Simran Jit Kaur Sandhu and Marthe Vågen

Baseline Estimation for Flexibility Validation

Master’s thesis in Energy and the Environment
Supervisor: Jayaprakash Rajasekharan
June 2022

Master’s thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Electric Power Engineering

NTNU
Norwegian University of
Science and Technology

Simran Jit Kaur Sandhu and Marthe Vågen

Baseline Estimation for Flexibility Validation

Master's thesis in Energy and the Environment
Supervisor: Jayaprakash Rajasekharan
June 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Electric Power Engineering

NTNU
Norwegian University of
Science and Technology

Preface

The research presented in this master's thesis was conducted by Simran Jit Kaur Sandhu and Marthe Vågen the spring of 2022 for the Department of Electric Power Engineering. We are in our last semester of the two-year MSc programme Energy and the Environment at Norwegian University of Science and Technology.

This thesis includes an introduction to the flexibility market, with its participants and action sequence in addition to system architecture. Further, theory and literature review regarding baseline estimation will be presented. This includes an introduction to low voltage load forecasting, load forecasting methods for baseline estimation and challenges regarding baseline estimation. Data preparation and analysis will be presented in the data processing. Further, methodology regarding two different load forecasting methods for baseline estimation with different strategies will be presented, followed by results from the methodology and discussion. The workload was equally divided between the two students of the group. This thesis has contributed to an increased understanding of important aspects regarding flexibility validation in the settlement process at low voltage levels.

We would like to thank our supervisor, associate professor Jayaprakash Rajasekharan, for assistance and feedback during the project. In addition, we would like to thank PhD student Surya Venkatesh Pandiyan for valuable contributions regarding the simulation experiments and guidance throughout the thesis. Finally, we would like to thank each other for a great collaboration throughout this year.

Trondheim, June 9, 2022

Simran Sandhu

Simran Jit Kaur Sandhu

Marthe Vågen

Marthe Vågen

Abstract

To deal with challenges posed by intermittent energy resources and the implementation of distributed energy resources, distribution system operators require flexible loads and energy sources that support the balancing of electrical energy supply and demand. An aggregator, acting as an intermediary, may purchase flexibility from consumers to aggregate and sell to a buyer, such as distribution system operators. The settlement process between aggregator and distribution system operator requires validation of the activated flexibility, which can be challenging as this activated flexibility can not be physically measured.

The main research question of this thesis includes how the distribution system operator can validate this demand-side flexibility at substation level activated by the aggregator in the settlement process and how validation can be implemented in a realistic scenario using consumption data available to the distribution system operator. Load forecasting methods for baseline estimation can be implemented for this purpose, as they estimate what consumers would have consumed in the absence of flexibility activation. Two regression methods were proposed in this thesis: artificial neural network and multiple linear regression.

Two strategies were implemented with the regression methods: recursive and rectifying. The recursive strategy was chosen to improve the estimation results and make the simulations reflect a real-world scenario, as only data available to the distribution system operator was used. The rectifying strategy was implemented to improve the accuracy of the recursive strategy. Artificially created substation data with both 1 and 5-minute frequency were used. Baseline estimation of individual households was also conducted to examine whether more information regarding the flexibility validation could be determined at a lower level.

The implementation of the recursive strategy showed more accurate results in artificial neural network than multiple linear regression. Both methods followed the trend of the actual baseline, but neither method was able to capture the high fluctuating frequency. The rectifying strategy improved the baseline estimation results to some degree. The accuracy of the methods after implementing the strategies is moderate. However, it has scope for improvement in the future by using appropriate explanatory variables and advanced machine learning algorithms, among other factors.

The frequency adjustment had little or some effect on the accuracy of the methods, and neither of the two frequencies might therefore be more favorable for the DSO. As baseline estimation is challenging on residential data, the contribution from the individual house estimations might be negligible to the DSO in the settlement process.

The complexity of validation using baseline estimation has been proven, where even the most common regression methods fail due to the nature of the problem. Further work should research the factors affecting the strategies and examine different strategies.

