capture the probabilities of transitions between states of consumption in adjacent time steps. A Markov model can be built to capture the behavior of each consumer, unlocking the possibility to model large populations with Monte Carlo Simulations.

The validation of the model has been carried out by comparing meaningful indicators for both the original and the synthesized profiles. Such indicators are annual consumption, experimental distribution of the load magnitude, distribution of the magnitude of the daily peak load, distribution of the hour of the day at which the peak load occurs, and temporal autocorrelation of the profiles.

According to these metrics, the model has been proved to be very effective in reproducing the characteristics of the original profiles. At the same time it allows the generation of an arbitrary number of profiles at an acceptable computational cost, thus being a valid candidate for the generation of scenarios and large datasets of profiles.

1.4 Structure of the report

Chapter 2 expands the topics of the Introduction.
Chapter 3 explores the characterization of load profiles and the indicators for validation. A novel notation is introduced.
Chapter 4 describes the methodology of the machine learning approach.
Chapter 5 discusses the preprocessing of the data.
Chapter 6 illustrates the algorithm for the generation of synthesized load profiles.
Chapter 7 presents the results and performs the validation, discussing strengths and limitations of the model.
Chapter 8 summarizes the results and outlines further developments.

Chapter 2

From smart grids to synthesized load profiles

This chapter expands the concepts and motivations presented in the introduction. The structure is maintained, concepts are presented from the broader to the most specific.

2.1 Smart grids

Explaining Smart grids The structure of the existing power grid is the result of the economic boom that took place in the second half of the last century. Technologies and procedures may slightly differ from country to country, but the underlying structure and topology are basically the same around the world, and lock-in effects determined this structure to be unchanged over time [6]. The current grid shows a highly hierarchical structure and a clear subdivision into the three areas of generation, transmission and distribution. The first two areas are being increasingly automatized and extensively monitored, but the distribution system still relies mainly on electromechanical devices with little or no automation at all. Automation in distribution grids is a recent innovation.

In the past, conventional distribution grids planning followed the "fit-and-forget" principle. When the infrastructure needed to be expanded, the maximum aggregated expected load was evaluated and the implementation followed accordingly. Afterwards, little had to be done more for the remaining lifetime of the network. The resulting system was therefore not optimized and overengineered to withstand infrequent events.

Nowadays, distributed renewable generation, electric mobility and storage are groundbreaking forces that promise to provide great benefits to society. Nevertheless, they need to be adequately managed to avoid disastrous failures. Since a great share of the new distributed capacity is expected to be connected to the medium- and low-voltage network, distribution grids have to evolve from a passive infrastructure to a flexible resource able to accommodate new technologies. The optimization requirements of the new smart grids are expected to shift the paradigm from a "fit-and-forget" to a "connect-and-manage" [7]. Resources and assets will no longer operate passively, but will have to be actively managed, to cope with the rapidly changing demand in the network. Table 2.1 summarizes the key differences between the existing distribution grid and the expected smart grid of the future.

Table 2.1: Comparison between existing grid and Smart Grid. Extract from [6].

Existing Grid	**Intelligent Grid**
Electromechanical	Digital
Mainly passive	Active
One–Way Communication	Two–Ways Communication
Centralized Generation	Distributed Generation
Hierarchical	Network
Few Sensors	Sensors Throughout
Blind	Self–Monitoring
Manual Restoration	Self–Healing
Failures and Blackouts	Adaptive and Islanding
Manual Check/Test	Remote Check/Test
Limited Control	Pervasive Control
Few Customer Choices	Many Customer Choices

Benefits The implementation of smart grids is expected to provide a broad range of benefits for customers, operators and society [8]:

- **Customers:** Information about consumption and new data visualization are likely to enable more informed choices and develop more awareness about electricity consumption, increasing the potential for energy savings and peak shaving.
- **Generation and storage:** New options for distributed energy resources and storage are expected to be efficiently accommodated, contributing to integrate more renewable energy in the system.
- **Market:** Information technology and data availability will likely stimulate the creation of new markets for energy related products and services. The electricity market itself is expected to develop a new structure to improve efficient use of resources.
- **Products and services:** New products and services will be available for consumers, increasing house automation and helping to reduce electricity bills and enhance life quality.
- **Optimization:** Large data availability is the key to optimize operating efficiency and target investments. Power will be re-routed where needed to minimize operational costs.
- **Power quality:** Distributed measurement systems will enable to improve the insight on power flows, voltage and harmonic content at the distribution level. Power quality is likely to be constantly monitored, ensuring better services for a range of identified needs.
- **Resiliency:** A higher amount of automation and an ubiquitous monitoring system will provide insight on the states of the system, allowing fast and rational decisions and enabling the network to react quicker to disturbances and identify in advance

possible faulty conditions. The network is expected to become more resilient to disturbances, attacks and natural disasters.

- **Sustainable infrastructures:** The smart grid is likely to provide a framework to boost and harmonically integrate other energy infrastructures, like district heating network, sustainable transportation.

2.2 Monte Carlo simulations

2.2.1 Reliability evaluation

As aforementioned, the future distribution grids will be characterized by a high penetration of distributed generation, storage and automation. It will be necessary to overcome a large number of challenges in the areas of generation, transmission, distribution, operation, planning, safety and environment to fulfill the task of electricity supply.

Component failures or bad management can lead to interruptions of supply or degradation of power quality below the acceptable limits. Failures, low power quality and interruption of supply can damage economic and social activities and threat safety and health of people. The effects span over an extremely wide range, from marginal effects affecting a small number of consumers or equipment to catastrophic cascade failures capable of throwing a whole country into chaos.

The probability and duration of interruptions have to be reduced as much as possible. Power quality has to be kept within the acceptable levels. Therefore, given a finite amount of financial, technical and human resources, it is of utmost importance to rationally plan, design and operate the system, accounting for uncertainties and future expansions. As it is pointed out from Billington [9], it is evident that the economic and reliability constraint can conflict and lead to difficult managerial decisions. The "energy trilemma", affordability, security of supply and low environmental impact, needs to be addressed by using a set of proper tools to support and help the decision making-process in the areas of design, planning, and operation. The reliability of a configuration of the system must be tested under different conditions, in a process known as reliability evaluation.

Probabilistic simulations are powerful tools and are widely spreading in the industry. These methods utilize probabilistic quantities as input, to provide insight on the behavior of the system. They allow to model the highly stochastic nature of customer demand, distributed generation and component failure. If proper exploited, these models can provide meaningful information to identify which circumstances may induce a component to fail and the likelihood to happen, deduce the consequences of a failure, identify the safety boundaries for the states of the system, test the robustness of the system and compare the outcomes of different system configurations to target the investments. The next section will introduce a increasingly important typology of probabilistic simulation called Monte Carlo Method.

2.2.2 Monte Carlo Method

The Monte Carlo Method (MCM) is the general designation for stochastic simulation using random numbers [1] [9]. Monte Carlo simulations are probabilistic methods that treat the problem as a series of random experiments. These methods aim to estimate the probability distribution of states and outputs of a system by computing the actual response of the system to a large number of different stochastic inputs. Because of the high computational intensity that the simulation of a large amount of random experiments requires, the extensive use of Monte Carlo methods became appealing only recently.

According to [9, 10], Monte Carlo methods have a number of advantages for the exploitation in real life systems and power system analysis:

- Effects or system processes may be included without necessity of approximation.
- The required number of samples for a given confidence level is independent from the size of the system. Therefore, the methods are suitable for large scale simulations.
- Non-electrical system factors, such as weather effects, can be simulated.
- The probability distribution of the possible outcomes can be estimated.
- MCM can be used to compute risk outcomes under a number of different model assumptions, showing their strength in scenario analysis.

Despite many appealing advantages, it is important to notice that the evaluated outcomes are only as good as the quality of both the model of the system and the quality of input data. It is therefore important to develop appropriate models to generate probabilistic quantities to fuel Monte Carlo Simulations.

2.2.3 Theoretical Background

As pointed out in [11], the Monte Carlo Methods pose their theoretical roots in two well known theorems.

Law of large numbers. The average of a sequence of independent and identical distributed random variables converge to its expectation as the numbers of trials goes to infinity. More formally, let x be a random variable taking values in the interval $[a, b]$ with probability density $p(x)$, let $f(x)$ be an arbitrary continuous function $f : \mathbb{R} \to \mathbb{R}$ and $y = f(x)$ a new random variable. Let's run n random experiments by evaluating the value of x for n times. Let $\{x_i\}$ be the set of different outcomes. Hence, the law of large numbers states that:

$$I \triangleq \frac{1}{n} \sum_{i=1}^{n} f(x_i) \quad \rightarrow \quad E[f(x)] \triangleq \int_a^b f(x)p(x)dx \tag{2.1}$$

The theorem ensures the stability of outcomes when averaging the results for a large number of trials.

[1] The method was named "Monte Carlo" after the suburb in Monaco made famous by its gambling casino. This was because the roulette wheel is an emblematic device for generating random numbers.

Central limit theorem. The sum of a large number of independent and identically distributed random variables is approximately normally distributed. More formally, for a sum of independent and identically distributed random variables z_i with mean μ and finite variance σ^2:

$$\frac{\sum_{i=1}^{n} z_i - n\mu}{\sigma\sqrt{n}} \rightarrow \mathcal{N}(0,1) \quad \text{as} \quad n \rightarrow \infty \tag{2.2}$$

It can then be shown [11] that the error ϵ is:

$$\epsilon \triangleq |I - E[f(x)]| = \frac{\sqrt{Var[f(x)]}}{\sqrt{n}} |\mathcal{N}(0,1)| \tag{2.3}$$

Properties The mathematical properties of the Monte Carlo Estimation can be summarized as follows:

- If the variance of $f(x)$ is finite, which is the common case for engineering problems, the MC estimation is consistent.
- The MC estimation is asymptotically normally distributed and asymptotically unbiased, i.e. the error of the estimations converges to zero.
- Let be σ_f the standard deviation of $f(x)$ and n the number of trials. Then the standard deviation of the MC estimate is

  $$\sigma = \frac{\sigma_f}{\sqrt{n}} \tag{2.4}$$

  Therefore, the accuracy of the method can be improved by increasing the number of samples n, although the convergence is slow.

Limits of Monte Carlo methods Some limits of MCMs are [10]:

- MCMs are computationally expensive.
- MCMs show a slow rate of convergence, especially because many real world applications are in situation which are far from "asymptotic".
- MCMs perform poorly in the simulation of rare events, for the reason that rare events do not show up in a typical simulation run.

2.3 Generation of Synthesized Load Profiles

To tackle the challenges set by distributed generation and allow the efficient and reliable integration of new technologies, load and generation has to be modeled at the very disaggregated level. A disaggregated load shows the following attributes:

- **High variability:** Consumption is very volatile at the household level. At the level of a single detached dwelling, for example, consumption can vary from almost zero to peak consumption in a few minutes.

- **High unpredictability:** Consumption of individual dwelling is extremely related to occupants habits and practices, which endorse a large share of randomness. Unpredictability characterize individual load profiles both on the short term, i.e. 15-min consumption and the medium term, i.e. daily or weekly, because of different activities or changing routines throughout time.

The ability to generate Synthesized Load Profiles (SLPs) to model this high variability and low predictability brings a number of benefits for both operators and consumers, for example:

- Identification of favorable configurations for investments and for implementation of demand response schemes.
- Ability to test control algorithm for storage or smart appliances at the residential level. Optimized control algorithms lower the overall costs and strengthen the robustness of the system.
- Ability to test reliability and adequacy of the distribution grid at the very capillary level, testing control algorithms to avoid line overloading and improve voltage control.

Kim et al. [12] point out that a large effort in research is being spent for exploiting the previously unseen amount of data provided by smart meters, in order to analyze loads and generate Synthesized Load Profile of non-automatic metered customers. Applying the SLPs to power utilities' GIS (geographic information systems) allow energy companies to improve operation and to plan for optimal grid reinforcement and expansion.

2.4 Correlation of load profiles and buildings

A good amount of studies has been conducted to find the most influencing factors on residential electricity consumption. Traditionally, the studied output variables has been macro-indicators like total annual electricity consumption, peak consumption, average peak consumption, and distribution of peak hours, i.e. Time of Use (ToU) of maximum electricity consumption. Input variables are usually socio-economic quantities and can be grouped into three main categories:

- **Technical features of the buildings**: dwelling type, number of rooms, size, insulation, year of construction, heating system, presence of air conditioning, water heating system, etc.
- **Appliances:** type of cooking facilities, presence and energetic class of washing machine, dishwasher, tumble dryer, oven, refrigerator and other different smaller appliances, etc.
- **Social Features:** number of occupants, age of occupants, age of the head of household (HoH), income, education level, ethnicity, home ownership (rented/owner), etc.

McLoughlin [13], in an Irish case study from a representative cross section of 4'200 domestic Irish dwellings, found that the maximum electricity demand is strongly influenced by householod composition, water heating and cooking type and the type of appliances installed. Conversely, the time of peak demand has been found to be correlated mainly with social variables: income, age of HoH and household composition. Young people has been (reasonably) found to be more inclined to consume late in the evening than older people. It is argued that this results are very country-specific, since in countries where it is common to posses electric heating systems and air conditioning, electricity consumption and peak times become also influenced by external temperature and weather conditions. Total electricity consumption has been found to be strongly correlated to the size of the dwelling, or the number of bedrooms, that can be used as a proxy for the size when this is not available. Apartment dwellings, being smaller and with less occupants and appliances, consumed less electricity compared to other dwelling types. The social class of the occupants significantly influenced the total electricity consumption, reflecting a possible income effect, with higher professionals consuming more than middle or lower classes.

Few studies explored the specific influence of the behavior of occupants, by analyzing energy consumption in identical dwelling. The results show variation in consumption up to 300% - 400% [14].

Gram-Hanssen [15], using a socio-technical approach to a case study of 30 identical houses in Albertslund, a suburb of Copenhagen, Denmark, highlights significant variation in consumption between similar buildings, due to different usage of both house and heating system. Interviewing the occupants, he uncovered the great complexity and variability in practices of even a simple practice like regulating the indoor climate or using electric cooking facilities. Measurements illustrated that different user behavior in identical houses may result in three times higher electricity consumption for heating.

Lutzenhiser [16] from a study of 1'627 Northern Californian household, reported a strikingly high variability in annual electricity consumption among the samples of the set, the distribution of consumption is highly skewed and the most consuming samples recorded over 30 times the consumption of the least consuming. A large proportion of the observed variation has been found to be explained by a relatively small set of variables, including: climate zone/temperature, dwelling type and size, building age, home ownership, household income, ethnicity and household composition. It must be pointed out that relations between indicators/predictor variables are not simple. Not unexpectedly, in fact, many of the predictors have significant correlations within each other, e.g. income with dwelling size, household composition with building type.

In a 2004 study, Gram-Hanssen [17] related yearly electricity consumption of 50,000 households in Aalborg (Denmark) with features of buildings and occupants. He found that indicators and predictors only explain a fraction of the variation found within the sample. The simple most important indicator is the number of persons living in the dwelling, followed by income and floor area. If accounted together, the number of persons, floor area and income level explain between 30% and 40% of the total variation in electricity use in the three different types of housing, which also means that 60%-70% of the variation in electricity is not explained by these variables.

The same author in 2002 [18] noticed that considerable differences in energy use can be

found even when considering similar households and removing the expected influence of all relevant variables, implying, once again, how patterns of activities vary among different individuals, even with similar lifestyles.

Morley [19], analyzed electricity consumption of three blocks of student apartments of a UK university, with similar social groups and homogeneous age. The study focused upon electricity that is not used for space heating and cooling, with large fixed appliances being equivalent across households (no tumble dryers, no washing machines, no dishwashers). It has been noticed that even in this very homogeneous configuration, the standard deviation within the sample set is between 12% and 24% of the mean, the ratio between maximum and minimum consumption is between 1.5 and 3.4.

Correlation between building features and daily consumption patterns Most of the cited studies on correlations between residential electricity consumption and socio-economic features of dwellings and occupants focus mainly on the relations between socio-economic macro-indicators and only three aggregated measurements of electricity consumption: total annual electricity consumption, peak demand and time of use of peak demand. This choice is probably influenced by the difficulty in the past to obtain more detailed and disaggregated measurements of electric load.

In this context, this work aimed at filling a gap in the literature, by developing a model able to link building features and patterns of consumption.

Chapter 3

Characterization of residential load profiles

A load profile is a time series of electrical load[1]. A critical approach to the generation of and validation of customer load profiles necessitates a good understanding of the basic dynamics governing their generation in real life. It also required to reflect upon what are the most important characteristics of a residential profile, from the point of view of the distribution network. After a brief survey on the main shaping dynamics for patterns of consumption, this chapter focuses on the analysis of the most significant parameters for the characterization of load profiles and the validation of synthesized profiles. For the sake of simplicity, the discussion focuses on purely residential profiles, but it is easy to generalize to profiles of offices and small commercial activities.

3.1 On the nature of residential load profiles

A residential load profile is the result of the aggregation of load of many different operating appliances and devices providing energy services to the occupants. Each appliance has an individual load profile, commonly depending on the user practice and the operation mode. For example, consumption of electricity for cooking depends on user behavior, which set time, intensity and number of electric stoves utilized. Therefore, some preliminary features and properties of a typical load profile can be inferred directly by reflecting upon the dynamics of human activities.

Domestic load profiles are usually cyclical, typically showing a morning and evening peak and a small base load overnight. Residential profiles normally show weekly cycles with 24-hour long sub-cycles. As it is easy to suppose, homologous days of different weeks tend to share similarities in consumption patterns. Depending on the typology of household and the degree of electrification in appliances, electricity consumption and load patterns can be also dependent on the season and the external temperature.

[1] In this thesis, it was decided not to adopt the term "power" to indicate the value of a data point of a load profile, because of the possible ambiguity of this term. In fact, smart meters typically do not provide measurements of power, intended as instantaneous power, but of energy, e.g. samples of 15-min energy consumption. For these reasons the more suitable terms "load" or "demand" are utilized.

A load profile is shaped by the on/off switching of individual electrical appliances, which in turn is influenced by the practices of occupants and the characteristics of the dwelling. The relevance of the contribution of the single appliance to the aggregated profile depends on various factors: frequency of utilization, magnitude of the load, average duration of utilization. The regularity of utilization is also an important aspect: for example, some appliances typically have cyclical utilization patterns, while other may appear to be switched on and off at random [20]. The next section will focus on the analysis of the nature of load profiles, to set a solid and intuitive base for the further analysis, highlighting the mechanisms shaping the profiles. With this mechanisms in mind, it will be easier to understand which characteristics can be expected from a residential load profile and which ones cannot.

3.1.1 Residential appliances and utilization patterns

Household appliances can be grouped in different categories according to magnitude of energy consumption and frequency of utilization. Dickert[1] describes four traditional categories:

- **Major appliances, or "white goods":** major domestic appliances accomplishing routine housekeeping tasks such as cooking, food preservation, cleaning, heating and air conditioning.
  - Kitchen appliances: stove, refrigerator, freezer, dishwasher, microwave, etc.
  - Laundry appliances: washing machine, laundry dryer, drying cabinet, etc.
  - Heating, Ventilation and Air Conditioning (HVAC): air condition, water heater, etc.
- **Consumer electronics, or "brown goods":** relatively light electronic consumer goods. Typical features of these appliances are changing quickly, following the fast innovation in the sector.
  - Office/Communication equipment: PC, LCD, printer, scanner, phone, etc.
  - Entertainment electronics: TV, CD/DVD player, hi-fi system, etc.
- **Small appliances:** In comparison to brown and white goods small appliances are portable or semi-portable. Many small appliances are kitchen appliances as well as for personal care.
  - Kitchen appliances: kettle, toaster, blender, coffee machine, etc.
  - Household appliances: fan, iron, sewing machine, vacuum cleaner, etc.
  - Electronic devices: mobile phone, tablet, laptop, radio, etc.
  - Personal care: hair dryer, curling iron, shaver, electric toothbrush, etc.
- **Lighting:** The illumination of homes during twilight and night as well as of windowless rooms or workplaces is also an important part of residential energy consumption. Due to energy-saving lamps their contribution to the overall consumption is decreasing.

- A new emerging category of electric goods are devices for electric mobility, such as electric bikes and cars. Given the expected magnitude of the load demand and the daily frequency of use, this category is expected to completely reshape consumption patterns in the not-too-far future.

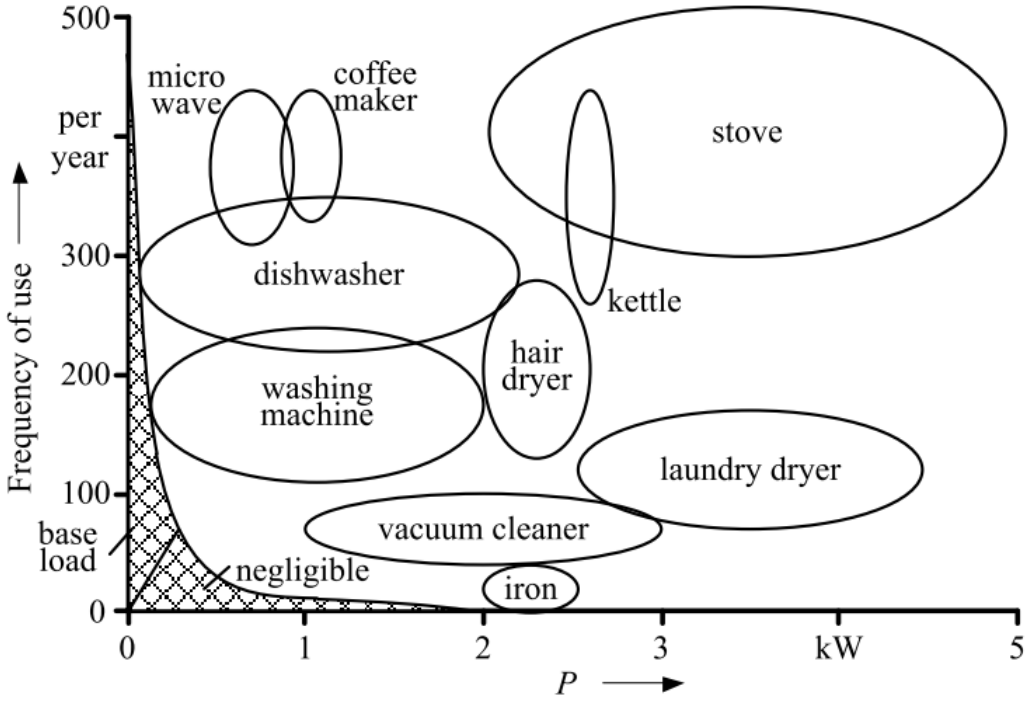


Figure 3.1: Sample of household load profile. Extract from [1].

Figure 3.1 displays the consumption magnitude of appliances versus the frequency of use per year of an average arbitrary customer [1]. The picture highlights how the main contributors to the household energy consumption are white goods, characterized by power requirements and high frequency of use.

Once the magnitude of consumption and frequency of utilization has been defined and described for the different groups, the predictability of such loads becomes another important parameter to understand. Wood and Newborough [21] used three characteristic groups to explain electricity consumption components in the home: “predictable”, “moderately predictable” and “unpredictable”.

- **Predictable load** consists of small cyclic loads occurring when a dwelling is unoccupied or all the occupants are asleep. For example, small cycles from refrigeration appliances and steady loads from security lighting and items on standby, such as TVs.

- **Moderately predictable load** relate to the habitual behavior patterns of the residents. For example, many people watch TV programmes at regular times each

day/week and switch lights on/off each weekday morning as they rise and then leave for work.

- **Unpredictable load** describes the majority of domestic energy use. It tends to be irregularly occurring at the users discretion, for example when the occupant wants to cook food or operate the dishwasher or the washing machine.

Appliances in standby mode or running continuously tend to form a small base load, while spikes in consumption are due to appliances with higher power requirements, such as stoves and washing machines. As aforementioned, consumption related to these appliances are distributed more randomly in time, amplitude and duration, with higher probability of use in localized periods of the day, such as early morning or evening. Although energy intense white goods might compose the majority of the energy consumption in a household, it can be unwise to completely neglect the coincidental effect on power request of relatively small loads, when strong behavioral components are involved. If many loads are synchronized, this may lead to important surges in power consumption. A famous example is the surge in power demand during the half-time break or at the end of important football matches in Britain, when thousands of spectators turn on the kettle for tea or coffee brewing at the same time, causing relevant spikes of additional load, with highest ever spike of 2'800 megawatts at the national level, equivalent to 1.1 million kettles, recorded after England lost the 1990 World Cup semi-final penalty shootout against West Germany.

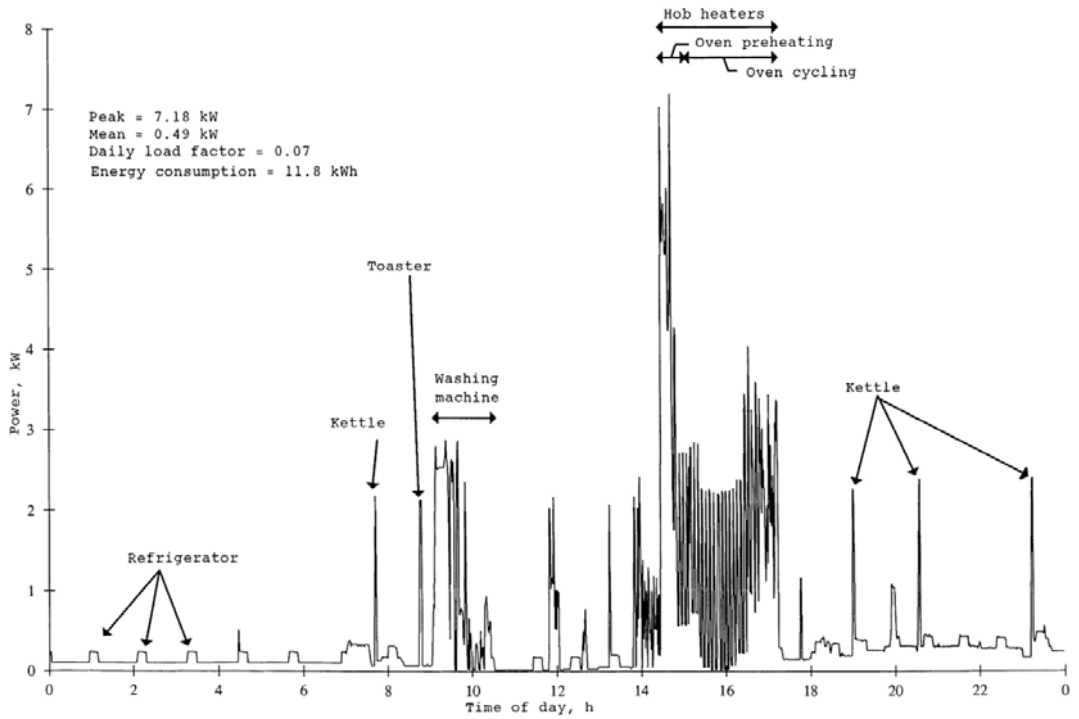


Figure 3.2: Example of a residential load profile from an individual household recorded on a 1-min time base. Extract from [2].

Figure 3.2 reports a sample 1-day load profile for a household with a 1-min sampling. The contribution of the individual appliances is highlighted. Some small appliances such as kettle or toaster can significantly contribute to peaks in power request, even if their share on total energy consumption is limited. When the load measurement has a lower time resolution, for example 15 minute, spikes are averaged, leading to a relatively smoother profile, as illustrated by Figure 3.3.