Sammendrag

For å håndtere utfordringer fra ukontrollerbare energiressurser og implementering av distribuerte energiressurser, krever distribusjonssystemoperatører fleksible laster og energikilder som hjelper balanseringen av elektrisk energiforsyning og etterspørsel. En aggregator, som fungerer som en mellommann, kan kjøpe fleksibilitet fra forbrukere for å så samle dette og selge videre til en kjøper, som distribusjonssystemoperatører. Oppgjørsprosessen mellom aggregator og distribusjonssystemoperatør krever validering av den aktiverte fleksibiliteten, noe som kan være utfordrende siden den aktiverte fleksibiliteten ikke fysisk kan måles.

Hovedformålet i denne masteroppgaven inkluderer hvordan distribusjonssystemoperatøren kan validere denne etterspørselsside-fleksibiliteten på nettstasjonsnivå aktivert av aggregatoren i oppgjørsprosessen, og hvordan valideringen kan implementeres i et realistisk scenario ved å bruke forbruksdata tilgjengelig for distribusjonssystemoperatøren. Lastprognosemetoder for grunnlinjeestimering kan brukes til disse formålene, da de estimerer hva forbrukere ville ha konsumert i fravær av fleksibilitetsaktivering. To regresjonsmetoder ble foreslått i denne oppgaven: kunstig nevrale nettverk og flerlineær regresjon.

To strategier ble implementert med regresjonsmetodene: rekursiv og korrigerende. Den rekursive strategien ble valgt for å forbedre estimeringsresultatene og få simuleringene til å reflektere et virkelighetsscenario, da kun data tilgjengelig for distribusjonssystemoperatøren ble brukt. Den korrigerende strategien ble implementert for å forbedre nøyaktigheten til den rekursive strategien. Det ble brukt kunstig nettstasjonsdata med både 1 og 5 minutters frekvens. Grunnlinjeestimering av individuelle husholdninger ble også utført for å undersøke om mer informasjon angående fleksibilitetsvalideringen kunne bestemmes på et lavere nivå.

Implementeringen av den rekursive strategien viste mer nøyaktige resultater i kunstig nevrale nettverk enn flerlineær regresjon. Begge metodene fulgte trenden til den faktiske grunnlinjen, men ingen av metodene var i stand til å fange opp den høye fluktuerende frekvensen. Den korrigerende strategien forbedret estimeringsresultatene til en viss grad. Nøyaktigheten av metodene etter implementering av strategiene er moderat. Det er imidlertid rom for forbedringer i fremtiden ved å blant annet bruke passende forklaringsvariabler og avanserte maskinlæringsalgoritmer.

Frekvensjusteringen hadde lite eller noe betydning for nøyaktigheten til metodene, og ingen av de to frekvensene vil derfor være mer gunstige for distribusjonssystemoperatøren. Siden grunnlinjeestimering er utfordrende på boligdata, kan bidraget fra de enkelte husestimatene være ubetydelig for distribusjonssystemoperatøren i oppgjørsprosessen.

Kompleksiteten til validering ved bruk av grunnlinjeestimering er bevist, der selv de vanligste regresjonsmetodene mislykkes på grunn av problemets natur. Videre arbeid bør undersøke faktorene som påvirker strategiene og undersøke ulike strategier.

Table of contents

List of abbreviations

AE	Absolute error
AI	Artificial intelligence
AMS	Advanced metering system
ANN	Artificial neural network
AR	Autoregressive
ARE	Average relative error
ARIMA	Autoregressive integrated moving average
ARIMAX	Autoregressive integrated moving average exogenous
ARX	Autoregressive exogenous
BRP	Balance responsible party
CART	Classification and regression trees
CNN	Convolutional neural networks
DER	Distributed energy resources
DR	Demand response
DSO	Distribution system operator
EV	Electric vehicle
GDPR	General Data Protection Regulation
GP	Gaussian process
ICT	Information and communication technologies
kNN	K-nearest neighbor
LFM	Local flexibility market
LSTM	Long short-term memory
MA	Moving averages
MAE	Mean absolute error
MAPE	Mean absolute percentage error
ME	Mean error
MLP	Multilayer perceptron