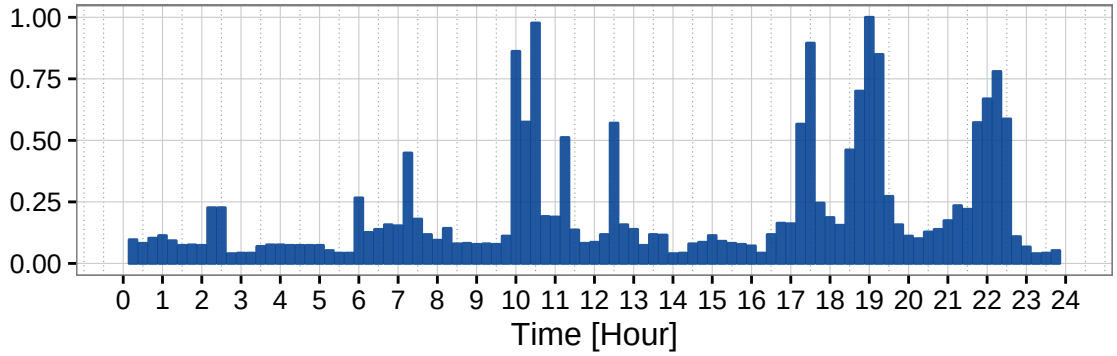


Figure 3.3: Example of a normalized residential load profile from an individual household recorded on a 15-min time base. Extract from Basel dataset.

The previous discussion depicted the level of volatility that it is reasonable to expect in a residential load profile. Magnitude of load, magnitude of peak load and the timing of spikes are normally very volatile and unpredictable. Therefore, it is fundamental to characterize these quantities in terms of probabilistic distributions.

3.1.2 Seasonal variability and influence of temperature

In countries with a high penetration of electric heating systems and air conditioning there is a clear correlation between temperature and daily consumption.

The correlation is typically dependent on the magnitude of the temperature. In a study involving 1000 meters from Canadian houses, Arkadanian at al. [22] found different coefficients of correlation, depending on the range of temperature:

- At low external temperature, consumption was found to be negatively correlated with temperature, with higher consumption at lower temperatures, reflecting the effect of heating systems.
- At intermediate temperature, consumption was found to be temperature independent, since almost no heating or air conditioning system is operating.
- At warmer temperatures, consumption was found to be be positively correlated with temperature, because of air conditioning systems.

The grade of correlation may obviously vary depending on the typology of appliances installed. All else being equal, the higher the penetration of electric HVAC systems in the household, the stronger the correlation. Similarly, insulation of the building play the

opposite role, with improved insulation leading to lower correlation of consumption with temperature.

Some countries are characterized by a low penetration of electric HVAC systems (Heat Ventilation and Air Conditioning) and electric water boilers. Concerning the frame of this thesis, in the City of Basel, very low level of penetration of electric devices for HVAC has been observed. Less than 1% of buildings are equipped with electric heat pumps or electric heating systems. Electric water boilers account for only 21.9% of penetration. See Figure 5.2 in Chapter 5 for a more detailed overview.

3.2 Characterization of profiles

Characterization of load profiles is a systematic identification of descriptive indicators and features that allows a rigorous comparison and classification of profiles. The identification must be carried out with an eye on the physical and engineering impact of profiles on the network. Key features and indicators are the ones allowing to recognize and group together profiles that affect the network in a similar way. An effective method of characterization is extremely important to properly validate load profile generators and develop performing machine learning approaches.

The terminology for the definition of indicators should avoid any ambiguity. The following section illustrates a novel rigorous notation developed for this purpose.

3.2.1 Notation

The definition of indicators for load profiles involves two dimensions:

- The time resolution of electric load to be considered. E.g. 15-minutes, one hour, one day, one week, etc.
- The time window (i.e., the subset of samples) to be considered for the evaluation. E.g. one day, one week, every day 6am–9am, every Monday, 1pm–6pm during summer, etc.

This distinction leads to different possible "flavors" of the same indicator. An illustrative example is given in table 3.1 for the peak load. The same applies also to other common indicators, such as standard deviation, median and quantiles of load distributions, etc.

Table 3.1: Different "flavors" of peak load.

		Time resolution	
		15-min load	daily consumption
Window	1 day	daily peak 15-min load	n.d.
	January	January peak 15-min load	January peak daily load
	1 year	yearly peak 15-min load	yearly peak daily load
	1 year, 8pm	yearly-8pm peak 15-min load	n.d.

To obviate this problem of ambiguity, it was developed a particular notation that take into account the two defining dimensions. Speed of reading has been sacrificed for the sake

of clarity and precision. After the definition of the structure of the notation (def. 3.1), a few examples will clarify the concept.

Let's first define:

E^{δ}	Set of energy measurements of time resolution δ. Hereafter, the term E^{δ} is referred to as "δ-energy", "δ-load" or "δ-demand" and expresses energy consumption in time slots of size δ. For example, $E^{15\text{min}}$ is referred to as "15-min load" and indicates consumption of 15 minutes; $E^{1\text{day}}$ is referred to as "daily-load" and indicates consumption of 1 day.
δ	Time resolution of the energy measurements.
f	Function to be applied. e.g. mean, peak, standard deviation, 3^{rd} decile, etc.
T	Time window defining the subset of samples to which f is applied. e.g. 1 year, 1 day, weekends, mornings, etc.

The notation will be of the form:

$$E^{\delta}|_{T}^{f} \tag{3.1}$$

The convention is to name quantities first by the time window of evaluation T, then by the function evaluated f, finally by the level of time resolution of the load δ.

Examples:

$E^{1\text{day}}\|_{1\text{year}}^{\text{peak}}$	"Yearly-peak daily-load": the highest value evaluated over one year of the daily consumption.
$E^{15\text{min}}\|_{1\text{day}}^{\text{peak}}$	"Daily-peak 15-min-load": the highest value evaluated over a day (daily-peak) of 15-min demand.
$E^{1\text{hour}}\|_{1\text{day}}^{\text{peak}}\|_{1\text{year}}^{\text{avg}}$	"Yearly-mean daily-peak hourly-load": the mean evaluated over one year (yearly-mean) of the highest value for each day (daily-peak) of consumption with one hour resolution (hourly-load).

Time of Use (ToU) The Time of Use (ToU) is the index of the time slot of the day at which a measurement is taken. For example, assuming to measure energy consumption

every 15 minutes, one day would be composed of 96 time slots: ToU = 1 indicates the time slot 00:00 – 00:14, ToU = 2 indicates the time slot 00:15 – 00:29, etc.

In this thesis the ToU refers to a 15-min time slot of the day, i.e. ToU $\in \{1, 2, .., 96\}$. In general, and depending on the context, the Time of Use can also indicate the hour of the day, i.e. ToU $\in \{1, 2, .., 24\}$, the day of the week, i.e. ToU $\in \{1, 2, .., 7\}$, the day of the year, i.e. ToU $\in \{1, 2, .., 365\}$, etc.

3.2.2 Meaningful parameters for load profile characterization

This section reviews some parameters that can be utilized for characterization of load profiles. It has been conceived to define the indicators utilized for load profile validation and as an overview of the approach to features engineering for load profiles, setting the base for further works. Some of the indicators are currently used in literature, others are original. Not all the listed parameters had the chance to be used in the project, because it was not possible to run the complete machine learning algorithm and test the characterization performance of each one.

Hereafter, unless stated otherwise, all quantities are evaluated over one year and refer to a single household.

Total Annual Electricity consumption Because of its aggregated nature, it is probably the most important parameter for consumer characterization. It is a well established parameter for distribution grid planning [13].

$$E^{tot} \tag{3.2}$$

Mean load Mean energy demand, considered over the time interval T.

$$E^{\delta}|_{T}^{\text{avg}} \tag{3.3}$$

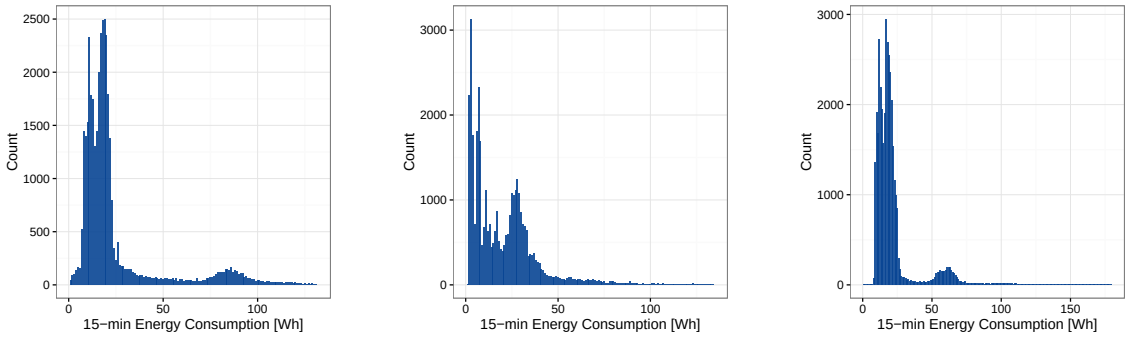
Common usage of Mean Load include:

- Mean 15-min load:
$$E^{\text{15min}}|_{\text{1year}}^{\text{avg}} \tag{3.4}$$
Yearly average of 15-min consumption.

- Mean daily load:
$$E^{\text{1day}}|_{\text{1year}}^{\text{avg}} \tag{3.5}$$
Yearly average of daily consumption.

- Mean ToU 15-min load:
$$E_{\text{ToU = i}}^{\text{15min}}|_{\text{1year}}^{\text{avg}} \tag{3.6}$$
Yearly average of energy consumption evaluated for each 15-min Time of Use i. For example, $E_{\text{ToU=57}}^{\text{15min}}|_{\text{1year}}^{\text{avg}}$ is the yearly average of all 15-min energy consumption in the 57^{th} time slot of the day (ToU = 57).

Load histogram The load histogram is an important fingerprint for characterization [23]. It is a representation of the distribution of load magnitude. It reports on the x-axis the load magnitude and on the y-axis the observed frequency of occurrence, absolute or relative. Usually, two or more modal peaks are present. Normally, a residential load histogram is hard to match with any known probabilistic distribution. In the literature, when an analytic form is required, the load histogram is usually modeled as a superposition of several density functions (Gamma-distr., Beta-distr., Weibull). Figure 3.4 reports three load histograms, extracted from load profiles of the Basel dataset.



(a) (b) (c)

Figure 3.4: Load Histogram of three sample load profiles of the Basel dataset. For the ease of visualization, only the distribution of the lowest 99% of the data in displayed, since the highest 1% load magnitude may differ of over an order of magnitude from the median.

As it is possible to notice, many modal peaks are present and distributions are highly skewed.

The width of the bins is an important parameter: bins that are too narrow or too wide might not reveal the shape of the distribution. Moreover, for purposes such as clustering, too many bins could lead to a harmful decrease of the efficiency of the algorithm. The choice of the proper number of bins has to be tailored for the specific dataset and algorithm.

Load histograms can also be used to characterize the distribution of other quantities, for example the Time of Use of the peak load, the magnitude of the daily peak load, etc.

In this project, various typologies of load histograms have been extensively utilized for the validation of synthesized profiles.

Duration curve The concept of duration curve is closely related to the load histogram. In fact, just like the load histogram is the discrete version of the probability density function (pdf), the duration curve is the discrete version of the cumulative distribution function (cdf). The duration curve expresses on the y-axis the percentage of time during which the load is lower or equal to the magnitude expressed on the x-axis. For this reason, it is completely equivalent to characterize the load profile with the load histogram or the duration curve, since the two representations can be obtained one from the other.

Autocorrelation Load is typically correlated within neighbor time slots. This short-term correlation is caused by the nature of household appliances. Load profiles usually

reveal strong autocorrelation at short lags, i.e. less than 2 hours, and peaks of autocorrelation when the lag is a multiple of 24 hours. If mean and variance of load magnitude are constant over time, the autocorrelation is only a function of the time lag and not also of the point of the time series at which it is evaluated. This is the case for load profiles that are not strongly dependent on the temperature.

The discrete autocorrelation R at lag l for the discrete-time signal $x(n)$ evaluated over a time-window T is defined as:

$$R(l) = \sum_{n \in T} x(n)x(n-l) \tag{3.7}$$

A more detailed discussion, combined with analysis of the Basel data is carried out in section 3.3.

Typical Load Profile (TLP) It is a load profile representing the annual mean load for every Time of Use [24].

$$\mathrm{TLP}^{\mathrm{15min}}_{\mathrm{1day}}(h) = E^{\mathrm{15min}}_{\mathrm{ToU}=h}|^{\mathrm{avg}}_{\mathrm{1year}} \qquad h \in \{1, .., 96\} \tag{3.8}$$

$$\mathrm{TLP}^{\mathrm{1hour}}_{\mathrm{1week}}(h) = E^{\mathrm{1hour}}_{\mathrm{ToU_w}=h}|^{\mathrm{avg}}_{\mathrm{1year}} \qquad h \in \{1, .., 168\} \tag{3.9}$$

$$\tag{3.10}$$

Figure 3.5 displays the daily TLP (Equation 3.8) from a random sample from the Basel Dataset. Figure 3.6 displays the weekly TLP (Equation 3.9) for the same profile.

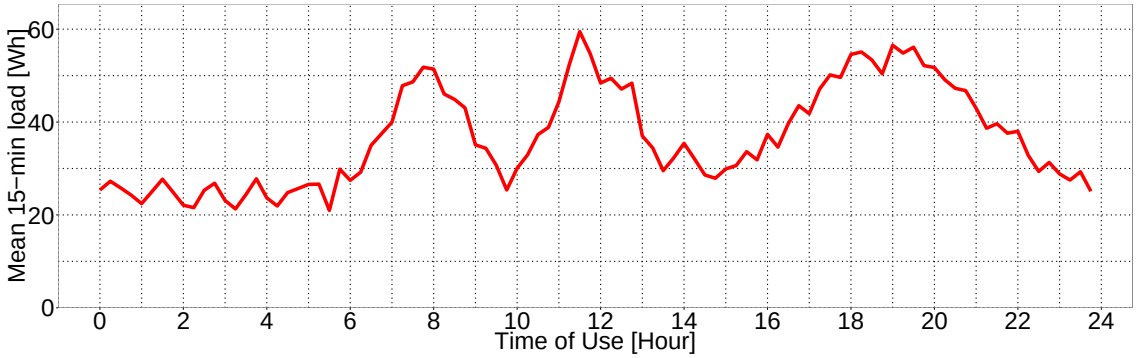


Figure 3.5: Example of a daily TLP (Equation 3.8) of a load profile randomly sampled in the Basel dataset.

TLPs are utilized for planning and engineering studies. They are used for transformer rating selection and management, for load diversity evaluation, and to determine the expected profiles at aggregated level in distribution networks [25].

Typical Load Profile are also been utilized as inputs for clustering techniques oriented to customer segmentation, for example in Kim [12], although, given the high number of points from which it is constituted, the direct use as features for clustering can lead to sub-optimal results [26].

In this thesis, TLP was utilized for validation of synthesized load profiles.

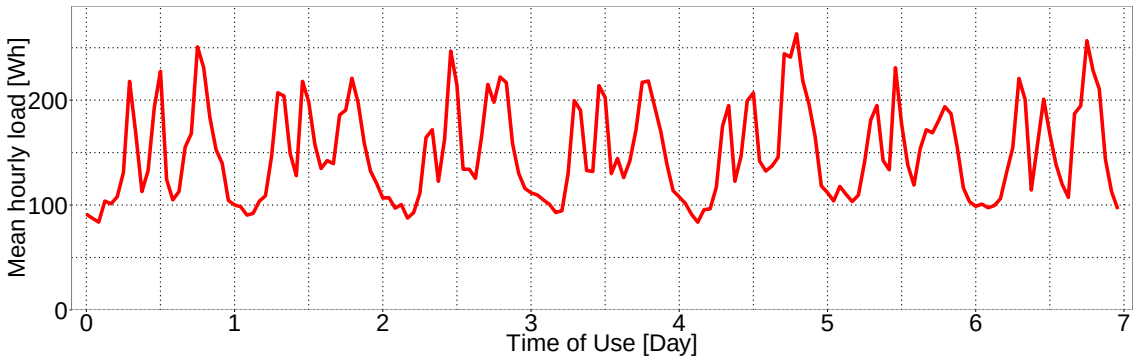


Figure 3.6: Example of weekly TLP (Equation 3.9) of a load profile randomly sampled from the Basel dataset.

Quantile Typical Load Profile (q-TLP) A quantile Typical Load Profile is a generalization of the Typical Load Profile. The TLP is a time series taking the value of the average load of each ToU. In this work, the concept has been generalized by introducing the q-TLP, being defined as a time series taking the values of a chosen quantile of the load distribution of each ToU. q-TLPs can be displayed in a ribbon chart as in figure 3.7.

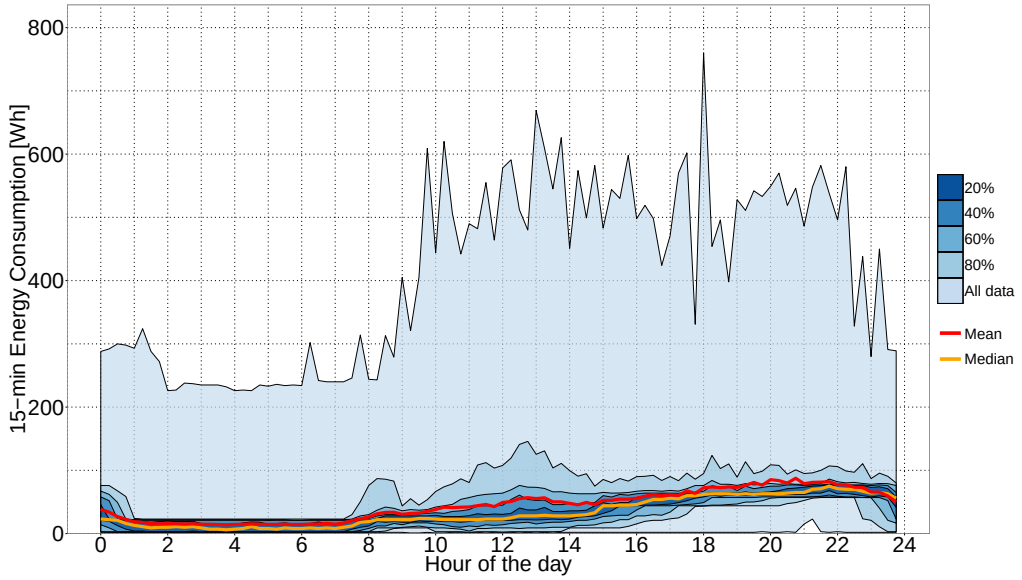


Figure 3.7: Example of q-TLP from the Basel Dataset, with colored bands determined by deciles of the load distribution, by Time of Use.

In the figure, each band of blue represents 10% of the data points, with lighter shades corresponding to the external tails of the load distribution. The profile in red is the TLP, while the one in yellow is the median profile. Ribbon charts of q-TLPs (as the one in Figure 3.7) are powerful tools to graphically illustrate the range of load variability of a load profile, that cannot be characterized by a simple TLP.

Furthermore, q-TLPs charts are effective for preliminary validation of SLPs against

the corresponding RLP. In fact, if two load profiles have very similar q-TLP charts, that ensures that other significant quantities, such as the mean load and the distribution of load magnitude, are very close for the two cases.

What cannot be inferred from two identical q-TLP charts is whether the autocorrelation of the two profiles is similar and if the two distributions of the magnitude of daily peak load and of the ToU of the daily peak load are actually matching. This must be checked separately.

Charts of q-TLPs have been used in this thesis for a preliminary qualitative validation of SLPs.

Peak load Maximum energy demand, considered over the time interval T.

$$E^{\delta}|_{T}^{\text{peak}} \tag{3.11}$$

Common usage of peak load include:

- Daily-peak 15-min load:
$$E^{15\text{min}}|_{1\text{day}}^{\text{peak}} \tag{3.12}$$
- Yearly-peak 15-min load:
$$E^{15\text{min}}|_{1\text{year}}^{\text{peak}} \tag{3.13}$$
- Yearly-peak daily load:
$$E^{1\text{day}}|_{1\text{year}}^{\text{peak}} \tag{3.14}$$

Mean peak load Is the mean evaluated over the period T_2 of the peak consumption evaluated for each time window of size T_1.

$$E^{\delta}|_{T_1}^{\text{peak}}|_{T_2}^{\text{avg}} \tag{3.15}$$

Common usage of mean peak load include:

- Yearly-mean daily-peak 15-min load:
$$E^{15\text{min}}|_{1\text{day}}^{\text{peak}}|_{1\text{year}}^{\text{avg}} \tag{3.16}$$

Time of Use of peak load It is the ToU of occurrence of the peak δ-load, evaluated over the time period T.

$$E^{\delta}|_{T}^{\text{ToUpeak}} \tag{3.17}$$

Common usage of Time of Use of peak load include:

- ToU of daily-peak 15-min load:

$$E^{15\text{min}}|_{1\text{day}}^{\text{ToUpeak}} \tag{3.18}$$

  ToU of the peak 15-min load

- Yearly-mode of ToU of daily-peak 1-hour load:

$$E^{1\text{hour}}|_{1\text{day}}^{\text{ToUpeak}}|_{1\text{year}}^{\text{mode}} \tag{3.19}$$

  Most common peak hour of the day, taken over a year.

- ToU of yearly-peak daily load:

$$E^{1\text{day}}|_{1\text{year}}^{\text{ToUpeak}} \tag{3.20}$$

  Day of the year of the peak daily load.

Load Factor It is the ratio between the average energy consumption and the peak energy consumption, evaluated over the time window T. It gives information on how much peaked is the profile.

$$LF_{T}^{\delta} = \frac{\text{avg } \delta\text{-load over period } T}{\text{peak } \delta\text{-load over period } T} = \frac{E^{\delta}|_{T}^{\text{avg}}}{E^{\delta}|_{T}^{\text{peak}}} \tag{3.21}$$

Mean Load Factor

$$LF_{T_1}^{\delta}|_{T_2}^{\text{avg}} \tag{3.22}$$

Common usage of the mean load factor include:

- Yearly-mean daily 15-min load factor:

$$LF_{1\text{day}}^{15\text{min}}|_{1\text{year}}^{\text{avg}}$$

3.2.3 Features engineering for load profiles

Domingos [26] highlights the fundamental importance of engineering meaningful features to capture important properties of data. In machine learning approaches, this is particularly important to allow the learning of interrelations and rules of classification. He points out:

"Often, the raw data is not in a form that is amenable to learning, but you can construct features from it that are. This is typically where most of the effort in a machine learning project goes. It is often also one of the most interesting parts, where intuition, creativity and "black art" are as important as the technical stuff. [...] Machine learning is not a one-shot process of building a data set and running a learner, but rather an iterative process of running the learner, analyzing the results, modifying the data and/or the learner, and repeating."

This section surveys an approach of engineering of features for load profile characterization, developed by Haben [3], to be applied on machine learning purposes, mainly clustering and classification. The exemplified process can be adapted to suit different needs.

Haben [3] exploited one year of data from an open Irish dataset of smart meter measurements of roughly 4000 Irish residential customers, sampled at time intervals of 30-min. He found that meaningful information for characterization of customers in the dataset is given by the average consumption in four time periods, namely:

Overnight period	(T_1)	: 22:30 - 06:30
Breakfast period	(T_2)	: 06:30 - 09:00
Daytime period	(T_3)	: 09:00 - 15:30
Evening period	(T_4)	: 15:30 - 22:30

The periods has been chosen according to the modal peaks in the load distribution, as it is possible to notice in Figure.

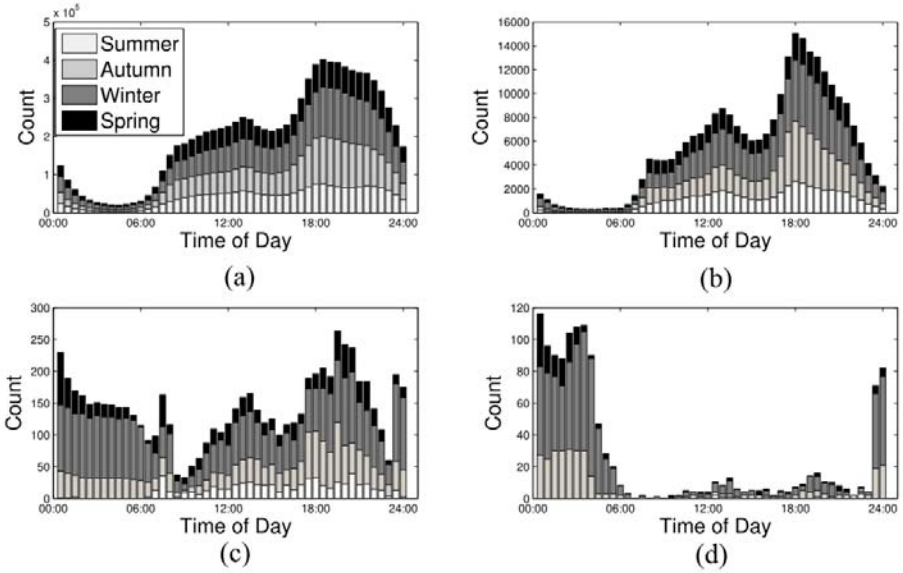


Figure 3.8: Count of number of half hours in the data set which exceed consumption of (a) 0.9, (b) 4.1, (c) 8.1, and (d) 10.7 kWh with respect to the time of day. The counts are broken down according to season. Extract from [3].

Haben [3] also defined other attributes, that were judged significant to characterize the load profiles for clustering purposes. he makes use of some preliminary definition (symbols are coupled with the newly introduces notation, for the sake of clarity):

$\hat{P}$	$E^{30\text{min}}\|^{\text{avg}}_{1\text{year}}$	yearly-mean 30min-load
M		set of time periods, $M \in \{T_1, T_2, T_3, T_4\}$
σ_i	$E^{30\text{min}}\|^{\text{avg}}_{T_i}\|^{\sigma}_{1\text{year}}$	yearly standard deviation of average 30min-load, for each time period T_i, $i \in \{T_1, T_2, T_3, T_4\}$
P_i	$E^{30\text{min}}\|^{\text{avg}}_{T_i,1\text{year}}$	yearly-mean 30min-load, for each time period T_i
$P_i^{\text{WE}}, P_i^{\text{WD}}$	$E^{30\text{min}}\|^{\text{avg}}_{T_i\text{ ,WE, 1year}}$	yearly mean of weekend (WE) and weekday (WD) 30min-load, for each time period T_i
$P_i^{\text{W}}, P_i^{\text{S}}$	$E^{30\text{min}}\|^{\text{avg}}_{T_i\text{ ,S, 1year}}$	mean summer (S) and winter (W) 30min-load over the entire year, for each time period T_i

Using the notation on the left column, Haben [3] engineered a total of seven features to be used for load profile characterization and clustering, namely:

- The relative average 30min-load in each time period over the entire year, P_i^R (4 features):
$$P_i^R = \frac{P_i}{\hat{P}} \quad i \in \{T_1, .., T_4\} \tag{3.23}$$
- Yearly mean relative standard deviation, $\hat{\sigma}$:
$$\hat{\sigma} = \frac{1}{|M|} \sum_{i \in M} \frac{\sigma_i}{P_i} \tag{3.24}$$
- A "seasonal score", S:
$$S = \sum_{i \in M} \frac{|P_i^{\text{W}} - P_i^{\text{S}}|}{P_i} \tag{3.25}$$
- A "weekday versus weekend difference score", W:
$$W = \sum_{i \in M} \frac{|P_i^{\text{WD}} - P_i^{\text{WE}}|}{P_i} \tag{3.26}$$

A small number of well engineered features has been proved to provide more significant results in machine learning approaches compared to a large number of raw features. It also improves computational efficiency and the robustness of the final outcome [26].

3.3 Effects of aggregation

When many residential load profiles are aggregated, the resulted profile shows properties that are significantly different from the ones of the individual components. As a rule, the aggregated profile is smoother and less volatile, with a higher autocorrelation and a more uniform distribution of load. Different typologies of real loads and customers show different coincidence factors. Therefore, synthesized profiles tailored on different typologies of customers has to respect the different aggregation properties. Else, the aggregated effects on the network might excessively diverge from the reality.

The aggregation effect is measured by the so-called coincidence factor, that measures how effectively peaks and valleys of individual profiles balance out at the aggregated level.

Coincidence factor The coincidence factor CF is defined as the ratio between the peak of the aggregated load and the sum of the peaks of n individual loads within a specified time period T:

$$CF(n,T) = \frac{\text{max. coincident load of n households}}{\Sigma \text{ peak load of every household}} \tag{3.27}$$

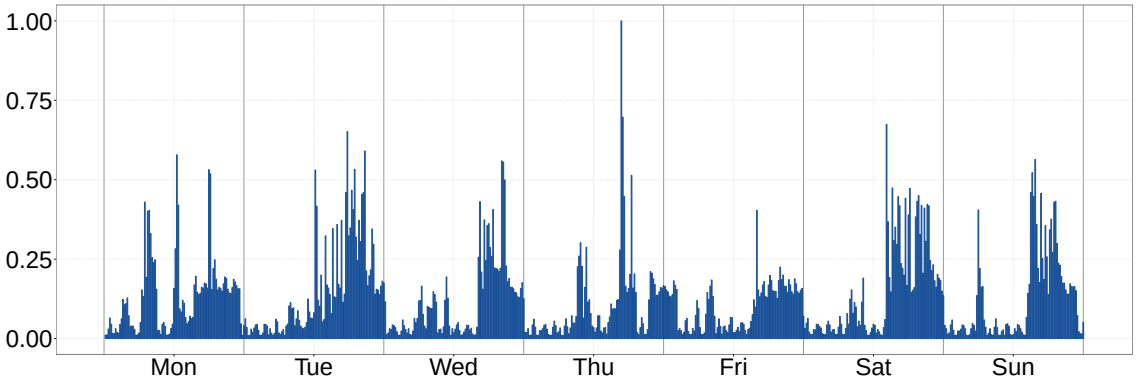
In the discrete-time context of this thesis, given n consumers, each demanding a load $E_i^\delta(h)$ at time h, it is:

$$CF^\delta(n,T) = \frac{\max_{h\in T} \sum_{i=1}^{n} E_i^\delta(h)}{\sum_{i=1}^{n} \max_{h\in T} E_i^\delta(h)} \tag{3.28}$$

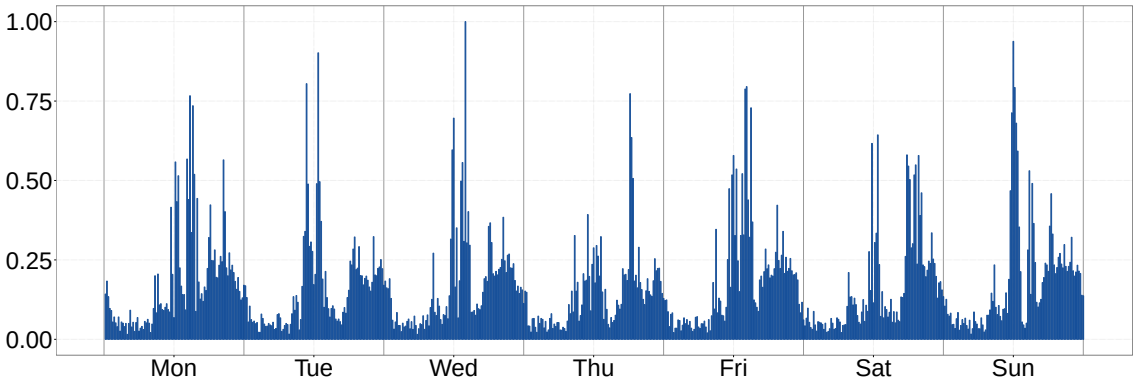
The next paragraph surveys the effects of aggregation for typical quantities: 15-min load, autocorrelation up to 10 days, load factor and coincidence factor.

3.3.1 Examples of aggregated load

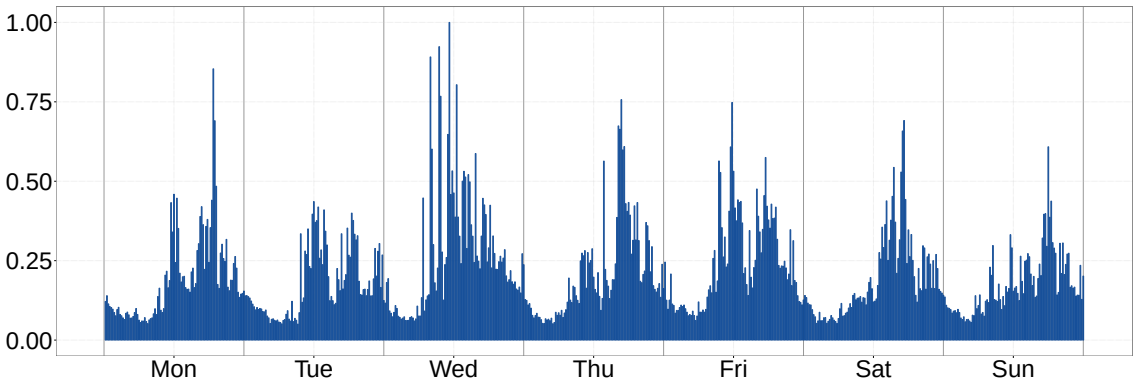
Figures 3.9-3.10 show different examples of aggregated load profiles, for a random week. Profiles are normalized according to the maximum load in the period. As the aggregation increase, the volatility decreases, the profiles becomes smoother and more autocorrelated. In addition, the relative magnitude of the base load increases, compared to the peak power.



(a) 1 Load Profile.

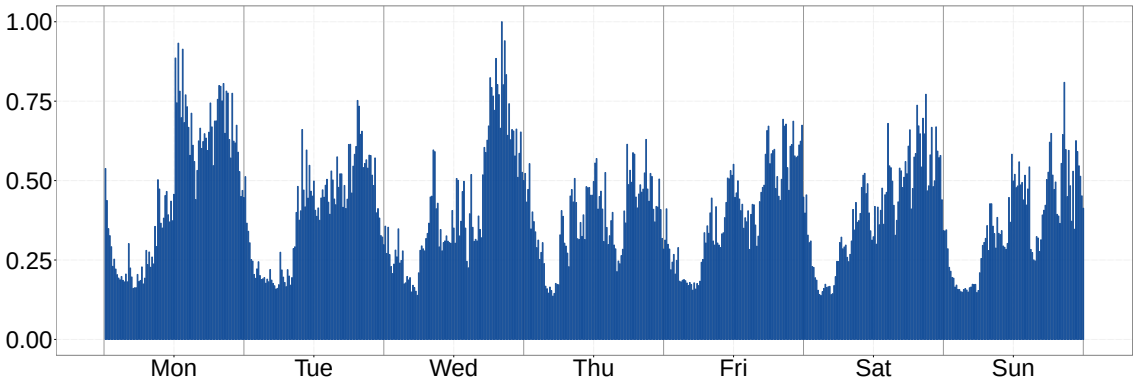


(b) 3 Load Profiles.

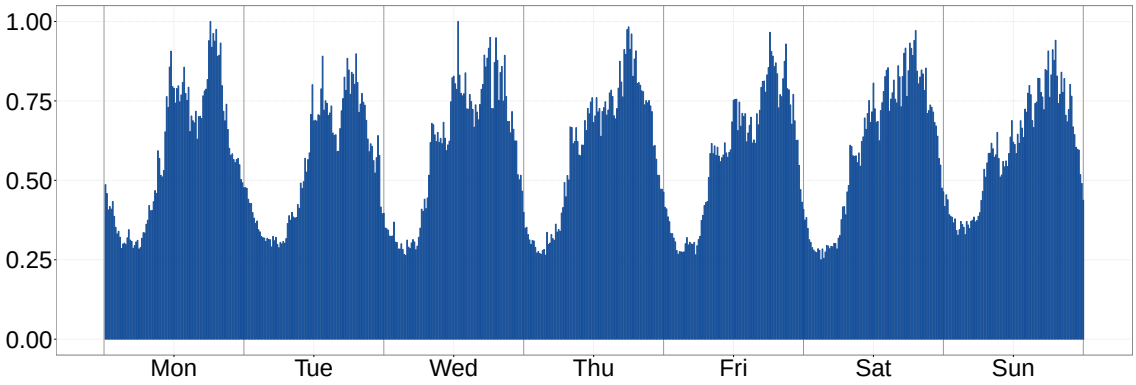


(c) 10 Load Profiles.

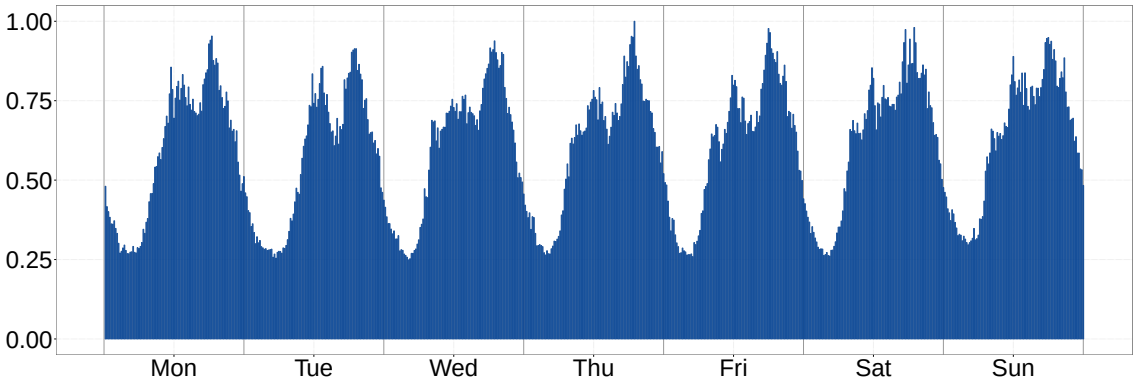
Figure 3.9: Example of aggregated load Profile, for small levels of aggregation.



(a) 30 Load Profiles.



(b) 100 Load Profiles.



(c) 300 Load Profiles.

Figure 3.10: Example of aggregated load profile, for higher levels of aggregation.

3.3.2 Autocorrelation

Autocorrelation has been defined in Section 3.2.2. As a rule, autocorrelation increases with the level of aggregation. The plots in Figure 3.11 illustrate the trends.

The x-axis spans 960 time lags, at time steps of 15 minutes. The y-axis displays, for each 15-min value of the lag, the distribution of the autocorrelation taken over a sample set of 200 aggregated profiles. Each aggregated profile of the sample set is generated by aggregating the requested number of individual profiles. Shades of green indicate different deciles in the distribution. The black line is the median value of autocorrelation for each time lag.

It can be noticed that autocorrelation approaches one for very small values of the lag, i.e. the first hour. Afterwards, autocorrelation decreases and becomes negative in correspondence of lags of around 12 hours, i.e. night consumption is generally much lower than daily consumption. Peaks occur at lags of multiples of 24 hours, as an effect of daily routines. Higher peaks occur at lags of 7 days, denoting the weekly cyclicality of patterns.

If a null level of aggregation is considered (Figure 3.11a) the distribution of the autocorrelation is quite wide. At lags of 1 day, few profiles have autocorrelation up to 0.5, a quite high value for a single profile, while others approach a correlation of zero.

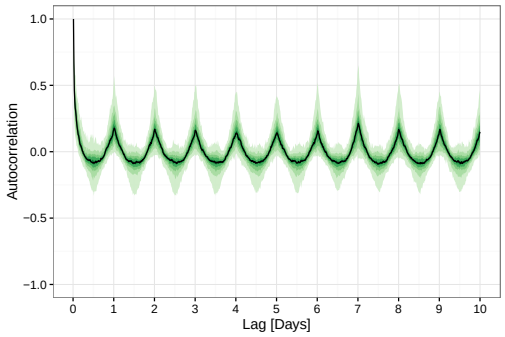
As the level of aggregation increases (Figure 3.11a-3.11f), the autocorrelation also drastically increases. Similarly, the width of the distribution of autocorrelation radically decreases. For 300 aggregated profiles, autocorrelation at multiples of 24-hours lag is approaching one, indicating that daily patterns are extremely regular and predictable.

3.3.3 Coincidence Factor

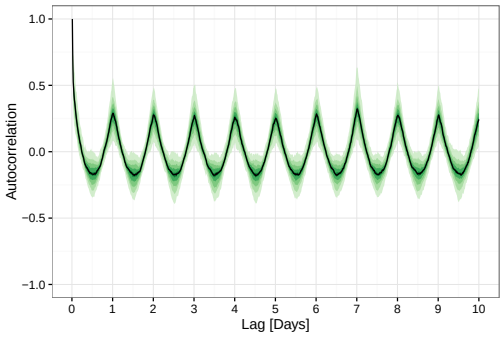
As a rule, the coincidence factor decreases with the level of aggregation, i.e. the more loads are aggregated, the smaller is the peak of the aggregated load compared to the sum of the individual peaks. Similarly, the wider the time window of evaluation is, the smaller the coincidence factor becomes, i.e. as the time window expands, the sum of the individual peaks grows more than the aggregated peak. Figure 3.12 shows the trends in the distribution of coincidence factors for samples of the Basel dataset. Darker shades of color represent central values in the distribution. The black line is the median value. It can be noted how the variance of the distribution of the coincidence factor decreases as the level of aggregation increase, and how the asymptotic value for the yearly CF (Figure 3.12c) is smaller than the one of the weekly CF (Figure 3.12b) and of the daily CF (Figure 3.12a)

3.3.4 Load Factor

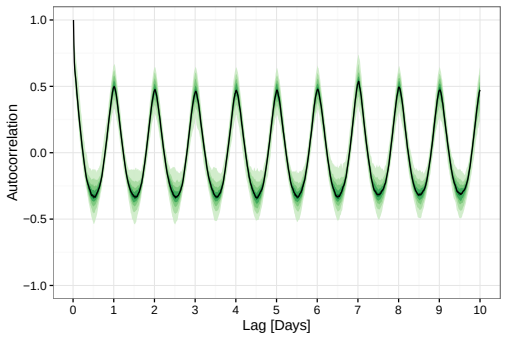
Similar considerations are valid for the load factor. As a rule, the load factor increases with the level of aggregation, i.e. the more loads are aggregated, the larger is the average load compared to the peak load. Similarly, the larger is the time window on which the load factor is evaluated, the smaller the load factor becomes, i.e. as the time scale expands, the peak load increases more than the average load. Figure 3.12 shows trends in distribution of the load factor for samples of the Basel dataset. Darker shades of color represents central values in the distribution. The black line is the median value.



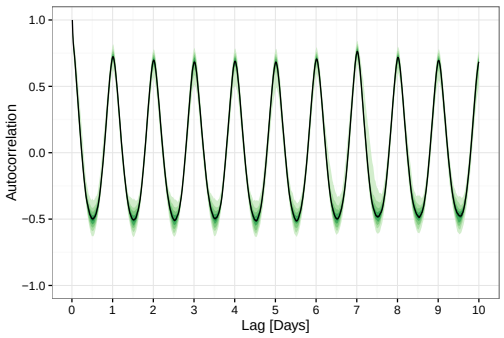
(a) Autocorrelation – 1 LP.



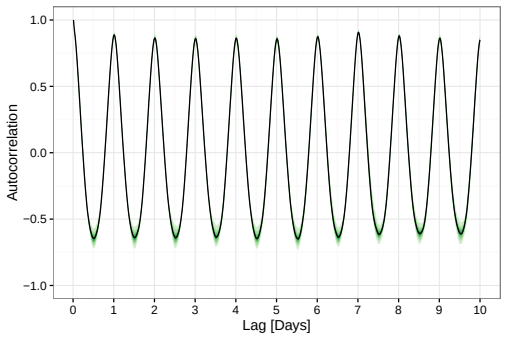
(b) Autocorrelation – Aggregation of 3 LPs.



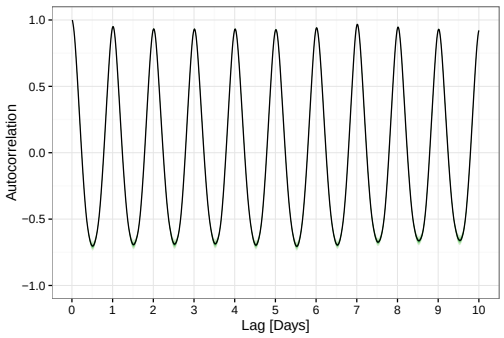
(c) Autocorrelation – Aggregation of 10 LPs.



(d) Autocorrelation – Aggregation of 30 LPs.



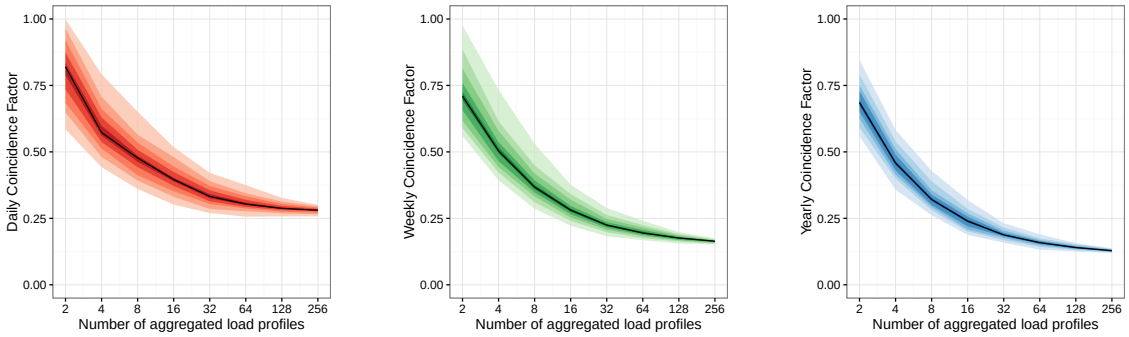
(e) Autocorrelation – Aggregation of 100 LPs.



(f) Autocorrelation – Aggregation of 300 LPs.

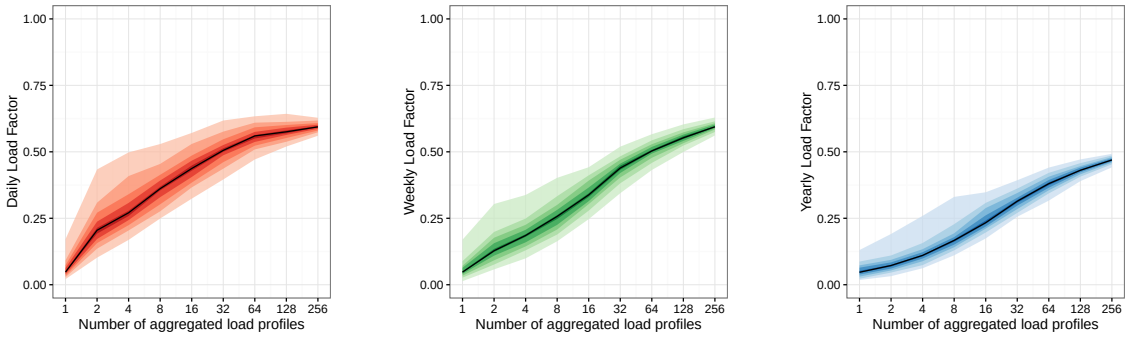
Figure 3.11: Distribution of autocorrelation with 10-day lag for 200 samples of aggregated Load Profiles. Notice the peak in autocorrelation for lag of 7 days.

As for the coincidence factor, the higher the level of aggregation, the narrower the distribution of values of the load factor becomes.



(a) Daily CF. (b) Weekly CF. (c) Yearly CF.

Figure 3.12: Trends in the distribution of the coincidence factor vs aggregation level, evaluated over time spans T = 1 day (red), T = 1 week (green), T = 1 year (blue). The distribution has been evaluated from 200 different samples of aggregated profiles.



(a) Daily LF. (b) Weekly LF. (c) Yearly LF.

Figure 3.13: Trends in the distribution of the load factor vs aggregation level, evaluated over time spans T = 1 day (red), T = 1 week (green), T = 1 year (blue). The distribution has been evaluated from 200 different samples of aggregated profiles.

3.4 Validation of Synthesized Load Profiles

As aforementioned, a reliable generator of realistic SLPs is a powerful tool for grid simulations, transformer sizing, forecasting and pricing models and allows the creation of very large realistic datasets for testing the scalability of management systems [22]. A rigorous method of validation of models for SLP generation is needed to ensure that results obtained from simulations exploiting SLPs are reliable and robust.

3.4.1 Validate a SLPs against the corresponding real load profile

In literature, comparison between a real load profile (RLP) and the corresponding SLP is carried out by confronting some parameters or parameter-ensembles, e.g. load histograms, histograms of peak ToU. Each parameter and parameter-ensemble can be defined in various flavors, as described in section 3.2.2. The following list reports indicators that has been

judged to be effective, based both on the literature and the results of this thesis.

Scalar parameters

- Total Electricity Consumption. (or equivalently, mean demand) [20, 27, 24].
- Standard deviation of load [20, 24]
- Load factor [27].
- Peak demand [27].

Parameter-ensembles

- Typical load profile (TLP) [24].
- Quantile typical load profiles (q-TLPs).
- Histogram of load magnitude [23, 24].
- Histogram of peak load magnitude.
- Histogram of Time of Use of peak load [27].
- Autocorrelation chart[24].

Unsuitable parameters for validation Typical forecast performance indicator as MAPE (Mean Absolute Percentage Error) or RMSE (Root Mean Square Error) are not suitable for validation of SLPs, since the goal is not to reproduce patterns point by point but instead to mimic an ensemble of properties. As a substitute to RMSE and MAPE, McLoughlin [27] adopted cross correlation between two load profiles X and Y as a performance indicator:

$$\rho_{X,Y} = \text{crosscorr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{3.29}$$

3.4.2 Validation at the aggregated level

Methods for generating SLPs allow the creation of huge datasets. When aggregating synthesized loads, the properties of the resulting profiles have to match to a given extent the ones from real measurements [23]. Parameters used to validate aggregation of SLPs are the same of ones utilized for individual level. In addition, aggregated SLPs should guarantee similar distributions of the coincidence factor.

3.4.3 A Synthesized Load Profile is not a forecast

It is very important to highlight that the generation of a realistic SLP has a different goal than the elaboration of a forecast. Forecasts aim at a point-by-point prediction. Conversely, a model for SLP generation aims to globally reproduce the probabilistic distributions of relevant quantities and patterns.

In other words, an ideal SLP is a profile that an expert should not be able to distinguish from a RLP. A well constructed SLP is a "fake" that is very difficult to unmask. The distinction between SLPs and forecast yields different approaches in methods for validation of the models, as pointed out in section 3.4.1. This concept is important to keep in mind for a critical validation of the model.

Chapter 4

Methodology

This chapter illustrates the main elements and intuitions of the chosen methodology. The complete machine learning pipeline of the project consists of many blocks, each block is designed to accomplish a specific task. The overall structure of the methodology will be described, trying to set a clear and solid framework. The discussion assumes a basic knowledge about machine learning concepts and methods, such as clustering and classification. For a more detailed discussion please refer to the book by Alpaydin [28]. A concise and insightful overview of useful concepts in machine learning has been written by Domingos [26]. The online course from Andrew Ng, Stanford University [29], was a great source of inspiration for this work.

Goal of the project As already stated, the goal of the complete algorithm is to generate a set of realistic load profile for buildings not equipped with smart meters, with a time resolution of 15 minutes, knowing some of their features (construction period, number of dwellings, heating system, etc.). The intuition is exemplified in Figure 4.1.

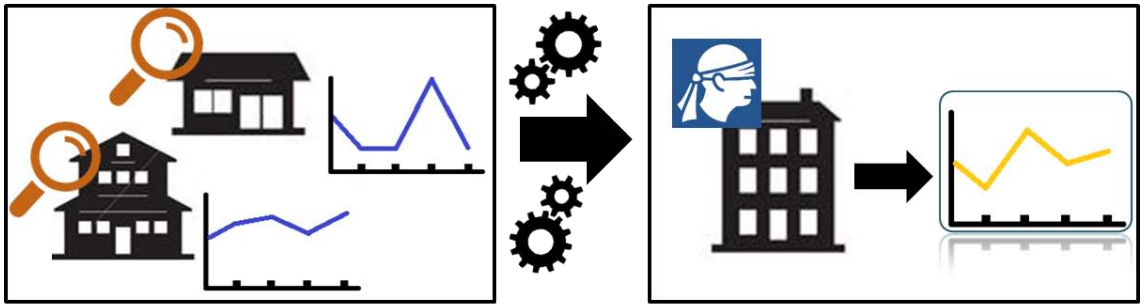


Figure 4.1: Conceptual representation of the goal of the project.

Terminology Before proceeding further in the discussion, it is useful to define some terminology that will help in the following.

- **Building features:** physical features of the building. They are of categorical or continuous type. Table 5.1 reports the complete list of features available.
- **Blind building:** a building not equipped with smart meters. It is "blind" because no measurement is provided about patterns of consumption.
- **Smart metered building:** a building equipped with smart meters.
- **Individual Load Profile (ILP):** a load profile measured by a single smart meter.
- **Building Load Profile (BLP):** In many cases, many smart meter are installed in a building. The BLP is the aggregation of all ILPs of meters of the building. In plain words, the BLP is the aggregated load profile of the building, as seen "from outside".
- **Real Load Profile (RLP):** a load profile (individual or aggregated) derived from actual measurements.
- **Synthesized Load Profile (SLP):** a load profile (individual or aggregated) artificially generated.

Basic assumption on data In the early stages of the project, data of load profiles and building features were divided into two different datasets. At the time, it was not possible to merge and skim the two datasets into a final "working dataset" because of a missing third dataset containing the connections between load profile IDs and building IDs. The third dataset was supposed to be provided by IWB in the second part of the project. Therefore, some basic assumptions had been made at the time to set the framework for the research, while waiting for all necessary data. The assumptions made regarding the structure and the content of the final "working dataset" are the following:

The "working dataset" contains only data of fully smart-metered buildings. Features of the buildings are known and linked to the building ID. All load profiles of smart meters installed in the buildings are available, and linked to the building ID. Load profiles embrace at least one year of measurements. Missing values are assumed to have been handled with appropriate methods.

A complete discussion on the structure of datasets and preprocessing can be found in Chapter 5. Here, it is sufficient to point out that three datasets had to be merged:

- **Building Features dataset (BFset)**, containing the features of the buildings.
- **Load Profile dataset (LPset)**, containing the load measurements.
- **Link Dataset (LKset)**, containing the matches between load profiles and buildings.

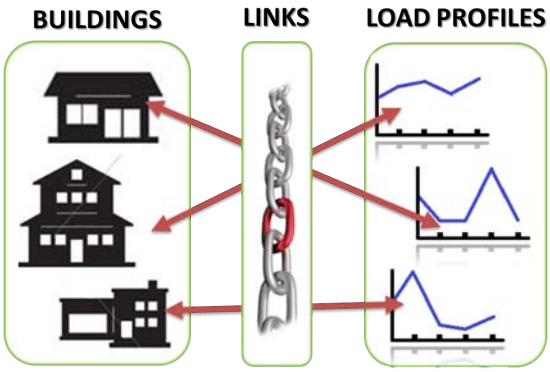


Figure 4.2: The three fundamental datasets. Features of the buildings, BFset (left); Load profiles, LPset (right); Links, LKset (center).

4.1 The machine learning pipeline

The key underlying idea of the methodology is to find correlations between the features of a building and the features of its load profile (or set of load profiles) and exploit the knowledge to generate profiles for blind buildings. More in detail, the idea is to first group buildings into classes of consumption (CoC). A CoC groups together all buildings having similar features of consumption. In the project, it is assumed that classes of consumption are non-overlapping, i.e. every building belongs to one and only one CoC. The second step is to learn how to classify a building into a CoC, exploiting only its features (and not the features of its load profiles). Third, knowing the CoC of a building, generate a realistic profile by miming load profiles of smart meters installed in buildings of the same CoC. The whole process of generating a set of realistic SLP for a blind-building is articulated in a so called machine learning pipeline constituted by different interlinked blocks, each designed to accomplish a different task: features engineering, clustering and labeling, classification, SLP-generation, validation. The complete machine learning pipeline is illustrated in Figure 4.7. The individual blocks are described in the following paragraphs.