MLR	Multiple linear regression
MSE	Mean square error
NARX	Nonlinear autoregressive exogenous
NRER	Normalized relative error ratio
NRMSE	Normalized root mean square error
OLS	Ordinary least squares
OPI	Overall performance index
P2P	Peer-to-peer
PL	Pinball loss
PICP	Prediction interval coverage probability
$\mathbf{R^2}$	Coefficient of determination
ReLu	Rectified linear unit
RMSE	Root mean square error
RMSD	Root mean square deviation
RNN	Recurrent neural network
RT	Regression trees
SAE	Stacked autoencoder
SSE	Sum of squared errors
SVR	Support vector machine
SVM	Support vector regression
tanh	Tangens hyperbolicus
TIC	Theil's inequality coefficient
TSO	Transmission system operator
WS	Winkler score

1 Introduction

The stability of the power system is being challenged with the increase of renewable energy. Flexibility can be used as a tool to maintain the stability of the grid. The purpose of this thesis is to research how the flexibility buyer can validate the demand-side flexibility activated by the seller of flexibility in the settlement process. In this chapter, a description of the background and motivation for this thesis is elaborated. The problem definition is described, followed by the contribution of the thesis. The thesis outline is presented towards the end.

1.1 Background

The Paris Agreement entered into force in November 2016, becoming the first legally binding international treaty on climate change in the world. The agreement aims to limit the global temperature rise to 1.5°C [1]. Norway is by 2030 expected to reduce its greenhouse gas emission by at least 50 %, compared to the reference year 1990. This is a direct effect of the Paris Agreement [2].

The world is, due to several global agreements and initiatives, rapidly moving towards a low carbon future, and the share of renewable energy generation is increasing in order to meet the growing power demand. To reduce carbon emissions, distributed energy resources (DER) such as photovoltaics, electrical vehicles (EVs) and demand response (DR) are being introduced. DER are defined as smaller resources directly connected to the distribution network [3]. The DER will have a great impact on the distribution network, as the distribution network will, for the first time, experience reversed power flow caused by distributed generation. This will further challenge and exacerbate grid stability issues in the existing distribution network, which is mainly designed for single state loads [4].

On the contrary, if appropriately managed, DER could provide ancillary services to system operators as well as local system services through price-based incentives. This could potentially solve issues related to voltage regulation, power quality and distribution network congestion [3].

To deal with challenges posed by intermittent energy sources and the implementation of DER, system operators require flexible loads and energy sources that support the balancing of supply and demand of electrical energy. Heating and cooling loads, batteries and EVs are among some potential flexibility resources, which are able to modify their consumption, e.g. curtail, regulate or shift, to provide a variety of services to the grid [5].

In Norway, there is a strict distinction between the grid operator, distributed system operator (DSO), on one hand and the electricity generator and consumer on the other hand. The DSO is hence dependent on a third party or intermediary, such as an aggregator, to provide DR. An aggregator can purchase flexibility from prosumers or consumers in order to aggregate and sell to a buyer, such as the DSO. One challenge in this context is that the buyer needs to verify and quantify the activated flexibility in the settlement process, before settling with the aggregator [6].

Settlement and validation of activated flexibility are however more complex. It is difficult to identify the exact amount of flexibility activated as a result of the flexibility trade. This can be obtained by comparing load measurements, such as advanced metering system (AMS) or substation measurements, with a so-called baseline. Baseline estimation is an important tool that can be applied for such services [5, 6].

1.2 Motivation

As the activated flexibility can not be physically measured, a DSO can estimate a baseline and compare it with the load measurements during a DR event. A baseline is an estimation of what prosumers or consumers would have consumed in the absence of flexibility activation. It is challenging to predict scenarios in the absence of activation. Hence, baseline estimation becomes an important aspect of the flexibility provision for the DSO, as the amount of activated flexibility is determined by the difference between the estimated baseline and actual consumption [7].

As the baseline estimation provides the basis for the economic compensation calculation, it is of importance for both the aggregator and DSO to obtain an accurate and agreeable estimation in order to verify and quantify activated flexibility. An incorrect baseline estimation can lead to over- or under-compensation of the activated flexibility for either parties and further affect the interests of all DR participants involved [7].