Features Engineering: The similarity between quantities is intrinsically linked to the metric chosen to define the resemblance. In the case of load profiles, these are just raw data and cannot be directly compared to expect reasonable results. First, some key features has to be extracted in order to allow a meaningful comparison. The same is true for the features of buildings. The process of constructing significant features is of utmost importance in machine learning. As it is pointed out by Domingos [26], the features used can make the difference between a successful machine learning project and a failure. The raw data is not in a form that is amenable to learning, but features can be constructed from it that are. Trial and error in features design is a constant in machine learning projects. A few approaches of feature engineering for load profiles has already been illustrated in section 3.2.3.

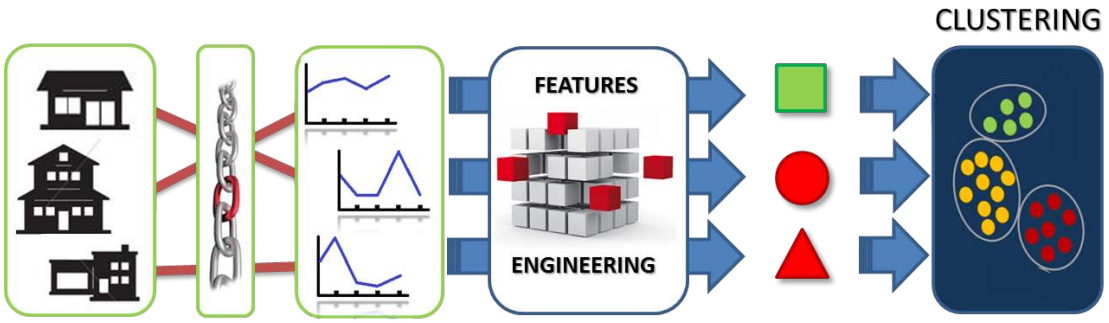


Figure 4.3: Features Engineering form BLPs, features extraction and clustering.

Clustering and labeling: The aim of this block is to group buildings with similar consumption characteristics into a same Class of Consumption. The load profile of each building is in general an aggregation of individual load profiles (ILPs). First, the ILPs related to a same building are aggregated to form a building load profile (BLP). Second, each BLP is characterized by its features, that are then used to cluster the BLPs and determine the Classes of Consumption (Figure 4.3). After clustering, each BLP receives a label according to the cluster. Then, labels are extended backwards to the buildings themselves and to the ILPs composing the BLP (Figure 4.4, ILPs are not displayed in the figure).

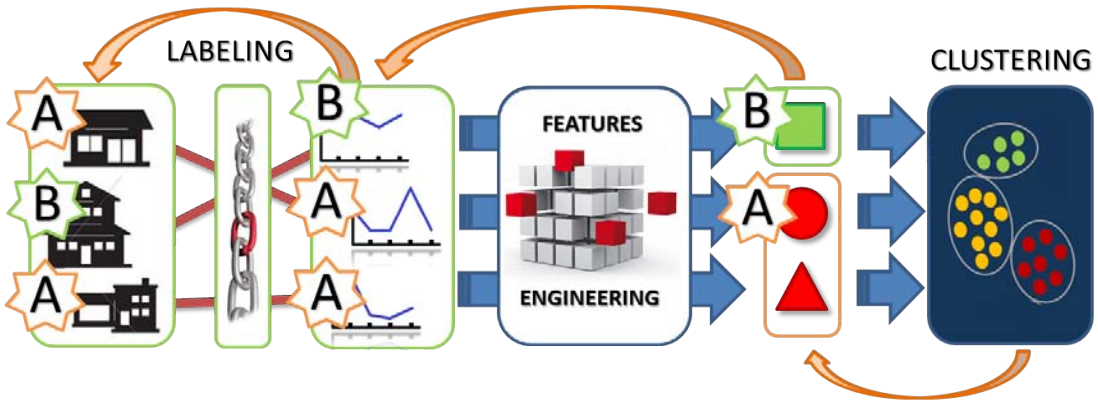
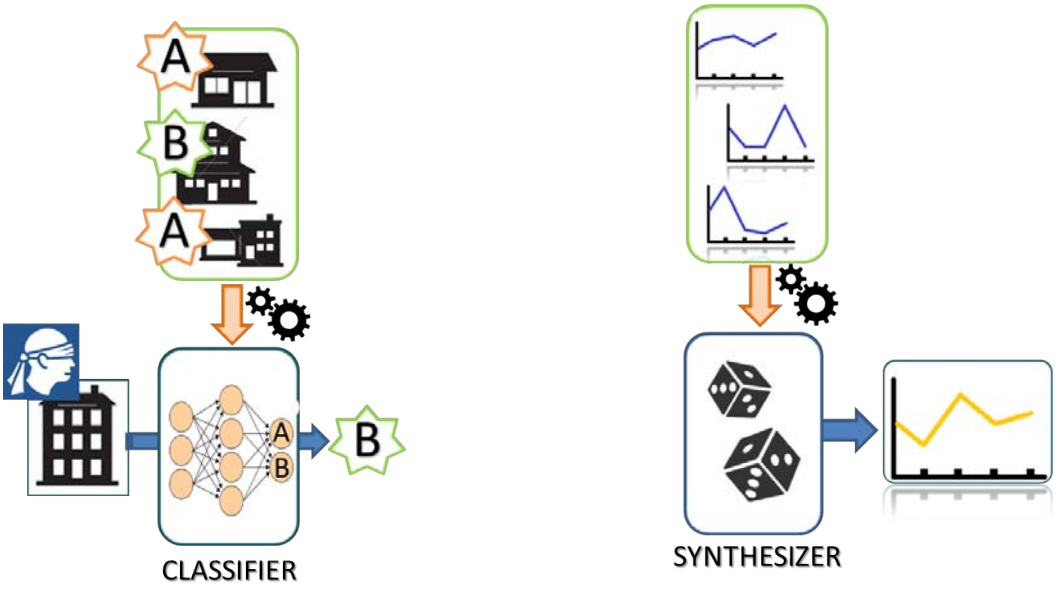


Figure 4.4: Backward labeling of BLPs and buildings, ILPs are not displayed in the figure.

Classification: After buildings are labeled by Class of Consumption, they can be used as learning examples to train a classifier that learns how to predict the Class of Consumption of a new blind building based on its features (Figure 4.5a). In this project, the chosen classifier is a neural network with one hidden layer that uses logistic regression as activation function. The number of input and output neurons is adjustable depending respectively on the number of CoC identified in the clustering phase and the number of features of the buildings chosen for classification. The number of hidden neurons is

adjustable according to a trade off between quality of the output, risk of overfitting and computational requirements. Tuning of the classifier can be carried out using the classic approach of performance evaluation based on training-set and cross-validation set.



(a) Representation of the training of the classifier used to predict the Class of Consumption of a blind-building.

(b) Representation of the load profile generator block.

Figure 4.5: Conceptual representation of the blocks of classification and SLP-generation.

SLP-Generation: The generation block has the task to generate a SLP for a blind building, by knowing the set of its features and the Class of Consumption. The underlying idea is divided into a few steps.

- First, design of an algorithm that is able to take as input a known load profile and to output an arbitrary number of synthesized load profiles that resemble the original (Figure 4.5b). A probabilistic Markov chain model has been developed for the purpose and Chapter 6 is entirely dedicated to the discussion about how this task is carried out. Chapter 7 reports the results and validates the model.
- Second, generation of a synthesized BLP by aggregation of many synthesized ILPs. Each synthesized ILP is generated from a real ILP of the same CoC of the blind building. The number of ILPs to be aggregated is a function of the number of dwellings of the blind building. In case of detached households, the number of dwellings (i.e. one) is equal to the number of meters installed. For buildings with more than one dwelling, there is usually an additional meter installed to account for consumption in shared areas. Rules for sampling the ILPs to be aggregated can be established by further analysis. Here, as a starting point, a naive approach has been chosen. The synthesized ILPs to be aggregated are randomly sampled within all synthesized ILPs of the according CoC.

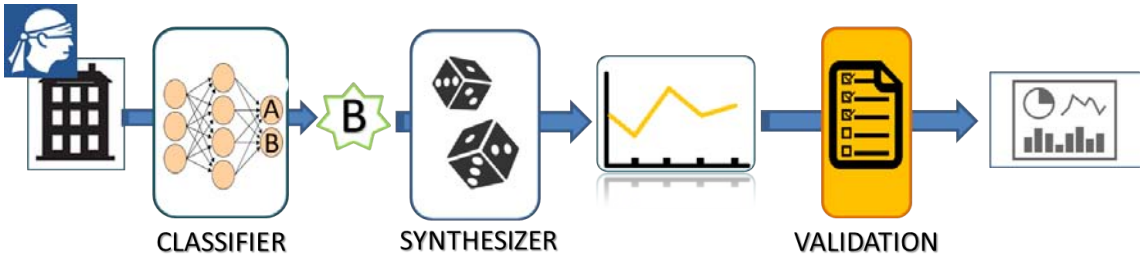


Figure 4.6: The three fundamental datasets: building features, load profiles and links.

Validation and iterative tuning: Validation of the whole project can be carried out on two different levels:

- **Block level:** the performance of some blocks can be evaluated without the need to run the algorithm entirely. For example, the SLP generation block can be validated by comparing the outcome of the generator to the input profile, its performances are independent by the performance of other block, i.e. clustering or classification.
- **Ensemble level:** for other blocks, i.e. the classification block, validation is more critical. For example, suppose that the features extracted from load profiles are not relevant to highlight similarities in consumption. Hence, features could not be different enough between each other to allow a meaningful clustering. The clustering algorithm will group load profiles in clusters that are not relevant. With this grouping, there may be no correlation at all between the building and the assigned label. When the labels in the training set are distribute randomly, even the best classification algorithm cannot learn anything and the classifier would seem to be unable to properly classify buildings, whilst the true problem is somewhere else.

  Validation must therefore be carried out holistically, considering many block simultaneously.

It is important first to keep the algorithms of the individual blocks as simple as possible. As pointed out by Domingos [26], the key factors are the quality of features and quantity of data. A simple and naive algorithm trained with good features and a large quantity of data typically outperforms a complex one trained with few noisy data.

In this work, the validation of SLPs has been carried out by comparing the parameters and parameter-ensembles illustrated in section 3.4: mean consumption, histograms of load, histograms of peak load and ToU of peak load, autocorrelation. Chapter 7 is entirely dedicated to this topic. Figure 4.7 illustrates the complete machine learning pipeline of the project.

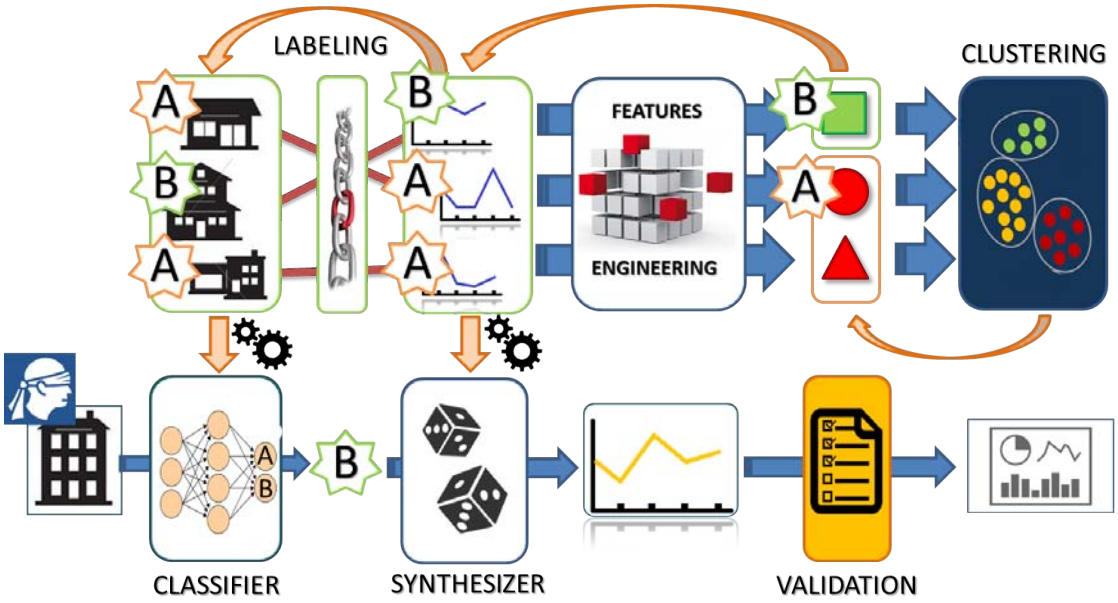
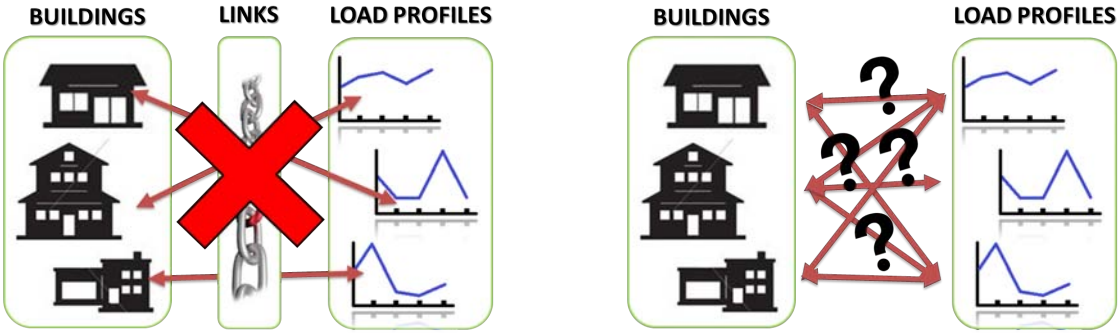


Figure 4.7: Complete overview of the machine learning pipeline.

4.2 The problem of the missing dataset

As pointed out at the beginning of this section, the whole machine learning pipeline has been designed with the assumption of having a working dataset in which IDs of load profiles were matched to the IDs of the buildings. The matches were contained in a dataset owned by IWB that was supposed to be provided by the second half of the project. Unfortunately, this dataset could not be provided by the end of the time window of the project. For this reason, although all blocks were in place, it has not been possible to run and validate the entire pipeline. Conversely, it has been possible to develop the algorithm for generation of SLPs more in detail, that could be validated independently from the rest, since the SLP has only to resemble the input profile. The algorithms for the complete machine learning pipeline has been designed and implemented and are available for test as soon as data will be at disposal.



(a) Missing dataset.

(b) Unknown association between datasets.

Figure 4.8: The problem of the missing dataset.

Chapter 5

Data preprocessing and exploration

5.1 Available datasets

At the beginning of the project, not all the needed datasets were available. This section describes the preliminary data processing and skimming of the existing datasets. The preprocessing of data stopped at the phase of association of the load profiles IDs with the buildings IDs, but it is intended to set a base for future works. Before moving forward in the discussion, it is worth to clarify the terminology.

Each building is identified by a "building ID". Each customer of IWB is identified by a "customer ID", independently from being equipped with a smart meter or not. Each smart meter is identified by a "smart meter ID". In the discussion, the terms "load profile ID" and "smart meter ID" are utilized equivalently. Smart meter IDs are different from consumer IDs and it is not possible to deduce one from the other.

The three datasets illustrated in the previous chapter and reported in figure 4.2 are actually already the result of an intermediate stage of preprocessing. An additional dataset was actually present in the raw data but was subsequently incorporated in the dataset of building features. The following list reports the raw datasets and to set the bases for the further discussion on preprocessing.

- **Load Profiles Dataset (LPset):** dataset provided by IWB, it contains the time series (Load profiles) of load measurements for 13 months of around 40'000 smart meters in Basel. Load is sampled at intervals of 15 minutes. Each load profile is labeled only by the smart meter ID, without any link to the physical address of the meter. The dataset was already been preprocessed in a previous work [30], that removed outliers and incomplete time series and integrated missing data in the load profiles using the k-nearest neighbors algorithm.

- **Building Features Dataset (BFset):** dataset provided by the Swiss Federal Statistical Office. It contains features and geographical coordinates of each building in Switzerland. The complete structure is reported in Table 5.1.

- **Link Dataset (LKset) [MISSING]:** dataset owned by IWB and reporting the

customer ID of each smart meter ID. It allows to match a load profile with the physical address of the smart meter. Using this dataset, a building can be coupled with all the load profiles of the occupant customers.

- **Address List Dataset (ALset):** dataset provided by IWB, it contains the list of the physical address of each customer-ID in Basel.

Building IDs are associated with geographical coordinates in BFset, coordinates can be converted into addresses, addresses can be matched to customer IDs through ALset, customer IDs can be matched to smart meter IDs through LKset (missing), smart meter IDs are associated with load profiles in LPset.

5.2 Overview of the Building Features Dataset

Although the building features could not be exploited due to the missing LKset, a brief exploration of BFset is worth to uncover some shared characteristics of the buildings. The analysis is carried out on the skimmed version of BFset, i.e. only for buildings with green signature in Figure 5.4.

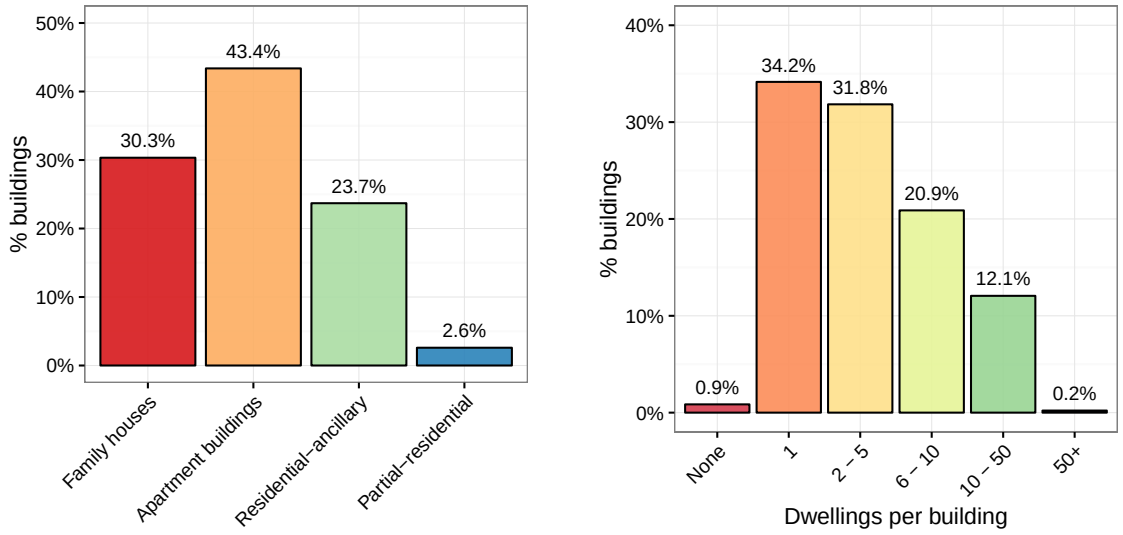
The building categories included in BFset are the followings:

- **Family homes.** Detached households, purely residential buildings occupied by a single family.
- **Apartment buildings.** Purely residential buildings, constituted by multiple units, such as row houses or multilevel apartment blocks.
- **Residential buildings with ancillary use.** Residential buildings with a minor share of non-residential units, normally located at the ground floor: small shops, offices, restaurants or other commercial activities.
- **Buildings with partial residential use.** Buildings with partial residential use, but consisting mainly of industrial, commercial or agricultural facilities.

Purely non-residential buildings are not included in the dataset. i.e. building without residential use, consisting exclusively of spaces assigned to industrial, commercial or agricultural activities.

Figure 5.1a displays the share of the different typologies. Figure 5.1b displays the distribution of the buildings by number of dwellings.

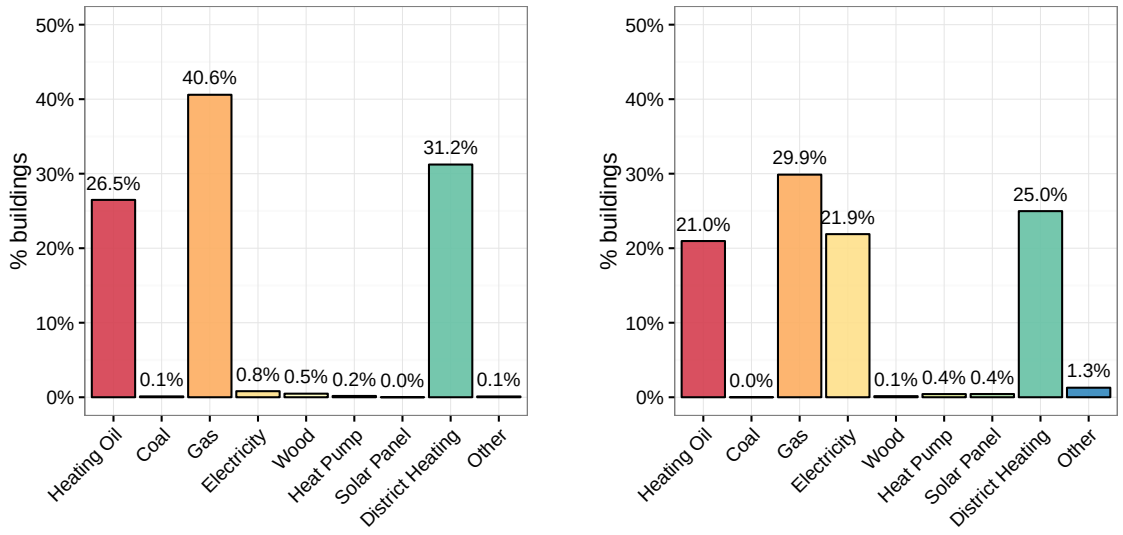
An overview of the typologies of space heating systems shows that buildings in Basel rely almost totally on district heating or fossil fuels (heating oil and gas), with absolutely negligible penetrations of heat pumps and electric space heating systems (Figure 5.2a). Concerning water heating, the situation is slightly different, with penetration of electric boilers slightly above 20% (Figure 5.2b).



(a) Share of building category.

(b) Share of dwellings per building.

Figure 5.1: Share of buildings per category and number of dwellings.



(a) Space heating.

(b) Water heating.

Figure 5.2: Share of fuels for space and water heating.

Table 5.1: Summary of building features dataset.

	Type	Range	Categories
Building ID	Categorical	–	–
Position: Latitude	Continuous	7.55 – 7.68	–
Position: Longitude	Continuous	47.52 – 47.60	–
Category	Categorical	4	Family houses Apartment buildings Residential buildings with ancillary use Buildings with partial residential use
Construction period	Categorical	13	Period before 1919 Period 1919 – 1945 Period 1946 – 1960 Period 1961 – 1970 Period 1971 – 1980 Period 1981 – 1985 Period 1986 – 1990 Period 1991 – 1995 Period 1996 – 2000 Period 2001 – 2005 Period 2006 – 2010 Period 2011 – 2015 Period after 2015
Heating System	Categorical	7	No heating Individual heating Central heating for the floor Central heating for the building Central heating for several buildings Public District heating supply Other heating
Fuel for heating	Categorical	10	No energy Source Heating oil Coal Gas Electricity Wood Heat pump Solar panel District heating Other energy sources
Fuel for hot water	Categorical	10	No energy Source Heating oil Coal Gas Electricity Wood Heat pump Solar panel District heating Other energy sources
n. floors	Integer	1 – 35	–
n. dwellings	Integer	1 – 180	–
n. persons (total)	Integer	0 – 269	–
n. pers. (main residence)	Integer	0 – 269	–
n. pers. (perm. resident)	Integer	0 – 269	–

5.3 Preprocessing of building data

The dataset containing the building features (BFset) has been provided by the Swiss Federal Statistical Office. BFset included data on all building in Switzerland, with geolocation expressed in Swiss Coordinates System (geodetic datum CH1903+), buildings are geolocated within a precision of ± 1 m.

First, buildings of Basel have been selected and the coordinates were converted into GPS coordinates of latitude and longitude (WGS 84) using a free online tool offered by the Swiss Federal Office of Topography (SwissTopo). Once coordinates have been converted, it was possible to locate the buildings onto an interactive map exploiting the JavaScript library *LeafLet*[1] and the online services offered by *OpenStreetMap*[2]. A first geographical representation of the spatial distribution of the buildings is reported in Figure 5.3.

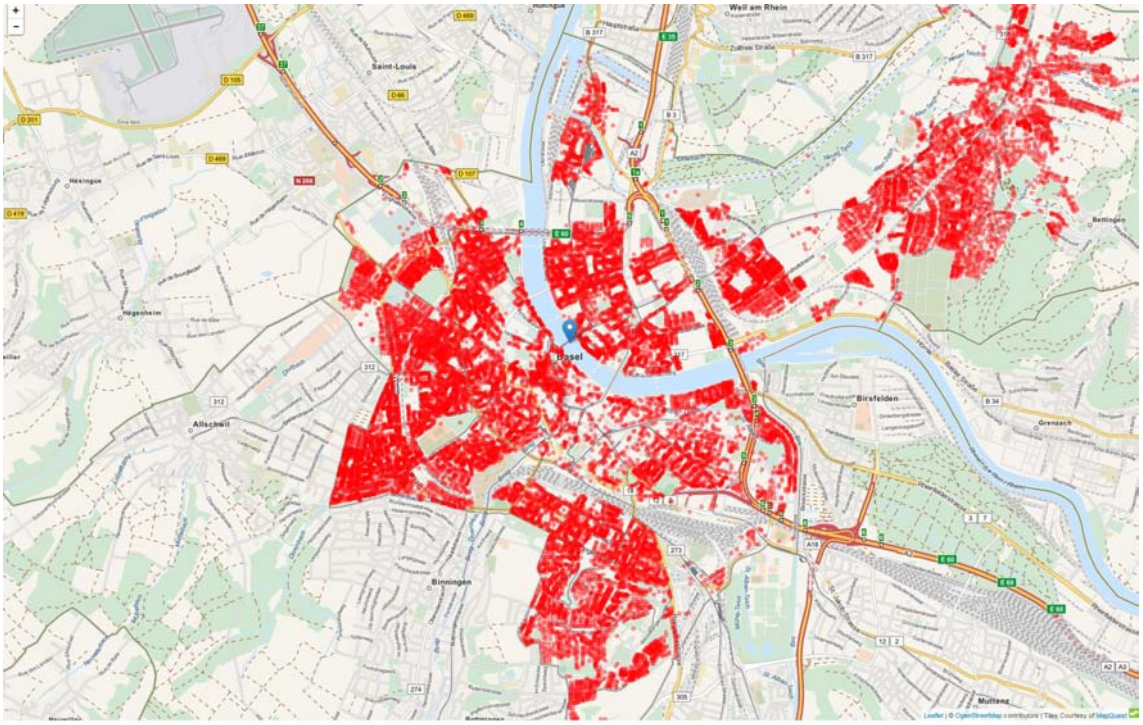


Figure 5.3: Position of buildings in Basel, before skimming by address matching.

Once all buildings had been successfully identified with latitude and longitude, the addresses of the buildings had to be determined using the GPS coordinates. This process is called reverse geolocation. A free online service offered by *Nominatim*[3] has been used for the purpose, using a geoJSON interface for R (package *jsonlite*). Some building coordinates could not be correctly reverse-geolocated and had to be discarded. This was due to at least one of the following reasons:

- The reverse-geolocated address was incomplete, i.e. the street name or the house

[1] leafletjs.com
[2] openstreetmap.org
[3] nominatim.openstreetmap.org

number was missing.

- More than one building was reverse-geolocated to the same physical address.

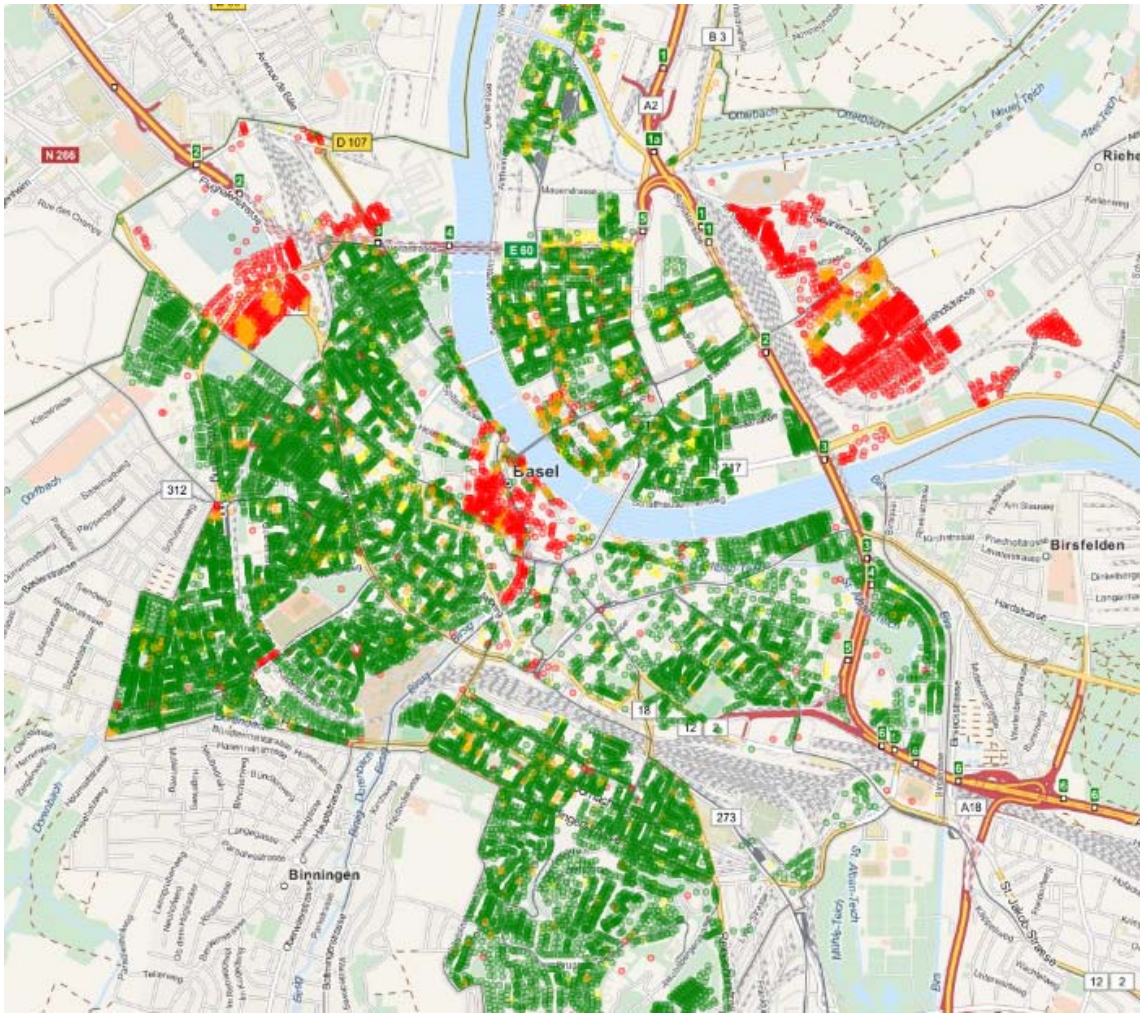


Figure 5.4: Position of buildings. Snapshot of Basel.
Green: Building matched with at least one customer ID.
Yellow: Building matched with zero customer ID.
Orange: Building with non-unique address.
Red: Building with no address available (incomplete reverse geocoding).

Once buildings were matched with physical addresses, the addresses could be matched with customer IDs, exploiting the ALset. In case of more customers living in the same building, their customer IDs are obviously matched to the same address. As for reverse-geolocation, also in this matching phase some data could not be matched and had to be discarded.

In the process of skimming and merging of ALset with BFset, 84% of buildings has been correctly matched with address and set of customer IDs. A bottle of wine will be awarded to the reader reaching this line.

Some considerations are worth noticing:

- The customer ID does not tell whether the customer is equipped with a smart meter or not and does not link the customer to the load profile recorded in LPset.
- Except from family houses, the number of customers IDs assigned to a building is typically different from the number of dwellings. This is because buildings with more than one dwelling are also equipped with a meter for common consumption, for example, for lighting of common spaces etc. In addition, a small fraction of customer IDs could not be matched because of typos in the spelling of the addresses.

As it can be noticed from Figure 5.4, few areas of the city are most affected by data mismatching. Figure 5.6 illustrates an example situation in the remaining of the city.

Figure 5.5: Position of buildings. Snapshot of a residential area in South-West Basel.
Green: Building matched with at least one customer ID.
Yellow: Building matched with zero customer ID.
Orange: Building with non-unique address.
Red: Building with no address available (incomplete reverse geocoding).

Figure 5.6: Position of buildings. Snapshot of a residential area in South-West Basel.

Chapter 6

Generation of Residential Load Profiles

6.1 The choice of the Model

The model for probabilistic generation of Synthesized Load Profiles (SLPs) must be flexible and data-driven, allowing to represent each consumer with its own model. It has to be able to generate an arbitrary number of load profiles that resemble the features of a specific real profile utilized as input.

As pointed out in Section 3.4.1, the resemblance of the SLPs to the original profile is judged by comparing different parameters or parameter-ensembles, in particular:

- Total annual consumption.
- Distribution of magnitudes of load.
- Distribution of magnitudes of daily peak load.
- Distribution of Time of Use of daily peak load.
- Autocorrelation.

The model has been optimized and validated according to these indicators.

6.1.1 Time scale of variability

The model has to be able to reproduce variability on typical time scales. For residential load profiles, four different time scales have been identified:

- **Intra-day.** Residential load is usually lower at night and higher during daytime, with peaks at breakfast, lunch and in the evening.
- **Intra-week.** Depending on occupation patterns and occupants practices, consumption may differ depending on the day of the week. For instance, some customers may have specific routines during some days of the week, and consumption at weekends may differ from consumption during weekdays.

- **Week-to-week.** Consumption might vary significantly from week to week. In some cases, this effect can be correlated to external temperature, in others, different patters appear in apparently random weeks, not directly imputable to temperature, as it can be seen from Figure 6.1. Occupation patterns and human activities are prone to be subject to high variability on this time scale (holidays, busier weeks at the workplace, etc.)

- **Season-to-season.** Consumption may evolve following year-long cycles, according to the average external temperature, as pointed out in section 3.1.2.

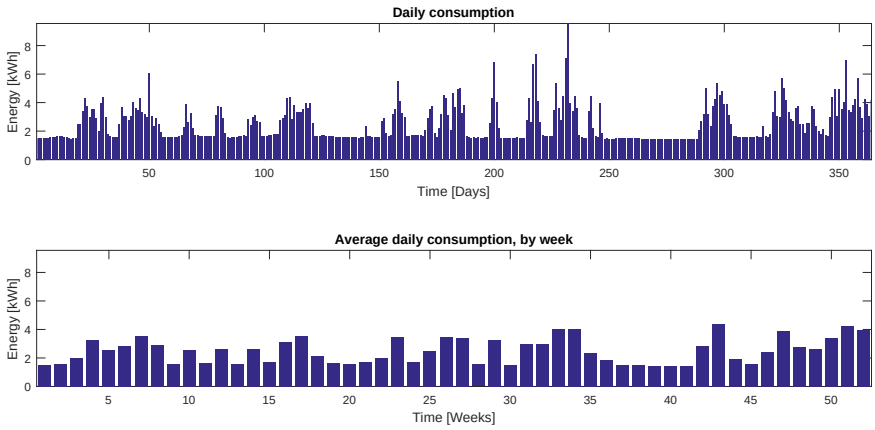


Figure 6.1: Example of an electricity demand profile from an individual household, showing large week-to-week variations.

It has been pointed out in Section 2.4 that the behavioral component has been found to account for variation in consumption up to 300–400% in identical dwellings [14]. In view of this, occupation patterns and practices of occupants are expected to play a much larger effect on the variation in consumption than the temperature. This is especially true for countries with low penetration of electric heating systems or air conditioning, as it is the case for Switzerland (for the City of Basel, the penetration of electric heating systems is quite low, see Figure 5.2). For this reason, the most influencing time-scales for the generation of a SLP has been judged to be the ones where the behavioral component is higher: the intra-day and the week-to-week. Hence, seasonal effects and the effect of the temperature are not considered. For the same reason, and for the sake of simplicity, no difference between days of the week is considered.

The model has to be able to cope with highly stochastic profiles, which are highly correlated on short term lags and are characterized by sharp spikes and little influence of temperature. For these reasons, it has been chosen to implement a probabilistic model based on Markov chains, that demonstrated to be very effective in reproducing autocorrelation and very volatile load profiles [24]. The next section will briefly define what Markov chains are and will illustrate advantages and disadvantages of the method.

6.2 Markov chains

A Markov chain is a model for representing a stochastic process between discrete states, whereby a transitions between states occur at discrete time steps. A finite set of states is defined, and the Markov chain is described in terms of its transition probabilities, t_{ij}, which determine the probability of transitioning from state i to state j, regardless of previous states that were visited [31]. The discussion about Markov chains and their application follows the exposition in [31, 32, 33, 34].

A Markov chain is a sequence of random variables $\{X_1, X_2, ...\}$ such that a future state X_{n+1} is dependent only upon the present state X_n. The transition probabilities between states can be represented in a squared matrix T, called *transition matrix*, such that the element t_{ij} is the conditional probability of transitioning to state j at time step $n+1$, being in state i at time-step n. More formally, for a Markov chain of M states:

$$t_{ij} = P(X_{n+1} = j | X_n = i) \tag{6.1}$$

$$T = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1M} \\ t_{21} & t_{22} & \cdots & t_{1M} \\ \vdots & \vdots & \ddots & \vdots \\ t_{M1} & t_{M2} & \cdots & t_{MM} \end{bmatrix} \tag{6.2}$$

Since the transition probabilities from a given state must add to 1, the sum of each row of a transition matrix T must also sum to one, that is:

$$\sum_j t_{ij} = 1 \tag{6.3}$$

If the transition matrix T remains constant over time, the associated Markov chain is said to be *time-homogeneous*. Conversely, if T is time-dependent, i.e. $T = T(n)$, the resulting Markov chain is called *time-inhomogenous*. Markov chains can be randomly walked to produce stochastic time series, taking values in the finite set of states. When utilized to model load profiles, each state of the Markov model represents a certain range of power consumption. In the following, the terms "state" and "load level" will be used equivalently.

6.2.1 Advantages

Markov chains are very effective in reproducing sudden spikes of load magnitude, compared to other methods based, for example, on Gaussian processes or multiple regression, that tend to create smooth outcomes, not suitable to reproduce the high volatility of individual residential profiles [27]. Compared to a simple Monte Carlo Method that samples load magnitude according to a given load probability distribution, Markov chains, by nature, also allow to capture temporal correlation. The effectiveness in reproducing autocorrelation increases if different time periods of the day are modeled separately with different transition matrices. In this case, the model becomes time-inhomogeneous.

6.2.2 Disadvantages

The use of a large number of transition matrices requires the model to calculate and store a large number of parameters, increasing memory and computational requirements. In addition, a simple Markov-chain model does not take into account the effects of exogenous variables. Nevertheless, for the scope of this work, known exogenous variables, such as temperature, have been considered to have a mild influence compared to other unknown exogenous variables, such as occupancy patterns or lifestyle of the occupants, as reported in Section 2.4.

6.3 Implementation of the Markov-chain model

The model was inspired by the work of Labeeuw and Deconinck [24], but it differs substantially in the implementation. The method operates in two phases.

- **Training phase:** given an input load profile, in the training phase the parameters of the corresponding Markov model are determined. Parameters include different type of transition matrices and a number of ancillary parameters to characterize the distribution of the load magnitude for each state of the model. Details are thoroughly illustrate hereafter.
- **Generation phase:** the model is exploited to generate an arbitrary number of probabilistic SLPs resembling the original RLP.

As required, the model is data-driven and input-specific, i.e. different model parameters are tailored on different input load profiles. The model is bi-level, i.e. it reproduces patterns of variability on two different time scales: week-by-week and intra-day, whereas the week-by-week states influence the intra-day ones.

- **First level – Weekly model:** variations on week-to-week basis are analyzed and reproduced. This is accomplished by randomly walking a transition matrix T^W of states of weekly consumption.
- **Second level – Intra-day model:** once the variation on the week time scale is accounted, the model aims at reproducing the intra-day variability, by randomly walking different transition matrices according on the Time of Use. Each state is coupled with ancillary information on the distribution of load within that particular state. These distributions are utilized to achieve a more accurate resolution of the output.

Input of the model The model is designed to adapt to an input load profile and reproduce its features. In the specific framework of this thesis, the input of the model is a load profile with the following properties:

- Measurements of energy consumption at sampling rates of 15 minutes, i.e. 96 samples per day. Resolution of 1 Wh.
- 52 weeks of measurements. i.e. 364 days, 34,944 samples, from Monday, April 7^{th}, 2014 to Sunday, April 5^{th}, 2015 .

6.3.1 Convention on nomenclature

- **Discretized profile:** a profile whose values can take a limited number of states.
- **Continuous profile:** a profile whose values are continuous. Both RLPs and SLPs are continuos load profiles.
- **Week-state:** each of the states of the weekly Markov model. Each week-state defines a certain range of weekly consumption.
- **Week-to-week profile:** a discretized profile, characterized by a time resolution of one data point per week. Values of data points are week-states.
- **ToU-state:** each of the states of the intra-day Markov model. Each ToU-state defines a certain range of consumption at the given Time of Use.
- **Intra-day profile:** a profile characterized by high time resolution. In this case, one sample per Time of Use of 15 minutes, 96 data points per day.

6.4 Training phase

This section describes in detail the steps to train the Markov chain model. As aforementioned, the model works on two levels: a weekly level, operating week-by-week and defined by a time-homogeneous model, and an intra-day level, operating at 15-min time intervals and defined by a time-inhomogeneous model.

6.4.1 Weekly time-homogeneous model

1. **Computation of the weekly consumption, week-by-week:** 15-min samples belonging to a same week are summed up to obtain a load profile of 52 weekly consumption points.
2. **Clustering of week-load points:** The 52 week-load data points are clustered according to the magnitude of consumption into three clusters: “Low” (L), “Medium” (M), “High” (H). Each cluster represents a “week-state”. The clustering is carried out with k-means, using euclidean distance. In this case the clustering algorithm operates only on one dimension (weekly consumption), therefore the risk of incurring in a sub-optimal cluster configuration is very low. For this reason, k-means is run only three times with different starting points. The final clustering configuration is the trial with the minimum total distance.
3. **Creation of the discretized week-to-week profile:** At this stage, every week data point has been assigned to a cluster and labeled with its week-state. The 52-data points weekly profile can be rewritten in discretized form, as a sequence of labels of week-states, to create the week-to-week profile. The transition matrix T^W, that models the week-by-week behavior, is built by calculating the relative frequency of occurrence of the transitions appearing in the week-to-week profile.

4. **Computation of the weekly transition matrix:** Each entry t_{ij}^W of the weekly transition matrix T^W expresses the conditional probability of the transition from a week of week-state i to a week-state j the next week. By considering the relative frequency of appearance f_{ij}^W of each type of transition in the discretized profile, it is possible to approximate the transition probabilities t_{ij}^W and fill up the transition matrix T^W. The relative frequency f_{ij}^W of transitions between week-states is defined as:

$$f_{ij}^W = \frac{\text{number of observations of transitions from state } i \text{ to state } j}{\text{number of observations of state } i} \tag{6.4}$$

The choice of considering three states for weekly consumption is a consequence of data availability. To properly approximate probabilities with relative frequencies, an appropriate number of events must be available. When evaluating the relative frequencies of transitions, the denominator is given by the number of observed weeks of the specific week-state. A higher number of states means less data to calculate the probability of each transition and more risk of over-fitting the data by a coarse estimation of probability. On the other hand, few states result in lower detail. The number of transitions has to be chosen as a trade-off between over-fitting and having enough detail [24].

For instance, considering a 1-year long load profile of weekly consumption, 52 data points are available. Assuming 3 states with the same likelihood of occurrence, in expectation there will be $52/3 \approx 17.3$ observations per state. A choice of 4 states would lead to an expectation of 13 points per state, which has been judged not to be enough.

In the generation phase, T^W is randomly walked to generate a synthesized week-by-week profile of week-states. Each week-state is associated to a specific intra-day model, that generates a profile with 15-min resolution for that specific week. The next section will discuss the details of the intra-day time-inhomogeneous model.

6.4.2 Intra-day time-inhomogeneous model

Once the model for the weekly behavior is trained, a more detailed intra-day model for the generation of a load profile with a time resolution of 15-min can be tailored on the data. The intra-day model consists of three sets of 96 $M \times M$ transition matrices T_h^w (one set for each week-state). M is the number of states of each ToU in the intra-day model (ToU-states), in this case, $M = 5$. $h \in \{1, 2, .., 96\}$ is the considered Time of Use and $w \in \{\text{L, M, H }\}$ is the week-state. $t_{h,ij}^w$ is the estimated probability of the transition from state i in ToU $= h$ to state j in ToU $= h + 1$, for a week having a week-state w. For each input load profile, $96 \times 3 = 288$ transition matrices are generated. Despite the higher requirements in computational power and memory compared to a time homogeneous model, the creation of an individual transition matrix T_h^w for each ToU and each week-state allows to better reproduce the temporal correlations between load magnitude of two adjacent data points. The detailed algorithm is described in the followings.

1. **Grouping of data according to week-state:** A set of 96 transition matrices has to be generated to characterize the transitions between ToU-states for each of the

three week-states. All data points belonging to weeks of the same week-state are grouped together to form three different subsets. Different transition matrices are be trained for each subset.

2. **Further grouping of data according to the Time of Use:** For each subset, data points related to the same Time of Use are grouped together.

3. **Clustering ToU-load into ToU-states:** Similarly as for the weekly model, for each ToU the energy consumption is clustered into M groups of consumption (ToU-states). Each ToU-state is then associated to a range of load magnitude, with the maximum being the maximum magnitude of load, measured in that ToU-state, and the minimum being the maximum magnitude of load, measured in the ToU-state of immediately lower magnitude, Figure 6.2 clarifies the concept. Three clustering repetitions are carried out to ensure robustness of the clustering. It has been found that a number of ToU-states $M = 5$ is a good compromise between speed, quality of the output, memory requirements and accuracy in probability estimation.

   It is important to point out that the clustering is performed individually for each Time of Use. In fact the distribution of load magnitude varies for each ToU throughout the day, as it is clear from the q-TLP chart of Figure 3.7, i.e. dinner-time ToUs are likely to show very wide distributions, whilst for nighttime ToUs the distribution is usually very narrow.

4. **Creation of discretized intra-day load profile:** Once ToU-states has been determined for each ToU by clustering, the whole load profile can be rewritten in discretized form as a sequence of ToU-states. This discretized profile allows the creation of the set of transition matrices T_h^w that capture the transition probabilities between ToU-states.

5. **Creation of intra-day transition matrices $\boldsymbol{T_h^w}$:** Each entry $t_{h,ij}^w$ of the intra-day transition matrices expresses the probability of the transition from the state of consumption i at ToU $= h$ to a state of consumption j at ToU $= h + 1$ (obviously if $h = 96$, $h + 1 \triangleq 96$). Using the same approach utilized for the weekly model, it is possible to approximate the transition probabilities and fill up the transition matrix T_h^w by computing the relative frequency of occurrence of each transition in the discretized profile.

   To illustrate the point, a numerical example follows. The entry $t_{36,(2,4)}^{\mathrm{M}}$ is the probability, in a week of medium consumption ($w =$ "M"), to have a transition from the ToU-state $i = 2$ at the ToU $h = 36$ to ToU-state $j = 4$ at the ToU $h = 37$. Suppose now that 40 weeks of medium consumption occur in the input week-to-week load profile. Then, suppose that among these $40 \times 7 = 280$ days, 80 days have a ToU-state $i = 2$ at ToU $h = 36$. Of these 80 days, suppose that 10 have $j = 4$ at ToU $= 37$. Hence, $t_{36,(2,4)}^{\mathrm{M}} = 10/80 = 0.125$

6. **Subdivision of each ToU-state into sublevels:** Each ToU-state associated to a range of load magnitude. At this stage, as a further refining of the discretization, the distribution of the available load data within a ToU-state is evaluated: each ToU-state is further split into 10 sublevels or "bins" of equal width of load magnitude

that span the range of the ToU-state. Figure 6.2 exemplifies the subdivision in ToU-state (by clustering) and sublevels (by splitting each ToU-state into bins of uniform width).

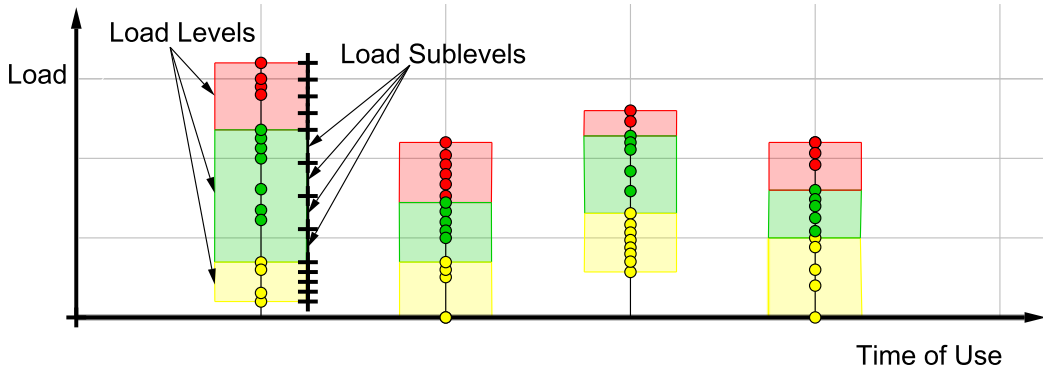


Figure 6.2: Conceptual procedure of the intra-day grouping into load levels (using k-means clustering) and load sublevels (using uniform binning). The number of clusters (ToU-states) and sublevels depicted are fictitious.

7. **Estimation of probability of occurrence of each sublevel:**

   The relative frequency of occurrence of each bin is evaluated, approximating the probability that a load of a given ToU-state is observed in a given sublevel.

   The choice of approximating the probability of occurrence of a sublevel with the relative frequency of occurrence, instead of fitting a probability distribution, has been dictated by the observation of data. For example, Labeeuw [24] proposes a piece-wise fit of two Weibull distributions, but given the often cumbersome shape of the distributions of load magnitude found in the data of the Basel Dataset (for example see figure 3.4), this method has been judged to be unjustified in this scope.

   It is also interesting to highlight why it has been chosen to further split each ToU-state into sublevels, instead of considering a Markov chain with more states. There are two main reasons:

   - Markov models have the advantage to reproduce temporal correlation between adjacent time intervals. Nevertheless, it has been noticed that a five-state transition matrix already appropriately reproduces the autocorrelation. Bigger matrices do not lead to improvements in this context, but increase the computational and memory requirements.
   - Large transition matrices tend to show zero-valued entries. Zero-valued entries are unwanted, since they represent forbidden transitions between ToU-states, which is an unrealistic assumption. A model with relatively small transition matrices and a relatively large number of sublevels has been found to increase the quality of the output.

At this point, the generation of the Markov model for a single load profile is completed. The parameters are stored in a data structure defined hereafter.

6.4.3 Data structure of the model

For a single input load profile, the final outcome of the process is a data structure, depicted in Figure 6.3, composed by the following parameters:

- One transition matrix T^W for the weekly model, of size 3×3.
- For each state w of the weekly model, 96 transition matrices T_h^w of size 5×5 for the intra-day model are computed, $w \in \{H, M, L\}$, $h \in \{1, .., 96\}$. Each T_h^w matrix is combined with two data structures, providing additional information on the load range of ToU-states and the probability of occurrence of sublevels:
  - One array defining the delimiting values of load magnitude of each ToU-state.
  - One matrix of size 5×10 containing the estimated probability of occurrence of each of the 10 sublevel of the 5 ToU-states.

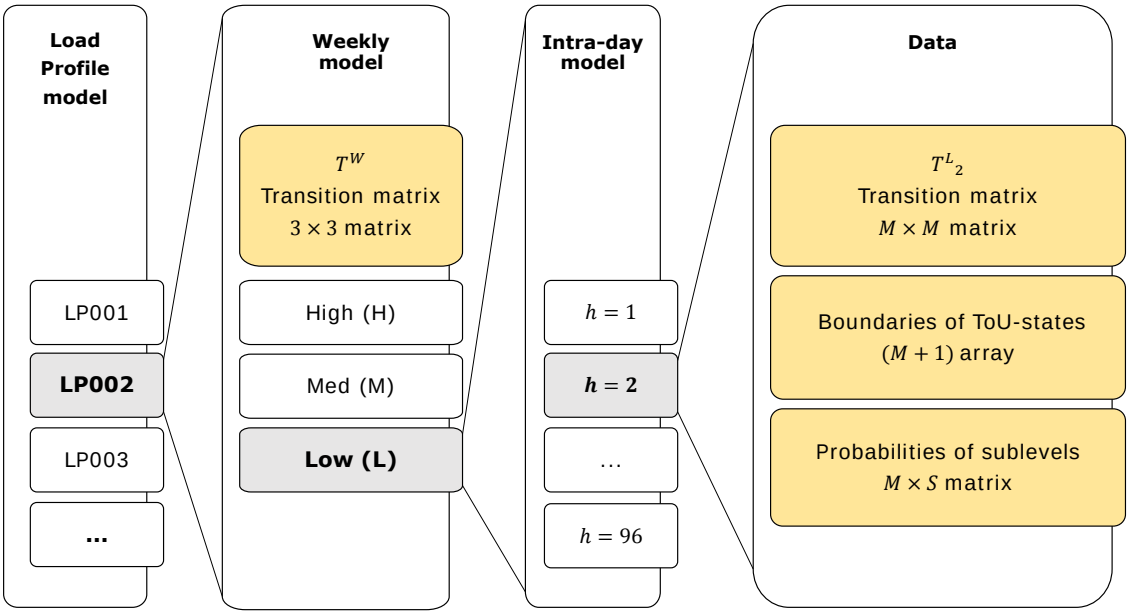


Figure 6.3: Data-Structure of the Markov Chain model. M : number of ToU-states, S : number of sublevels per ToU-state

6.5 Generation phase

The hierarchical structure of the model allows to generate a discretized load profile, where every data point is a signature indicating respectively the week-state, the ToU-state and the sublevel. In this case the ToU of the data point is automatically defined by the position of the data point in the time series.

The signature can be expressed in the form "$w.l.s$", where :

- $w \in \{\text{L, M, H }\}$ is the label of the week-state.
- $l \in \{1, 2, .., 5\}$ is the index of the Intra-day ToU-state.
- $s \in \{01, 02, .., 10\}$ is the index of the sublevel of the specific ToU-state.

Thanks to this structure, the generation of new profiles is straightforward. Matrices are randomly walked and states and load sublevels are assigned in a four step process:

1. **Generation of the discretized week-to-week profile:** By randomly walking the T^W matrix, a profile of week-states is obtained, i.e. a sequence such as:

   M, M, H, H, M, L, M, M, M, M, L, L, ...

   This sequence defines the week-state pattern of the SLP. The fist sample is chosen according to the probability distribution of the week-states in the original profile. It must be noted that since the number of steps in the random walk for a year is not very large, the deviation of the distribution of the three week-states in the generated profile compared to the distribution in the original time series can be quite high. Asymptotically, the distribution of synthesized transitions approaches the t_{ij}^W.

2. **Generation of the discretized 15-min profile:** Once the profile of week-states is determined, the intra-day load profile can be generated. For each week, the set of transition matrices of the according week-state, $T_{h,ij}^w$ is utilized for the random walk. The ToU-state of the first data point is chosen at random. ToU-states within a week are generated by randomly walking the transition matrices of the corresponding ToUs and the corresponding week-state. The first ToU-state of each week is determined by randomly walking the last transition matrix, from the last ToU-state of the previous week. To give a concrete example, a generated sequence may be:

   M.1, M.1, M.2, M.3, M.5, M.5, M.1, M.5, M.2, M.3, ...

   Representing the first ten ToU-states related to a week of medium week-state ($w =$ M).

3. **Sampling of the sublevel:** For the sampled ToU-state, a sublevel is sampled according to the estimated probability of occurrence. Recalling the previous example, the outcome at this stage may be:

   M.1.06, M.1.01, M.2.07, M.3.04, M.5.01, M.5.07, M.1.08, M.5.04, M.2.06, M.3.05, ...

4. **Back to continuous:** At this point the generation of the load profile is almost complete. The profile is still a discretized profile. To go back to a continuous profile, for each discrete data point a load magnitude is uniformly sampled within the range of the specific sublevel.

   As a numerical example, let's assume that for a generated data point at a certain ToU the generated signature is M.1.06. Let's assume that the signature corresponds to a load in the range 57–64Wh. The final load is then sampled randomly with uniform distribution in that range, and then rounded to the closest integer.

Chapter 7

Results and discussion

This chapter displays the results and validates the performance of the model. First, results are qualitatively displayed to provide a first grasp on the output of the model. Second, a more formal validation of is carried out thanks to the numerical indicators of performance and histograms defined in Section 3.2.2.

7.1 Qualitative validation

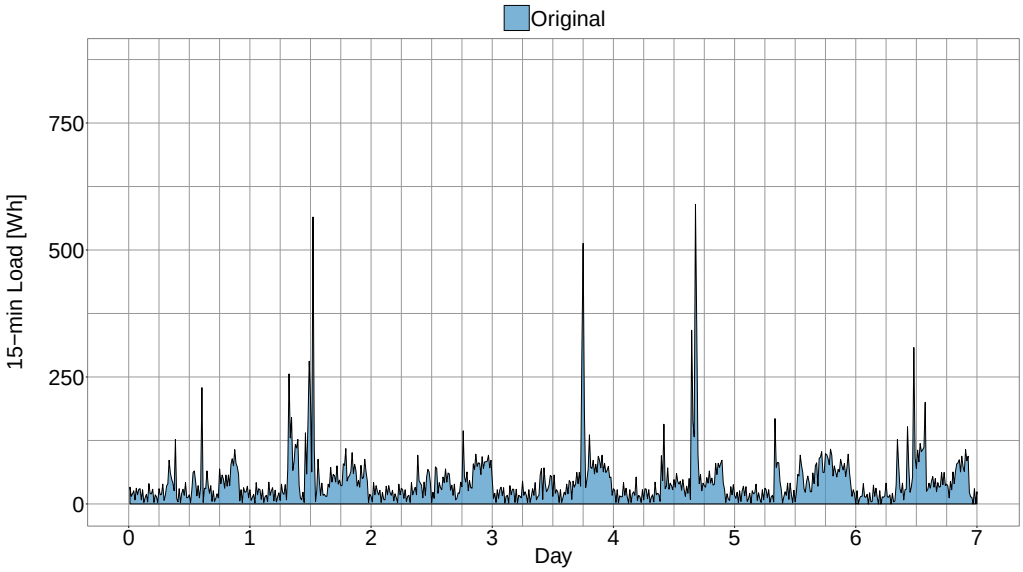
A first assessment of the goodness of the model is carried out by visually comparing the outcome of the model, when tailored on two different random load profiles, here labeled "LP1" and "LP2". As aforementioned, the model stores parameters to reproduce the behavior of the load profile for weeks of "High", "Medium" and "Low" week-state. Therefore, for the sake of completeness, for each of the two sample profiles, the outcomes for a sample week of each week-state are compared. It has to be noted that the visual and qualitative comparison of original and synthesized load profile cannot be used by itself to validate the model, but allows to intuitively understand some possible flaws. In the plots to follow, quantities related to SLPs are displayed in shades of orange, while quantities related to the original RLPs are displayed in shades of blue.

Please note that it is not desired to have a point-by-point reproduction of the patterns. Instead, the SLP should globally resemble the ensemble of patterns of the RLP. In particular, the distribution of the magnitude of peaks should be resembled. Conversely, it is not expected to have similar spikes on homologous days of the week, since the model, by design, does not separately model the days of the week.

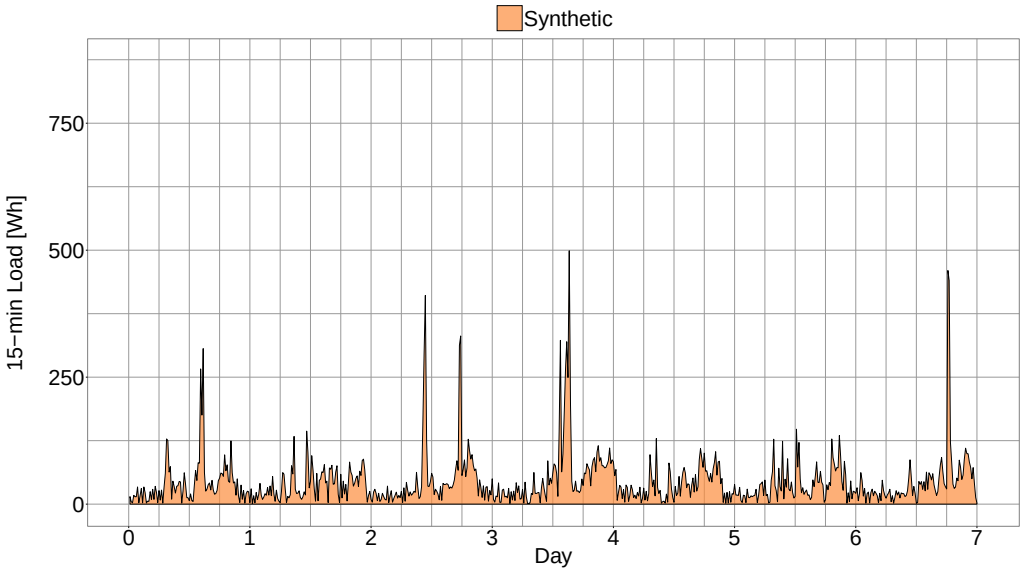
7.1.1 Comparison of patterns

Comparison of patters – LP1 Figure 7.1 compares an outcome of the model trained on the first sample load profile (LP1), for a week of high consumption ("high" week-state).

Globally, both amplitude and duration of peaks appear to be well reproduced. Please note that in the comparison spikes do not to appear in homologous days. This is a desired outcome of the model because, first and foremost, the SLP is not a forecast, therefore a reproduction point-by-point would be an undesired outcome. Second, by design the model does not treat days of the week individually, therefore it is normal to have mismatches



(a) Load Profile 01 – RLP – week of “High” week-state.



(b) Load Profile 01 – SLP – week of “High” week-state.

Figure 7.1: Load profile 01 – Comparison for a week of “High” week-state

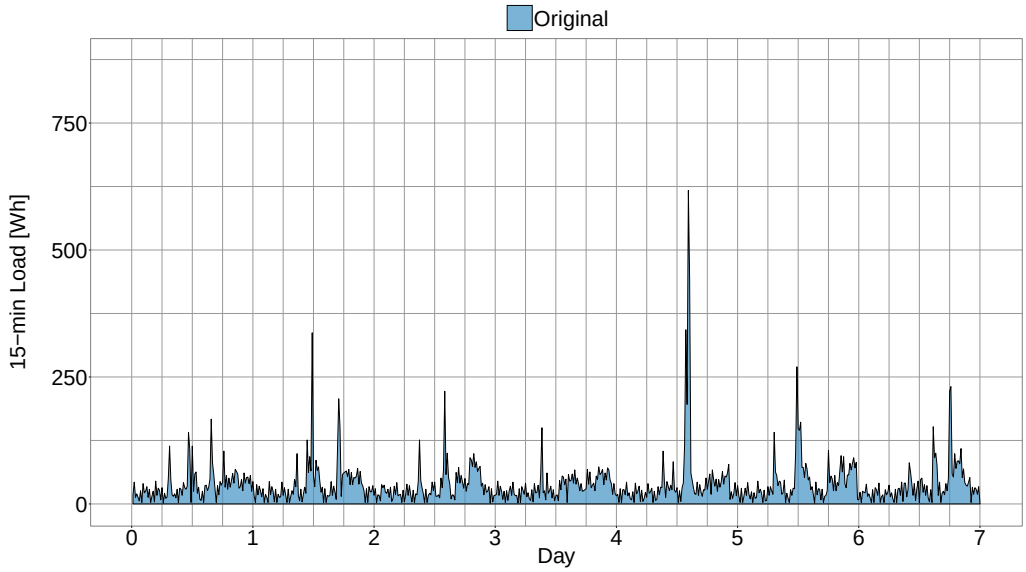
of patterns on homologous days. With the availability of longer RLPs, the model can be easily extended to individual day of the week. With these key concepts clear in mind it can be stated that the two outcomes of Figure 7.1a-7.1b appear to be very similar and have a good match, because daily patterns appearing in the RLP profile are also likely to appear similarly reproduced in the SLP. For example, the pattern of day 2 in the RLP of Figure 7.1a is reproduced in a similar fashion on day 4 of the SLP of Figure 7.1b. In the same way, the pattern of day 4 of the RLP is reproduced on day 7 in the SLP.

In the case of this specific load profile (LP1), patterns in the RLP show a high resemblance to a "block" structure, whereas the SLP shows a more "peaky" and variable profile. Similarly, while the RLP shows an intermittent and regular base-load (similar to the demand curve of a refrigerator), this pattern is not well reproduced in the SLP. A partial explanation is that the model have a memory of only one previous state. It is therefore intrinsically unable to reproduce patterns that appear cyclically with fixed amplitude and duration. This aspects is not crucial, because in the vast majority of residential load profiles, regular patterns appear only at night and have small amplitudes.

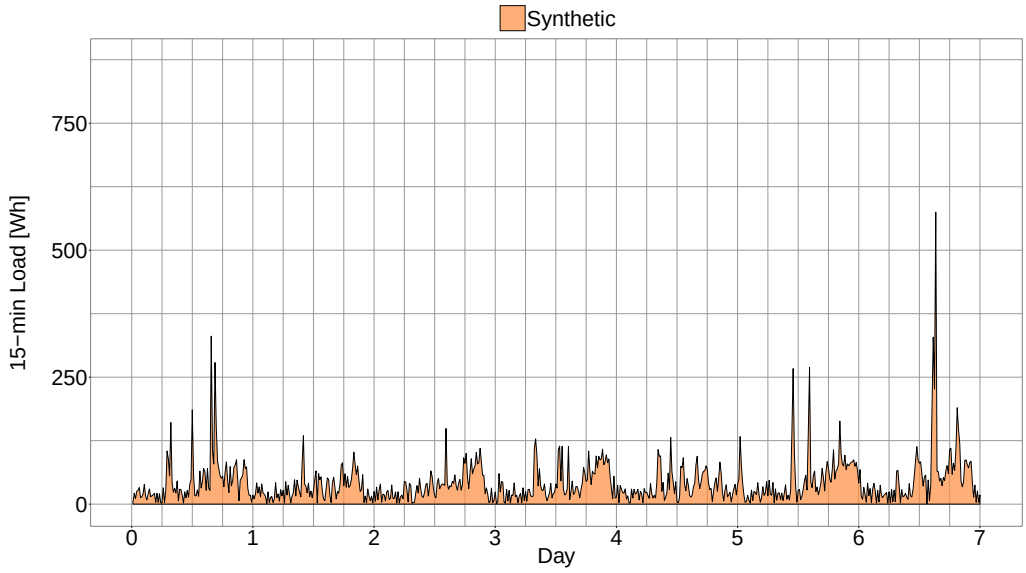
The not perfect performance in reproducing patterns at low load magnitudes can be explained by noticing that the load range of the small-amplitude pattern may fall within the load range of the lowest ToU-state of two adjacent ToU. Therefore, the algorithm will not identify the patterns as transitions between two different ToU-states and they will only be modeled by the probability of occurrence of sublevels, that does not model temporal correlation. This issue does not appear for patterns of higher load magnitudes, since larger variations of load are likely to be recognized as transitions between ToU-states, thus characterizing temporal correlation.

Figure 7.2 displays the outcomes for a week of medium week-state of LP1. First, it can be noticed how daily patterns for this medium week-state week are quite similar of the ones formerly displayed for the week of high week-state of Figure 7.1. This comes from the fact that some load profiles have quite constant weekly consumption over the year but the number of week-states of the model is fixed and equal to three. Under these conditions, the sets of ToU-transition matrices related to different week-states will tend to be very similar and so will be the outcomes. Once again, note how spike in day 5 of the RLP is successfully reproduced in a similar ToU on day 7 of the SLP.

Figure 7.3 compares original and synthetic profiles for LP1 in a week of low consumption. Here it can be noticed that even if the average load appear to be similar, the original profile shows less time-varying patterns, i.e. plateaus at different levels of consumption, whereas the synthetic profile is a continuous series of "peaks" and "valleys". This, once again, is a consequence of the short time-memory of the model.

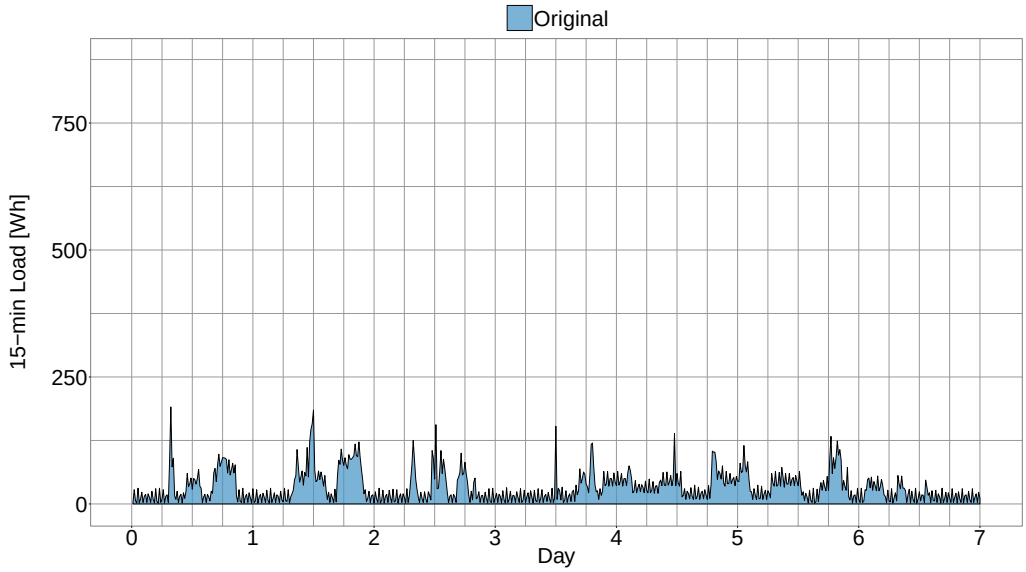


(a) Load Profile 01 – RLP – week of “Medium” week-state.

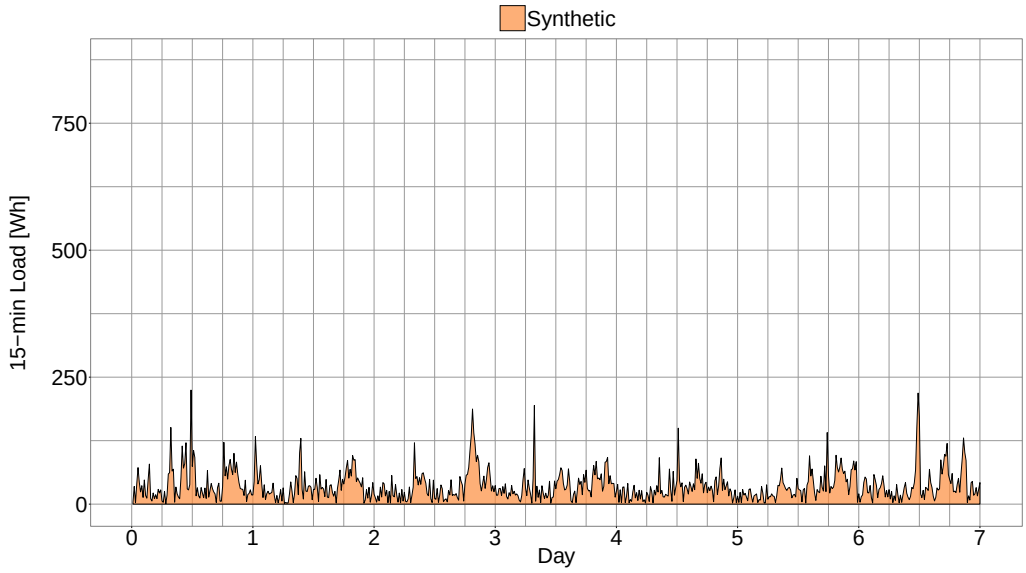


(b) Load Profile 01 – SLP – week of “Medium” week-state.

Figure 7.2: Load profile 01 – Comparison for a week of “Medium” week-state



(a) Load Profile 01 – RLP – week of “Low” week-state.



(b) Load Profile 01 – SLP – week of “Low” week-state.

Figure 7.3: Load profile 01 – Comparison for week of “Low” week-state.

Comparison of patters for LP2 Figure 7.4 shows the output of the model when reproducing a week of high week-state for the second sample load profile, LP2. It can be seen how the model manages to reproduce plateaus of highly time correlated load, at loads magnitudes of 80 Wh/15min, 200 Wh/15min, 280 Wh/15, resembling the RLP. Nevertheless, please note that the temporal duration of periods of high load is not reproduced, since the model has a memory of only one time step and it is therefore intrinsically unable to reproduce long patterns of uninterrupted high load.

To illustrate the concept, let's suppose to have a homogeneous Markov chain in which the transition probability of remaining in the highest state H is p. If at time $h = 0$ the process is in state H, then the probability of obtaining a pattern of n further consecutive states H is p^n. Even for values of p very close to 1 the probability p^n drops quickly. As a corollary, the higher the number of ToU in a day, the less probable is to reproduce long uninterrupted patterns of high consumption.

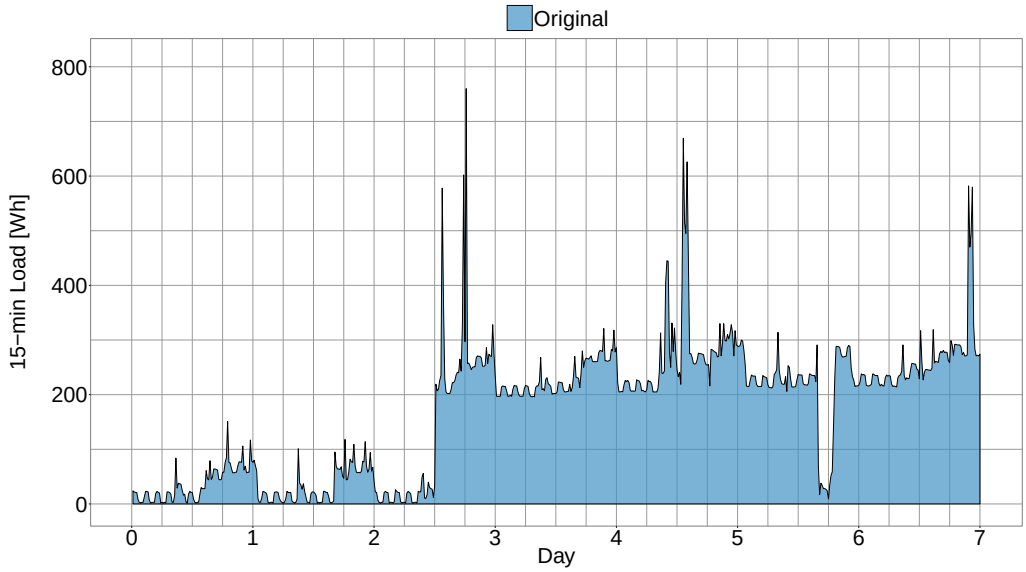
Figures 7.5 - 7.6 show the comparison for outcomes of weeks of medium and low consumption for LP2. In both cases, no uninterrupted patterns of high consumption are present in the RLP, therefore the performance of the model increases. Magnitude and frequency of spikes appear to be well reproduced, patterns at intermediate load magnitudes also appear to be similar, patterns at low load levels are not well reproduced, but this does not represent an issue.

7.1.2 Comparison of quantile Typical Load Profiles (q-TLPs)

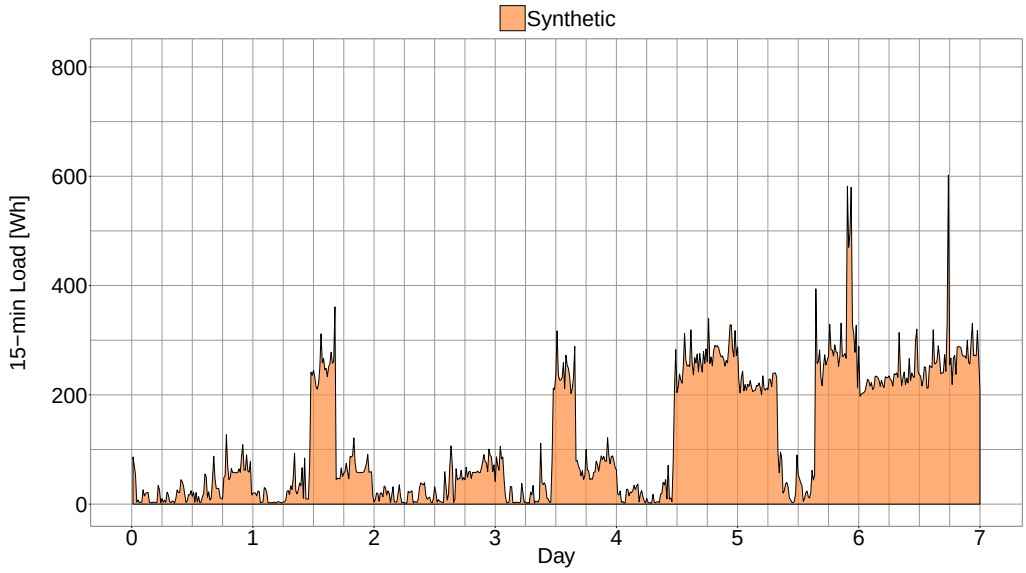
Another qualitative way of comparing the two outcomes is to display the quantile Typical Load Profiles (q-TLP), defined in section 3.2.2. A q-TLP of a good Synthesized Load Profile has to resemble the q-TLP of the original one. As already stated in section 3.2.2, if there is a close resemblance between the two plots, this suggests that also other indicators are likely to be similar.

Figures 7.7–7.8 display the comparison of the q-TLPs for the two sample profiles LP1 and LP2. It is evident that the q-TLP of both SLPs closely resembles the one of the respective RLPs.

It extremely important to remark that it is not necessary to implement a Markov-chain model to generate a profile having the same q-TLP of a given profile. It would be sufficient to generate the profile by sampling load at each ToU according to the distribution of the magnitude of load in the original profile in the homologous ToU. Nevertheless, in this way most of the information about the correlation between two adjacent time intervals would be completely lost, and this is the most important advantage of using a Markov chain model.

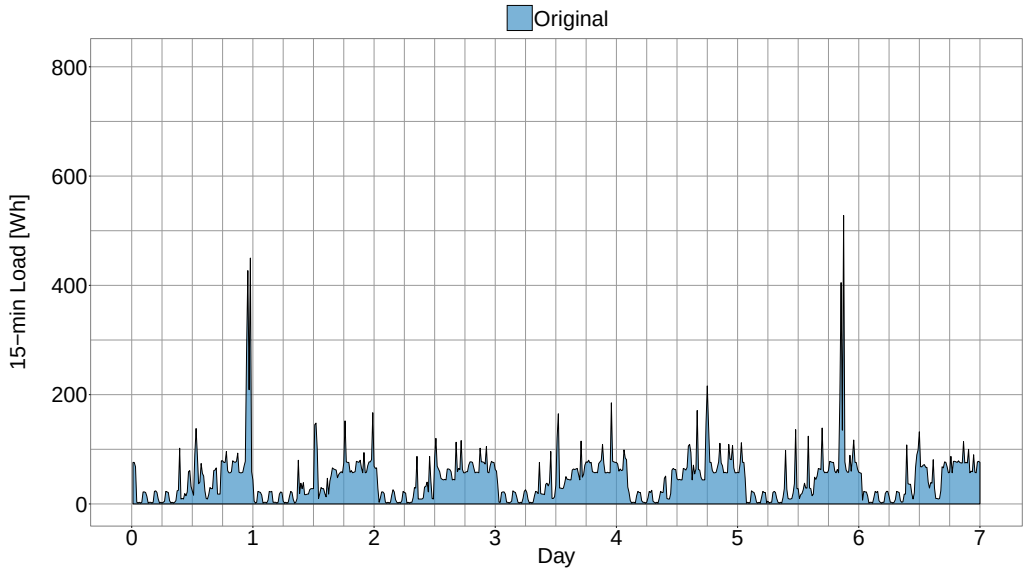


(a) Load Profile 02 – RLP – week of “High” week-state.

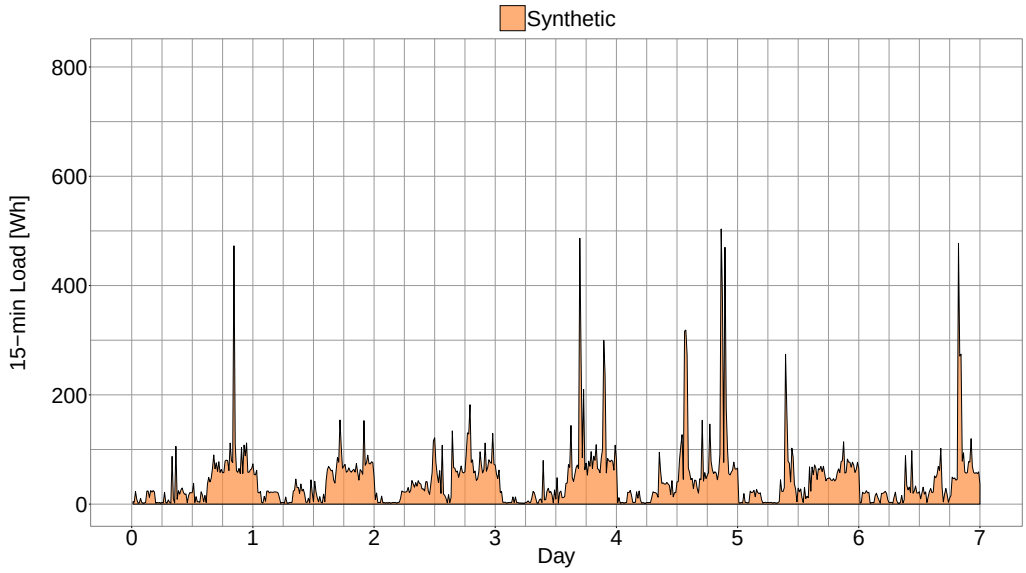


(b) Load Profile 02 – SLP – week of “High” week-state.

Figure 7.4: Load profile 02 – Comparison for a week of “High” week-state.

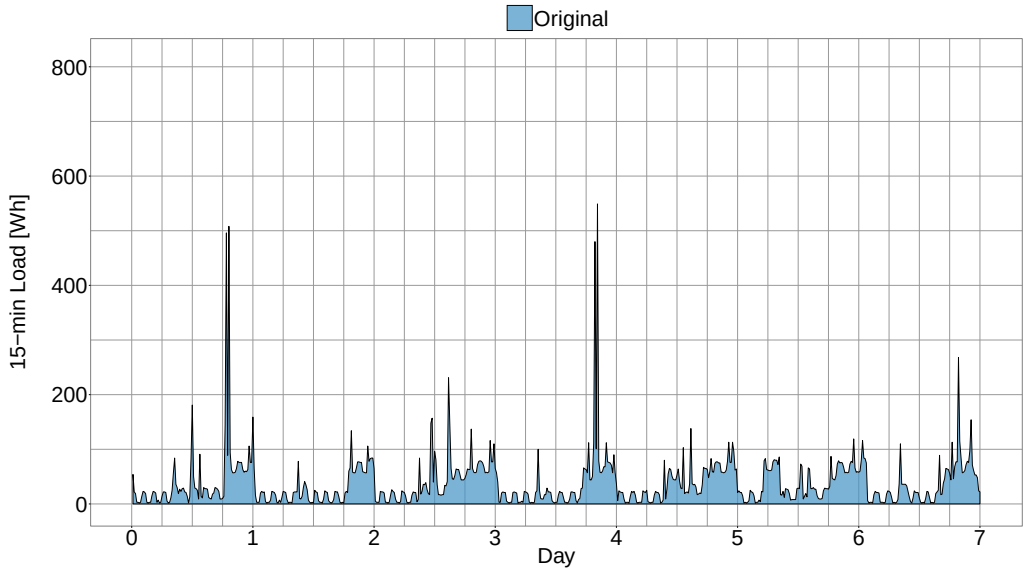


(a) Load Profile 02 – RLP – week of “Medium” week-state.

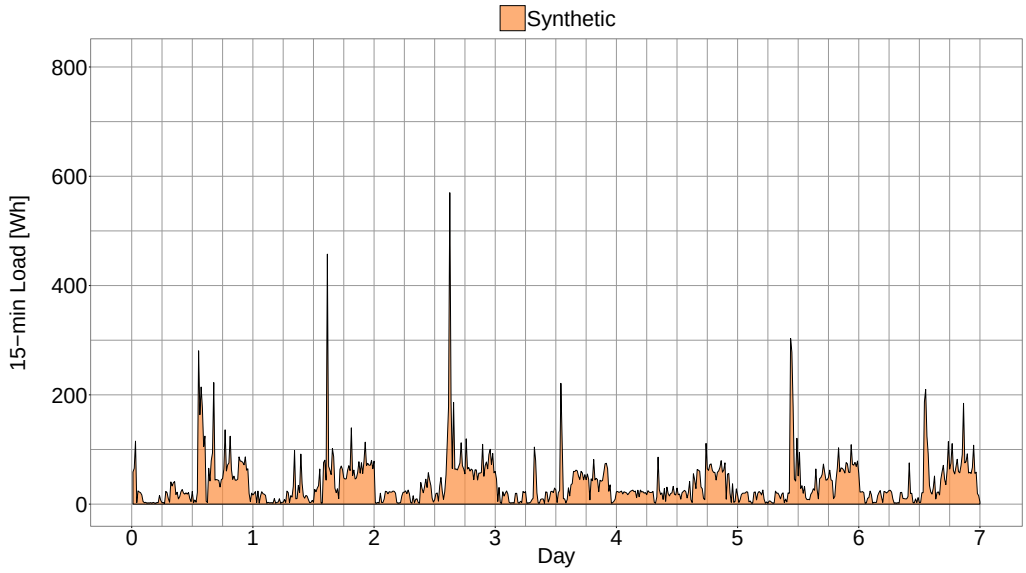


(b) Load Profile 02 – SLP – week of “Medium” week-state.

Figure 7.5: Load profile 02 – Comparison for a week of “Medium” week-state.

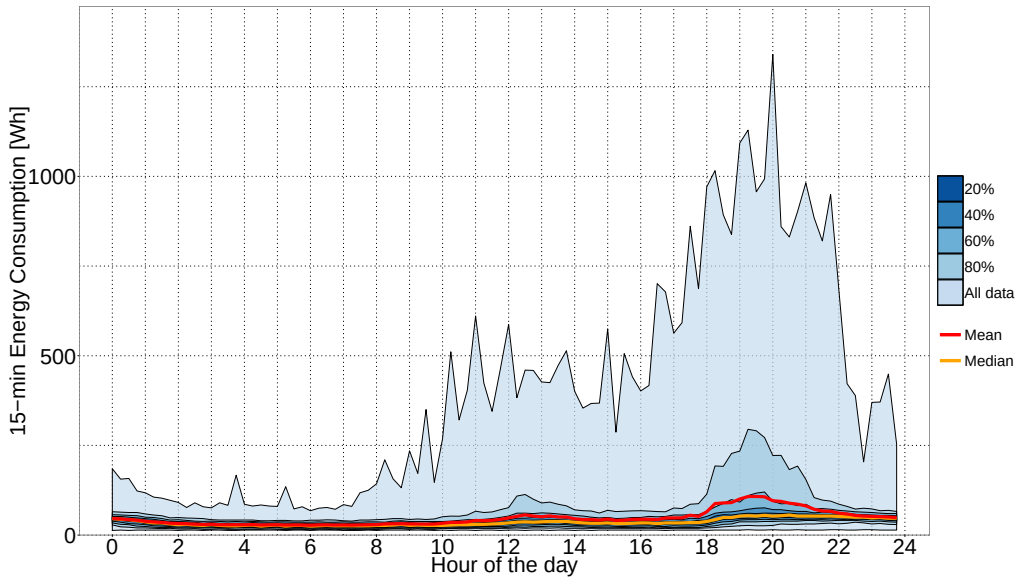


(a) Load Profile 02 – Original – week of “Low” week-state.

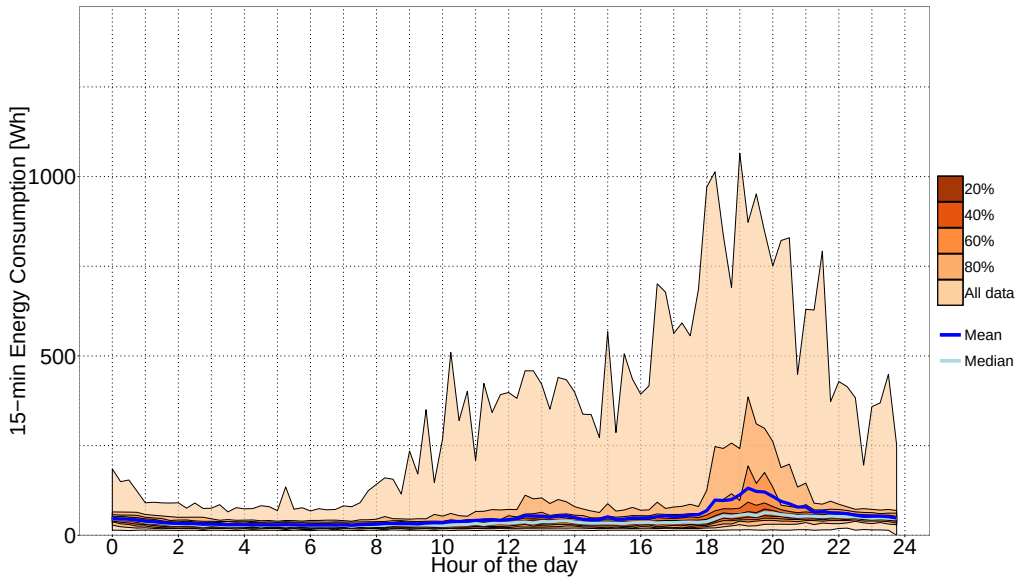


(b) Load Profile 02 – Synthetic – week of “Low” week-state.

Figure 7.6: Load profile 02 – Comparison for Week of “Low” week-state.

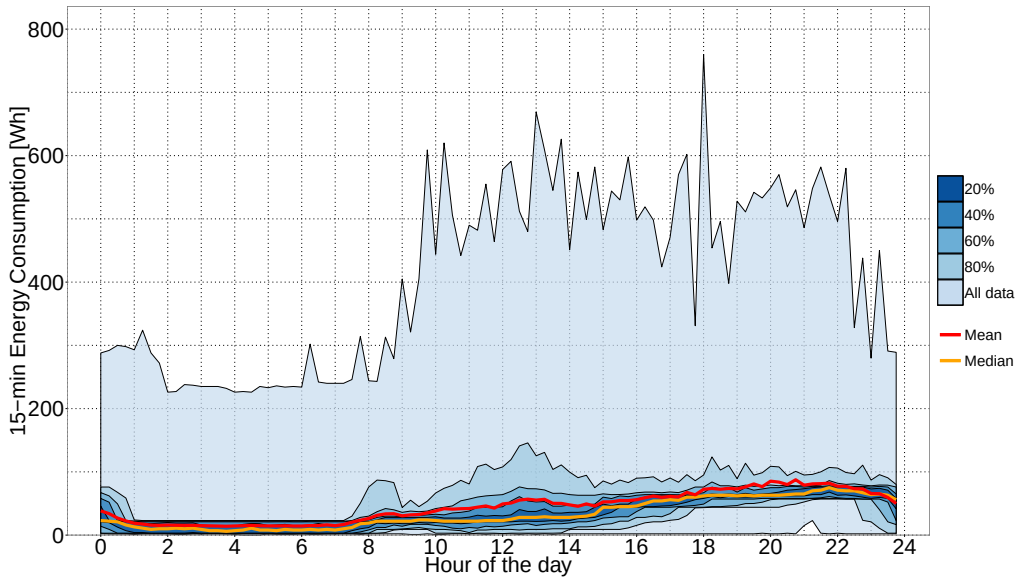


(a) Load Profile 01 – RLP – q-TLP chart.

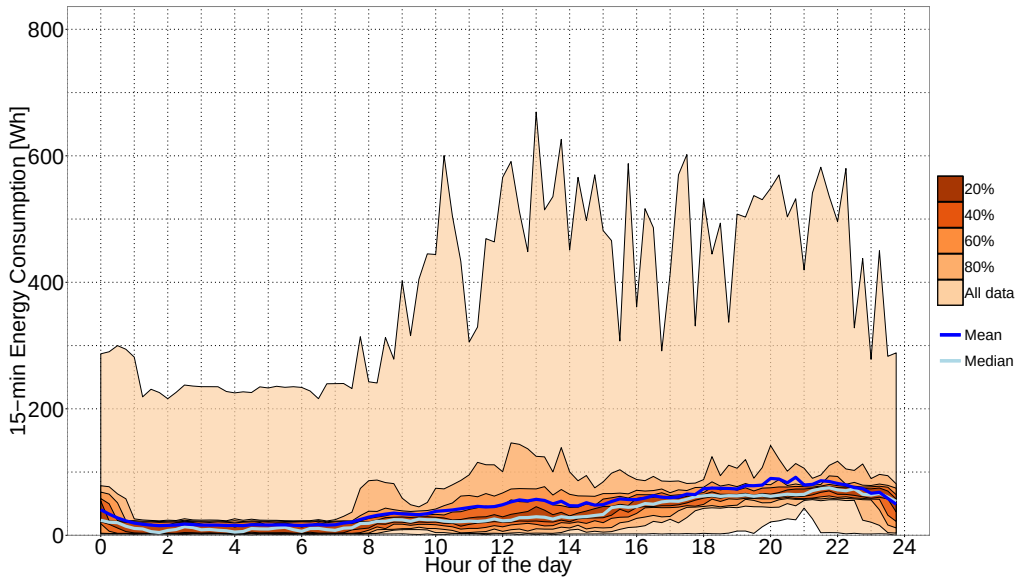


(b) Load Profile 01 – SLP – q-TLP chart.

Figure 7.7: Load profile 01 – Comparison of q-TLP charts.



(a) Load Profile 02 – RLP – q-TLP chart.



(b) Load Profile 02 – SLP – q-TLP chart.

Figure 7.8: Load profile 02 – Comparison of q-TLP charts.

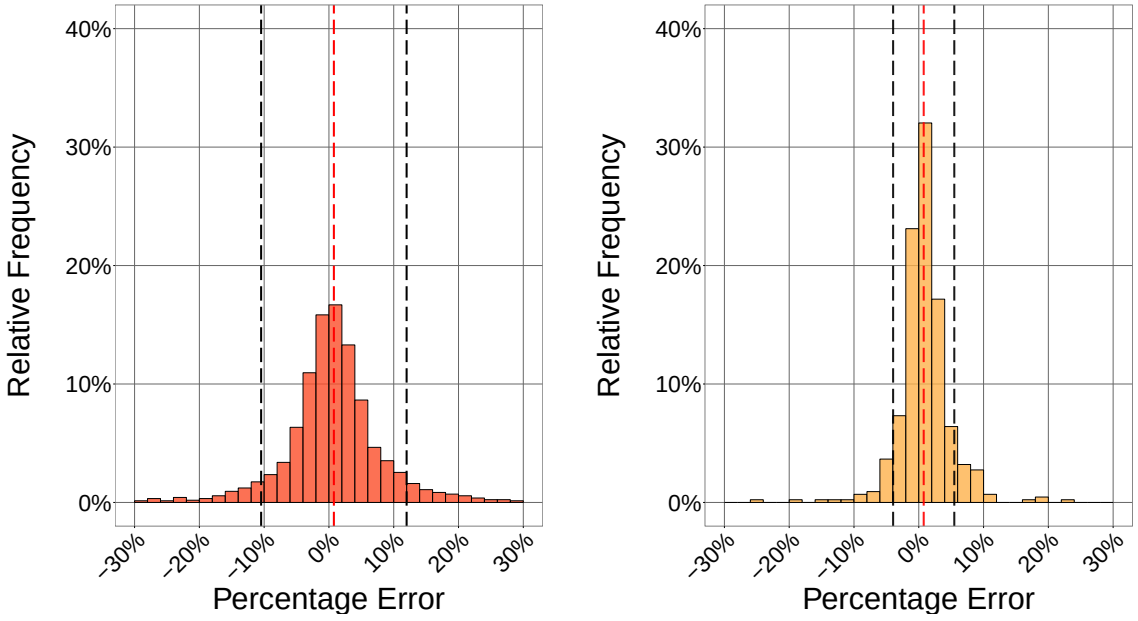
7.2 Quantitative validation

After the qualitative validation of two sample results gave a first grasp of potentialities and limits of the model, this section focuses on a quantitative and more rigorous validation of the generated profiles. The validation process indicates that in most of the cases the algorithm performs well in reproducing all indicators, with poorer performances only in case of profiles strongly correlated with temperature.

7.2.1 Error on average annual consumption

Figure 7.9 displays the distribution of the percentage error of the yearly consumption of the SLPs, compared to the yearly consumption of the corresponding 1-year long RLPs. The distribution of the error has been evaluated with the following steps:

- A Markov-chain model is tailored on each one of 400 different RLPs.
- Each Markov model is run to generate one 1-year long SLP and one 5-year long SLP.
- The annual consumption is evaluated for each RLP and SLP. In case of 5-year long SLPs, the average yearly consumption is evaluated.
- The percentage error on yearly consumption (or average annual consumption for multi-year SLP) is evaluated for each couple of RLP-SLP.



(a) Distribution of error on yearly consumption.

(b) Distribution of error on 5-year mean yearly consumption.

Figure 7.9: Distribution of error on yearly consumption.

Figure 7.9a reports the distribution of the percentage error for 1-year long SLPs, with the width of bins set at 2%. The red-dashed line represents the mean of the distribution, the two black dashed lines indicate one standard deviation from the mean. As it is possible to infer from the picture, the mean error is less than 1%, therefore the model can be considered as unbiased. Almost the totality of the mass of the distribution falls in the interval of ±20%. It is worth to notice that a certain variability in the error is a desirable feature of the SLP. In fact, the annual consumption of an individual customer is a random variable. Since data at disposal were collected for only a year it was not possible to estimate mean and variance of the random variable, but this consideration is enough to point out that a certain variance in the distribution of the percentage error is a realistic outcome of the model, i.e. a model that outputs SLPs with a perfect match in annual consumption would reproduce a constant and unrealistic behavior of the consumers.

If each SLP is generated to cover a time span of five years, the yearly fluctuations in annual consumption tend to balance out and the resulting average annual consumption is much closer to the original one, with a smaller standard deviation (Figure 7.9b). In the 5-years case, almost the totality of the mass of the distribution falls in the range ±10%.

It is worth to point out that for the central limit theorem (see section 2.2.3) the standard deviation of the error on the annual consumption is expected to decrease according to the square root of the number of years considered for the average, i.e. the standard deviation of the percentage error evaluated considering the average annual consumption of a 5-year SLP is expected to be $\sqrt{5} \approx 2.24$ times smaller than the standard deviation of the error evaluated for the 1-year consumption.

7.2.2 Worst-case vs average-case analysis

The Markov chain model can generate an arbitrary number of load profiles. Many typologies of load profiles are very well reproduced by the model, for others the performance is poorer. Validation has been carried out by comparing various feature-ensembles:

- Histogram of distribution of magnitude of load.
- Histogram of distribution of magnitude of daily peak 15-min load.
- Histogram of distribution of Time of Use of daily peak 15-min load.
- Autocorrelation with 10 days lag.

Two approaches for validation have been considered:

- **Validation by global performance score:** One option for validation could be to consider a large number of RLPs and assign a score to the corresponding SLPs, according to how well they resemble the original. The score can be evaluated by weighting the performance on different feature-ensemble of the SLP. The distribution of the scores can then be displayed. Nevertheless, this approach has been discarded. In fact, the score itself should be tailored on the task in which the synthesized time series are needed.

  For example, if SLPs are used to test a control algorithm for storage, accuracy in reproducing the time and amplitude of spikes of peak demand is more important

than reproduce accurately the complete load distribution. On the other hand, for the estimation of the load factor of a single building, a good resemblance of the load histogram is very important, while a good reproduction of the time of use of spikes is not.

- **Validation by worst-case vs average-case analysis:** Instead of evaluating a total performance score, it is possible to compare the outcomes of samples showing average performances vs worst performances, concerning some validation parameter (or parameters). This method allows to define the range of performance that is reasonable to expect from the output of the model.

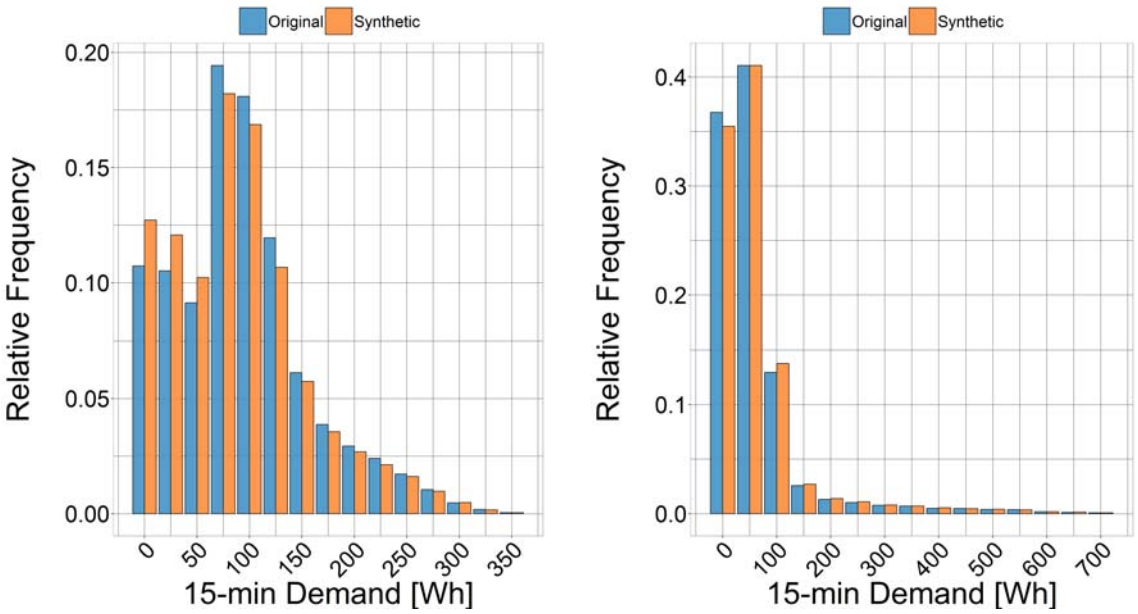
Validation by worst vs case-average case analysis has been chosen as validation method. It has been further chosen to separately analyze the performance for different type of indicators. The worst and average cases has been chosen using the following criteria:

- Worst-case performance has been selected by visual inspection of the validation parameters for a set of 400 sampled SLPs, each one resembling a different RLP. One worst-case example has been chosen for each of the indicators chosen for validation.
- Average-case has been selected by randomly sampling 10 RLPs from the initial LPset and by building a model on each of them. Then, the indicators of the SLPs are evaluated. The SLP with the median performance is then selected for display.

7.2.3 Comparison of load histograms

Distribution of load is highly skewed and a histogram embracing all values of load would be equally skewed and difficult to display. For this reasons, in the following plots the x-axis is split into 15 bins, of a width such that at least 97% of data are represented in the histogram. The performance at high loads is measured in the following by the histogram of magnitude of peak loads.

7.2.3.1 Magnitude of load



(a) Magnitude of load - worst performance, LP415.

(b) Magnitude of load - average performance, LP010.

Figure 7.10: Histograms of distribution of load magnitude - worst vs average performance.

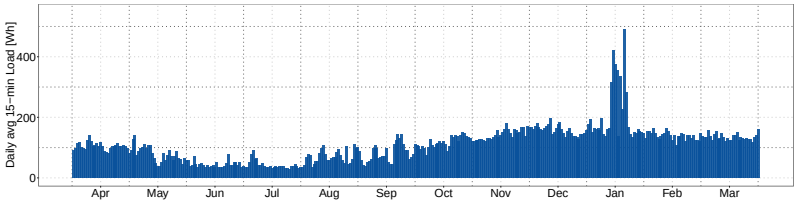
Figure 7.10 shows a comparison of the histograms of distribution of load magnitude between the worst and average performance. The x-axis shows the magnitude of the 15-min load. Y-axis shows the relative frequency of the magnitude of the load, taken over a year. It can be noticed how in the worst case the relative frequency is overestimated at low load magnitudes and underestimated at higher load, with maximum error that still remain within 20%. On the other hand, the average performance of Figure 7.10b displays a very close resemblance, with errors within 10%.

To better understand the reason of this difference in performance, it is worth to observe the load profiles on which the model is tailored.

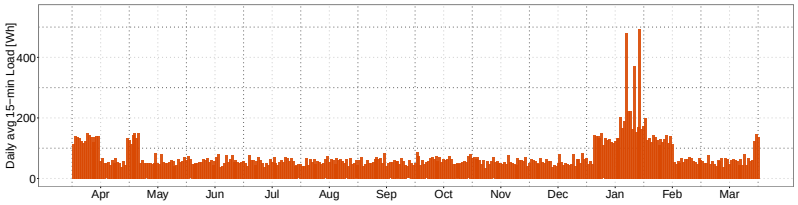
Worst Performance: Figure 7.11a reports the daily-mean 15-min consumption of the profile LP415 over the year. It is immediately clear that in this case consumption is correlated with the season, with lower consumption during the summer and higher during winter. The Markov model is by design neglecting this trend. In fact, Figure 7.11b shows a step shaped profile, the three week-states are evident. In this case, the bi-level structure led to poor results.

Average Performance: Figure 7.12a reports the measured daily consumption of a profile leading to average performance (LP010). Differences with the previous profile are evident: consumption in LP010 is much less correlated with the time of the year. The profile shows peaks in consumption every two weeks. The synthetic profile of figure 7.12b

qualitatively resembles much more the original. Quantitatively, the distribution of load magnitude is almost identical to the original (Figure 7.10b. Please note how the bi-week peaks in the RLP are not reproduced. Once again, this is expected, since the model does not distinguish between different days of the week.

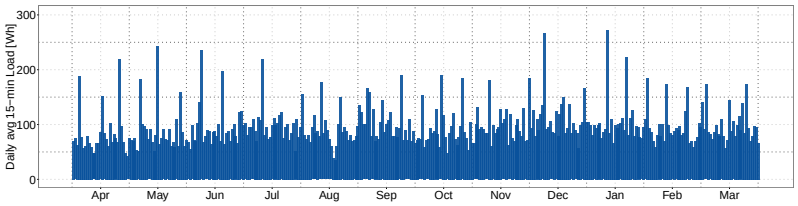


(a) LP415 - RLP - Note the dependence on the time of the year.

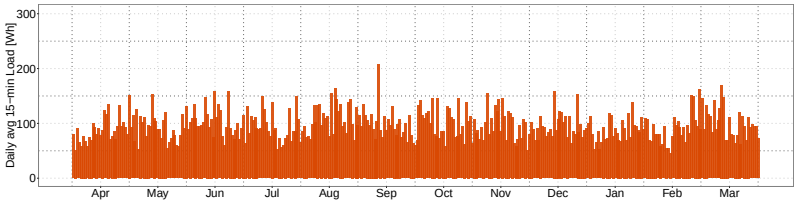


(b) LP415 - SLP - Note that the model performs poorly in reproducing long term trends in consumption.

Figure 7.11: LP415 – Worst performance in distribution of load magnitude.



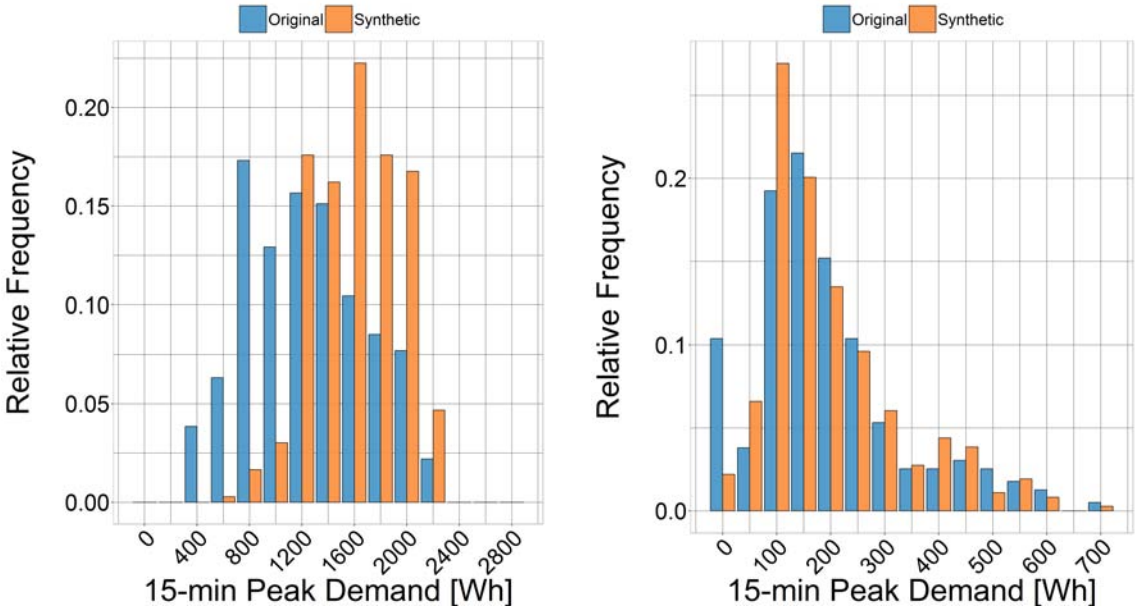
(a) LP010 - RLP - Note that the profile is more stationary. Note biweekly spikes.



(b) LP010 - SLP - Note that biweekly spikes are not reproduced.

Figure 7.12: LP010 – average performance in distribution of load magnitude.

7.2.3.2 Magnitude of daily peak load



(a) Daily peak load - worst performance, LP395.

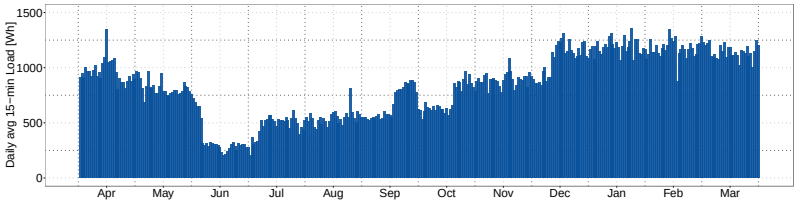
(b) Daily peak load - average performance, LP001.

Figure 7.13: Histograms of distribution of magnitude of daily peak load - worst vs average performance.

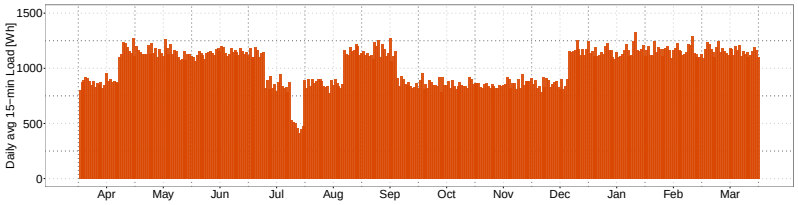
The distribution of the magnitude of peak load is important for transformer rating and evaluation of load factor and coincidence factor. Peak loads, in case of one sample every 15-min, represents a fraction of 1/96 of the data, i.e. slightly more than 1%. Therefore, they are by nature less frequent events, the sample set is much less populated (only 364 data points for a profile of 52 weeks) and it is reasonable to expect higher differences in distribution. Moreover, it is important to notice that the property of being a "peak" load depends by the magnitudes of all 96 samples in a day. Therefore, two profiles with the same distribution of magnitude of load can have two completely different distribution of magnitudes of peak loads. Figure 7.13 reports on the $x-$axis the magnitude of the daily peak load, and on the $y-$axis the relative frequency. Histograms embrace 100% of data.

Worst performance: Figure 7.13a shows the worst performance in the validation set, peak loads are misplaced and the mode of peak load of the SLP is almost three times the mode of the RLP. By observing the original load profile in Figure 7.14a it is clear that also in this case the consumption is dependent on the time of the year, consumption is highly varying throughout the weeks. In this case, the clustering of weekly consumption into three week-states led to a coarse approximation in the case of the synthetic profile. It is worth to notice how in this case the random walk through the weekly transition matrix led to a year with a high number of high consumption weeks. Once again, the main cause of the bad performance is judged to be the strong dependence of load magnitude on the period of the year.

Average performance: Distributions of peak load reported in Figure 7.13b for profile LP001 display a close match between RLP ans SLP. At intermediate peak loads the error is within 25%. At the lowest magnitude, a peak is mismatched. This can be explained by observing Figure 7.15, displaying the two load profiles. The period of low consumption of the original profile has been reproduced, but of shorter duration in the SLP (this is a contingency of the weekly random walk). Therefore, the RLP has more days of lower consumption, hence the higher frequency of peak loads of low magnitude.

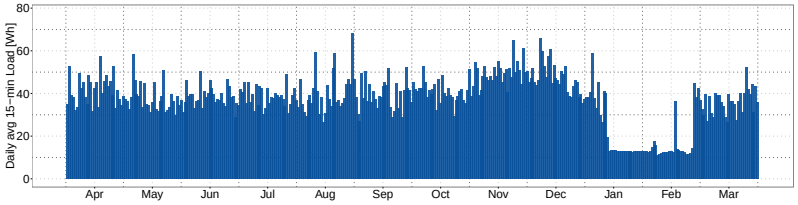


(a) LP395 - RLP - Note the dependence on the time of the year.

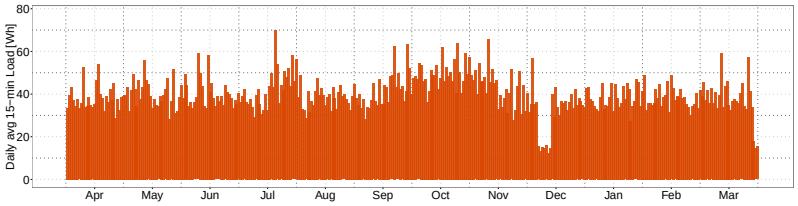


(b) LP395 - SLP - Note the clear effect of the subdivision into three week-levels.

Figure 7.14: LP395 – Worst performance in distribution of daily peak load.



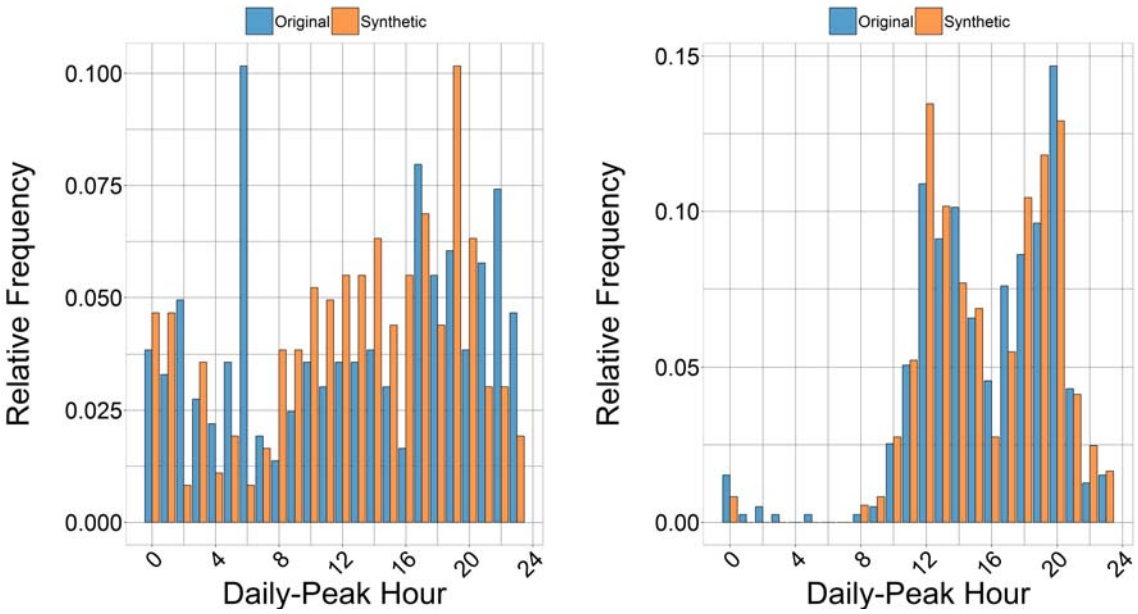
(a) LP001 - RLP - Note the period of low consumption.



(b) LP001 - SLP - Note that the zone of low consumption has been reproduced of a shorter duration.

Figure 7.15: LP001 – Average performance in distribution of daily peak load.

7.2.3.3 Time of Use of daily peak load



(a) Time of Use of daily peak load – worst performance, LP042

(b) Time of Use of daily peak load – average performance, LP005

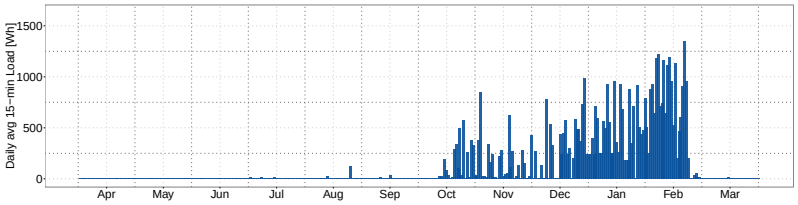
Figure 7.16: Histograms of distribution of Time of Use of daily peak – worst vs average performance

A proper resemblance of the Time of Use of the daily peak load is important to ensure that the aggregation of many SLPs behaves in the same way as the aggregation of the original RLPs, with respect to the coincidence factor. Histograms of Figure 7.16 display on the $x-$axis the 24 hours of a day and on the $y-$axis the relative frequency of peaks occurring in that hour.

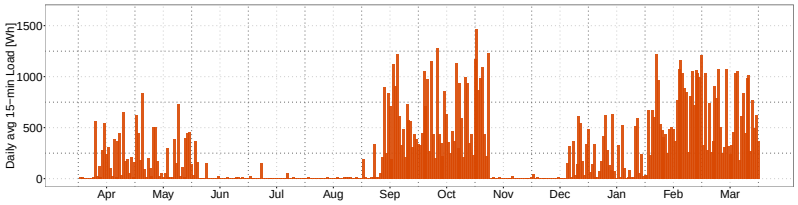
Worst Performance The worst performance of the model has been produced in sample LP042, where the SLP shows a 50% of error for peaks in the period 8am-4pm and completely misses the peak at 6am. This poor performance is easily explained by observing Figure 7.17a, the original load profile shows a consumption of almost zero from March to October, while consumption increases during the winter period, most likely due to electric heating, but also to partial seasonal occupancy of the dwelling. Since the histogram displays information regarding all peaks during the year, it is reasonable to assume that peaks at 6am are due to negligible peaks during this near-zero consumption period. This obviously influenced the relative frequency of peaks in other periods of the day. Once again, this poor performance is a consequence of the fact that the model is by design not considering seasonal variability and exogenous variables, such as temperature. Suggestions on how improve the model regarding this issue are reported in section 7.3.2

Average Performance If the original profile is more stationary, with more constant mean and standard deviation, as it is the case for LP005 shown in Figure 7.18b, the

accuracy of peaks hours greatly improves, as illustrated by the histogram of Figure 7.16b.

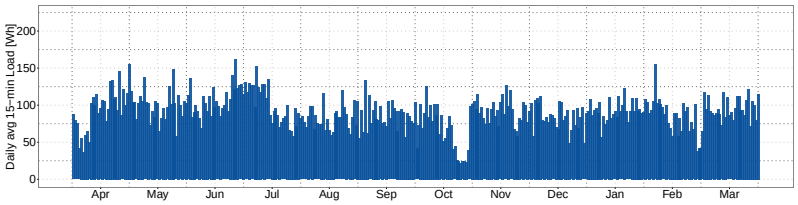


(a) LP042 – RLP – Note that consumption is correlated to the period of the year, negligible consumption from March to October

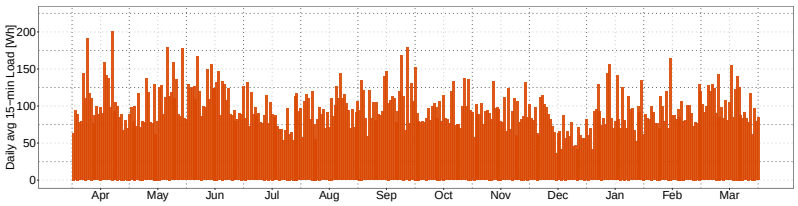


(b) LP042 – SLP – Note the reproduction of the three levels of consumption

Figure 7.17: LP042 – Worst performance in Time of Use of daily peak load.



(a) LP005 – RLP – Note that consumption is not strongly dependent on the time of the year.



(b) LP005 – SLP – Note the good reproduction of weekly variation

Figure 7.18: LP005 – Average performance in Time of Use of daily peak load.

7.2.4 Autocorrelation

The autocorrelation expresses the resemblance of the load profile within shifted instances of itself. For a discrete-time time series, it has been more formally defined in section 3.2.2. It is a validation method that has been utilized in other studies to verify the goodness of the model [24]. Since the autocorrelation is a function of the lag time, also in this case, for the sake of clarity of visualization, the valuation has been carried out by a Good-Average-Worst performance analysis. in the following figure, the value of the autocorrelation is displayed time steps of 15 minutes lag, with a maximum lag of 10 days. The autocorrelation at Lag = 0 has been dropped, since it is equal to one by definition. After an initial peak, the typical shape of the function for most load profiles resembles a series of "V"s, see Figure 7.19. This "VVV" shape highlights the positive correlation of the consumption with the level of consumption exactly 24 hours later, and the negative correlation with consumption at odd multiples of 12 hours lag. Obviously, this is due to the difference in consumption between day and night.

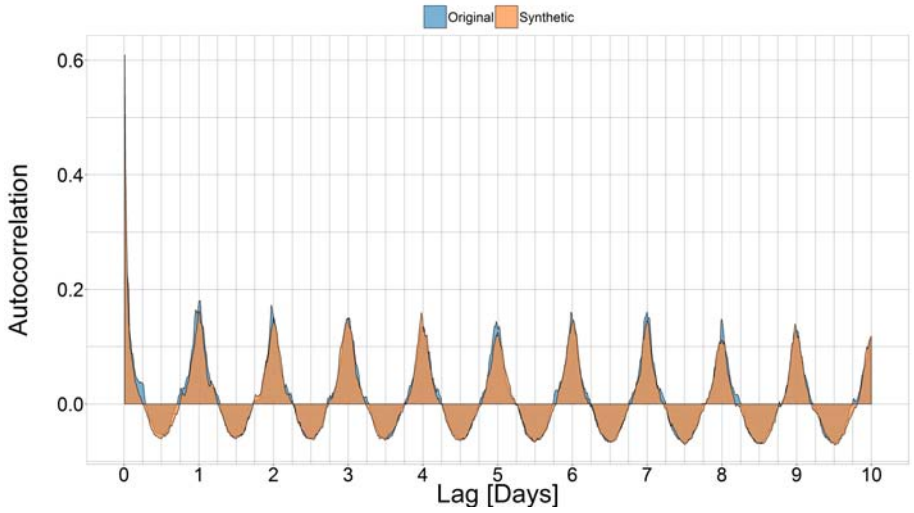


Figure 7.19: Autocorrelation – LP002 – Average performance

Average Performance Figure 7.19 shows an average performance of the model, the corresponding RLP has a typical autocorrelation function. Although the overall autocorrelation in not extremely strong, as it is possible to notice, the autocorrelation of the SLP is almost identical to the one of the original.

Worst Performance Figure 7.20 displays the worst performance of the algorithm within the validation set of SLPs. First, the autocorrelation at some time steps does not exactly match the one of the original load profile, with errors of around 30%. Second, it is possible to notice how the spike in autocorrelation at Lag = 7 days could not be reproduced. This result was expected, since consumption is more correlated within lags of exactly one week, i.e. people can have specific habits and practices in homologous days of the week. The model does not incorporate sub-models for individual days of the week, although with

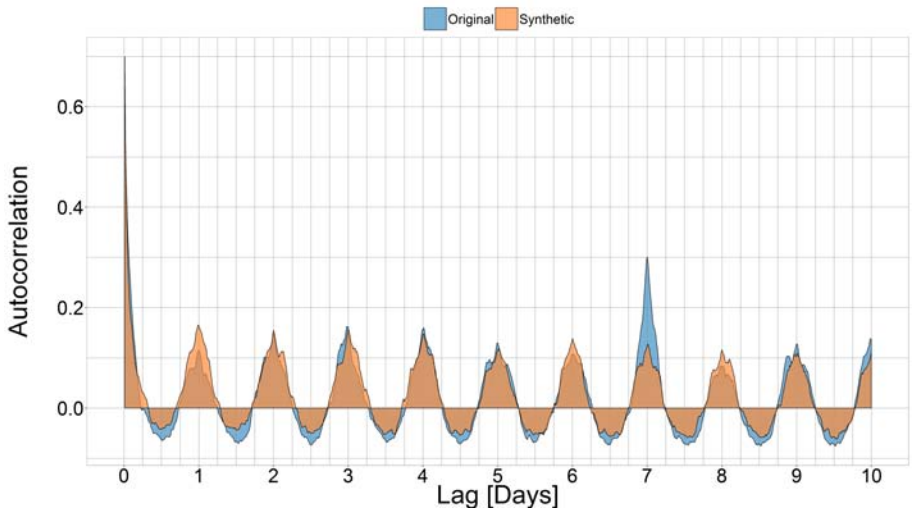


Figure 7.20: Autocorrelation – LP068 – Worst performance

availability of longer time series this could be fruitfully implemented. Since the example of Figure 7.20 is the worst found in the sample set of 400 load profiles, it can be concluded that the algorithm performs acceptably well on the whole dataset.

Good Performance Figure 7.21 shows how the autocorrelation is reproduced properly even when the original load profile is characterize by a more cumbersome autocorrelation function. Peaks of positive and negative correlation are well reproduced, except for the usual peak at 7 days of lag.

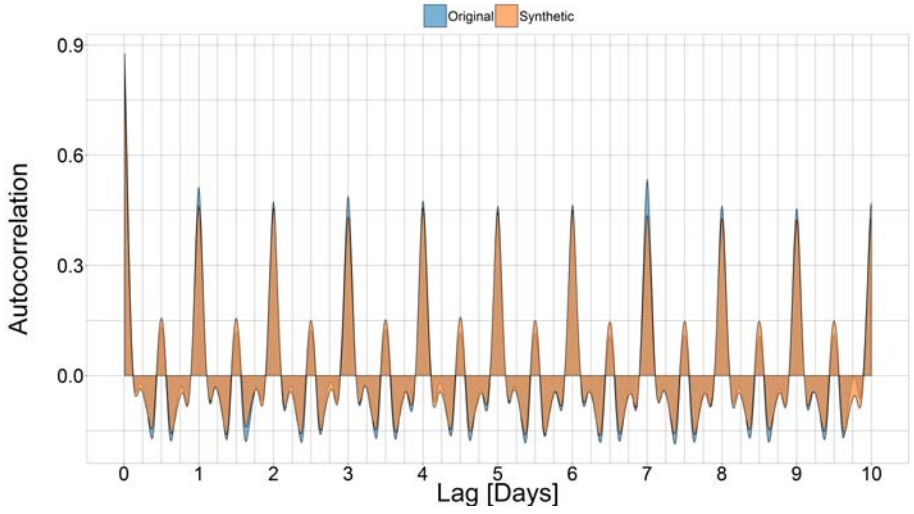


Figure 7.21: Autocorrelation – LP099 – Good performance

7.3 Strengths and weaknesses

This section summarizes the strengths and weaknesses of the model that emerged from the validation.

7.3.1 Strengths

- The model achieves significant performances in the average case, which is the most important to be considered, regarding the reproduction of annual consumption, distribution of load magnitude, distribution of peak load magnitudes. In particular, the model performs especially well in the reproduction of autocorrelation. Reproduction of distribution of the Time of Use of peak load shows a good average performance, but reports poor results in the worst-case. Nevertheless, poor performances has been found to be limited to RLPs with high seasonal variations.
- The model is easy to interpret and with only three tuning parameters: the number of week-states, the number of ToU-states and the number of sublevels. This facilitates the choice of the optimal configuration.
- Given a larger availability of data, the model can be easily extended to individually characterize each day of the week.
- The model allows the generation of an arbitrary large number of SLPs that can be stored separately and utilized at need.
- The model is purely data-driven, i.e. it does not assume beforehand any specific probabilistic distributions of load magnitudes or peak loads.
- Patterns at high load magnitudes are well reproduced in magnitude and frequency of appearance. Temporal duration of peaks is less accurately reproduced.
- The model is easily parallelizable and in can be easily implemented to run on multi-core machines or cloud computing, because of the modular structure (every individual model is trained on only one RLP). In fact, each model is tailored only on the input RLP. The training of the n different models using n different input RLPs can be independently carried out on n different processing units.

7.3.2 Weaknesses

- The model is not designed to reproduce seasonal patters and to take into account the effect of temperature. In the contingent case of the implementation for the City of Basel, this was a minor issue, because of the low penetration of electric heating and air conditioning devices, but it has to be taken into account in case of implementation in different context. Two solutions are suggested:

    1. Develop a filter able to detect whether a RLP is suitable to be reproduced with the Markov model by analyzing seasonal variations, for example with Fourier analysis. This method allows to implement the model only on RLPs that can ensure a minimum level of accuracy, thus improving reliability of the outcome.

2. Integrate the effect of exogenous variables in a preprocessing step. For example, a regression model could remove the temperature dependent component from the load profile before the training of the Markov model. The effect can be re-applied afterwards on the SLP.

- By design, the model is also not considering the differences between different days of the week. This issue can be easily fixed by feeding the model with longer load profiles. A greater abundance of data unlocks the possibility to train more transition matrices.

- The model is not suitable for the reproduction of patterns at low load levels, but the reproduction of these low-magnitude patterns is not of interest for distribution grid operation and planning.

- The model has medium requirements of memory. A data structure to model one RLP occupies around 100 kBytes of memory. If thousands of RLPs has to be modeled, this results in a memory requirement of a few GBytes. These levels of computational requirement are of a concern only if the model is run on a single machine. The issue can easily be solved by cloud computing.

- The number of states of the weekly and intra-day model has to be decided beforehand. A better performance could be achieved by designing an adaptive algorithm that can tailor the number of states of the model on the features of the input RLP.

Chapter 8

Conclusion

This work aimed at generating Synthesized Load Profiles for buildings not equipped with smart meters, to be exploited for simulations and operation of distribution grids. The project is based on dataset regarding the City of Basel (Switzerland). The methodology envisioned to accomplish the task by exploiting machine learning techniques, such as clustering and classification, to learn correlations between features of buildings and load profiles.

Nevertheless, because of a missing dataset, the focus had to be shifted towards the only phase of the machine learning pipeline that could be run in standalone mode: the generation of Synthesized Load Profiles that resemble the features of an input Real Load Profile.

In this thesis, in order to avoid ambiguity in the definition of features of load profiles, a novel notation has been developed.

Generation of SLPs has been carried out with a model based on Markov chains that demonstrated to be very effective in reproducing the most significant features of the original Real Load Profile utilized to train the model: annual consumption, distribution of magnitude of load, distribution of magnitude of peak load, distribution of the Time of Use of the peak load and autocorrelation.

The model is easily scalable and parallelizable, allowing the generation of large datasets of profiles that remove issues related to privacy and allow for large scale Monte Carlo Simulations.

The main drawback of the developed Markov model is that it has not been designed to take into account the effect of exogenous variables. Although this results in minor drawbacks for the Basel case, because of low penetration of electric heating and air conditioning in Switzerland, it is an aspect to be considered for the extension of the model to other countries.

8.1 Further research

Concerning the Markov model, further improvements should focus on the modeling of seasonal patterns and the influence of the temperature. The advantages of an adaptive algorithm able to optimally determine the number of states for each RLP should be explored.

Further research could explore the option to exploit information on the distribution of the duration of a load of a given magnitude. This would allow to differently model the temporal persistence of a state in the intra-day model, focusing on the reproduction of long-lasting patterns.

The test and validation of the complete machine learning algorithm relies entirely on the availability of the dataset linking buildings and load profiles. With the assumption of filling this gap, effort should be mainly put into the engineering of significant features and the development of robust validation methods. The specific algorithms for classification and clustering has been judged to play a secondary role and are suggested to be kept as simple as possible.

The dataset of building features could be expanded to include also non-residential buildings and, possibly, economic and sociological information on the occupants, i.e. age, profession, average income in the neighborhood, average price per m^2.

The dataset of load profiles should be expanded to account for multiple years of consumption. This would allow to better represent the RLPs and will provide information on individual yearly variations in consumption. More data would also allow to individually model each day of the week. Further expansion could include industrial loads, PV production and external temperature. Concerning PV production, the Markov model could be coupled with a thermal model of the PV panel and a generator of clear sky radiation to learn how to reproduce patters of cloud shadowing and, therefore, PV production.

Bibliography

[1] J Dickert and P Schegner. Residential load models for network planning purposes. *Modern Electric Power Systems (MEPS), 2010 Proceedings of the International Symposium*, pages 1–6, 2010.

[2] M. Newborough and P. Augood. Demand-side management opportunities for the UK domestic sector. *IEE Proceedings - Generation, Transmission and Distribution*, 146(3):283, 1999.

[3] Stephen Haben, Colin Singleton, and Peter Grindrod. Analysis and clustering of residential customers energy behavioral demand using smart meter data. *IEEE Transactions on Smart Grid*, 7(1):136–144, 2016.

[4] Wenyuan Li et al. *Reliability assessment of electric power systems using Monte Carlo methods*. Springer Science & Business Media, 2013.

[5] Ricardo Torquato, Student Member, Qingxin Shi, Student Member, and Wilsun Xu. A Monte Carlo Simulation Platform for Studying Low Voltage Residential Networks. *Smart Grid, IEEE Transactions on (Volume:5 , Issue: 6)*, 5(6):2766–2776, 2014.

[6] Hassan Farhangi. The path of the smart grid. *IEEE Power and Energy Magazine*, 8(1):18–28, 2010.

[7] Chris Beard. *Smart Grids for Dummies*. 2010.

[8] IEA. Smart Grids in Distribution Networks. Technical report, 2015.

[9] Wenyuan Billington, Roy; Li. *Reliability Assessment of Electric Power Systems Using Monte Carlo Methods*. 1994.

[10] Dirk P. Kroese, Tim Brereton, Thomas Taimre, and Zdravko I. Botev. Why the Monte Carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6):386–392, 2014.

[11] Gabriel a Terejanu. Tutorial on Monte Carlo Techniques. pages 1–15, 2013.

[12] Young Il Kim, Jong Min Ko, and Seung Hwan Choi. Methods for generating TLPs (Typical Load Profiles) for smart grid-based energy programs. *IEEE SSCI 2011 - Symposium Series on Computational Intelligence - CIASG 2011: 2011 IEEE Symposium on Computational Intelligence Applications in Smart Grid*, pages 49–54, 2011.

[13] Fintan McLoughlin, Aidan Duffy, and Michael Conlon. Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study. *Energy and Buildings*, 48(July 2009):240–248, 2012.

[14] Kirsten Gram-Hanssen. Standby consumption in households analyzed with a practice theory approach. *Journal of Industrial Ecology*, 14(1):150–165, 2010.

[15] Kirsten Gram-Hanssen. Residential heat comfort practices: understanding users. 2010.

[16] Loren Lutzenhiser and Sylvia Bender. The Average American Unmasked: Social Structure and Differences in Household Energy Use and Carbon Emissions Problem and Research Strategy. 2008.

[17] Kirsten Gram-Hanssen, Casper Kofod, and Kirstine N Petersen. Different Everyday Lives - Different Patterns of Electricity Use. *Proceedings of the 2004 American Council for an Energy Efficient Economy Summerstudy in Buildings.*, page 13, 2004.

[18] Kirsten Gram-hanssen and Danish Building. Technology and Culture As Explanations for Variations in Energy Consumption Social Construction of Technology. *Proceedings of the 2002 American Council for an Energy Efficient Economy Summer Study in Buildings*, pages 79–90, 2002.

[19] Janine Morley and Mike Hazas. The Significance of Difference: Understanding Variation in Household Energy Consumption. *eceee 2013 Summer Study.Rethink, renew, restart*, pages 2037–2046, 2011.

[20] Fintan McLoughlin, Aidan Duffy, and Michael Conlon. The Generation of Domestic Electricity Load Profiles through Markov Chain Modelling. *Euro-Asian Journal of Sustainable Energy Development Policy*, 3, 2010.

[21] G Wood and M Newborough. Dynamic energy-consumption indicators for domestic appliances: Environment, behaviour and design. *Energy and Buildings*, 35(8):821–841, 2003.

[22] Omid Ardakanian, Negar Koochakzadeh, Rayman Preet Singh, Lukasz Golab, and S. Keshav. Computing electricity consumption profiles from household smart meter data. *CEUR Workshop Proceedings*, 1133(c):140–147, 2014.

[23] Christof Bucher and Göran Andersson. Generation of Domestic Load Profiles - an Adaptive Top-Down Approach. *12th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, pages 436–441, 2012.

[24] W Labeeuw and G Deconinck. Residential Electrical Load Model Based on Mixture Model Clustering and Markov Models. *IEEE Transactions on Industrial Informatics*, 9(3):1561–1569, 2013.

[25] José Antonio Jardini, Carlos M V Tahan, M R Gouvea, A.U. Ahn, and F M Figueiredo. Daily Load Profiles for Industrial Low Voltage Consumers. *IEEE Transactions on Power Delivery*, 15(1):375–380, 2000.

[26] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78, 2012.

[27] Fintan McLoughlin, Aidan Duffy, and Michael Conlon. Evaluation of time series techniques to characterise domestic electricity demand. *Energy*, 50(1):120–130, 2013.

[28] Ethem Alpaydin. *Introduction to Machine Learning, Third Edition*. The MIT Press, third edit edition, 2014.

[29] Andrew Ng. Machine Learning. In *Coursera, online course*. Stanford University, 2012.

[30] Thierry Zufferey. SmartMetering Data Analysis by Machine Learning Techniques. Master's thesis, ETH Zürich, Power System Laboratory, id: PSL 1512, 2015.

[31] Dimitri P. Bertsekas Tsitsiklis and John N. Introduction to probability. *British journal of cancer*, 26(4):239, 1997.

[32] Kevin Brokish and James Kirtley. Pitfalls of Modeling Wind Power Using Markov Chains. *Electrical Engineering*, pages 1–6, 2009.

[33] T Pesch, S Schröders, H J Allelein, and J F Hake. A new Markov-chain-related statistical approach for modelling synthetic wind power time series. *New Journal of Physics*, 17(5):055001, 2015.

[34] Zhe Song, Xiulin Geng, Andrew Kusiak, and Chang Xu. Mining Markov chain transition matrix from wind speed time series data. *Expert Systems with Applications*, 38(8):10229–10239, 2011.