To estimate baselines, various load forecasting methods can be implemented. Load forecasting methods predict load demand and can thereby be used to estimate consumption in the absence of a DR event. This is a well-studied research topic, and there exists a vast amount of different load forecasting methods for baseline estimation [8].

However, baseline estimation can be challenging, in particular when considering the consumption data available for the DSO in a real-world scenario. Developing a method that can accurately estimate a baseline to verify activated flexibility, given the data a DSO will have available during a DR event, is a key challenge in the implementation of DR programs [8].

Traditionally this baseline problem has mainly focused on large-scale, predictable and controllable assets. However, with the increased flexibility at lower voltage networks, the DSO needs to adjust their baseline methods to fit more small-scaled and less predictable energy resources. Hence, it is of relevance to examine baseline estimation on a substation level [9].

1.3 Problem definition

For this master's thesis, the main research question is:

How can the DSO validate demand-side flexibility at substation level activated by the aggregator in the settlement process? How can validation be implemented in a realistic scenario using consumption data available to the DSO?

In order to achieve this, the thesis aims to:

- Estimate baseline for an artificially created substation utilizing different load forecasting methods and evaluate them using performance metrics
- Adapt suitable forecasting strategies fitted for real-world implementation
- Examine different factors that may increase accuracy, such as frequency and consumption data on different levels

1.4 Contribution

The contributions of this thesis can be listed as follows:

- Two regression methods for baseline estimation were built and tested on an artificially created substation from Austin, Texas, with 1-minute frequency. The methods were evaluated using three different performance metrics.
- Recursive strategy was evaluated to create more realistic methods. To improve the estimations from this strategy, a rectifying strategy was implemented and evaluated.
- Consumption data with 5-minute frequency was tested for both methods, which had a small effect on the accuracy of the methods. The methods were also tested on individual households. Due to the challenges regarding baseline estimation of residential data, the contribution from the results may be negligible to the DSO in the settlement process.
- The complexity of validation using baseline estimation has been established, where even the most common algorithms fail due to the nature of the problem. Suggestions for further work have been provided.

1.5 Thesis outline

The thesis is structured as follows:

An overview of the flexibility market, the respective participants and the action sequence is introduced in chapter 2, to get an understanding of the current situation in the market. A literature review on previous papers regarding load forecasting for baseline estimation is further on conducted in chapter 3 to gain perspective on the existing research relevant to the thesis. Theory regarding state-of-the-art low voltage load forecasting is also included in this chapter.

Chapter 4 presents the methodology regarding data preparation, in addition to an analysis of the artificially created substation data. Baseline estimation methodology for two load forecasting regression methods with different scenarios and strategies is given in chapter 5, followed by the baseline estimation results for the strategies in chapter 6. The results are discussed in chapter 7. Finally, a conclusion of the thesis is presented in chapter 8, in addition to suggestions for further work.

2 Flexibility

The increasing integration of renewable generation and DER is introducing new operational challenges to the grid. Due to the rise in intermittency, price volatility and uncertainty, the DSOs are seeking new concepts and market tools to enable active system management and control using flexibility. With the introduction of local flexibility markets (LFM), equipped with nonrigid technology, it becomes possible to aggregate small generation resources allowing performance at an adequate level of flexibility. The LFM provides an economically efficient way to trade flexibility between market participants in addition to providing ancillary services to the grid. These participants may include the DSO, transmission system operator (TSO), aggregator, balance responsible party (BRP), flexibility market operator, prosumers and consumers [10, 11].

In this chapter, the participants in the LFM are presented. The action sequence and system architecture of the flexibility market are further introduced. The identification of this relevant background material was carried out in the specialization project preceding this thesis, given in [12]. In addition, the state-of-art and current projects regarding the flexibility market are reviewed.

2.1 Transformation of the traditional power system

Traditional power systems are undergoing significant changes, in particular driven by the growing concerns over energy depletion. Technologies such as DR, renewable generation and information and communication technologies (ICT) are hence drawing attention and transforming the traditional power system. This transformation takes place on the generation-, grid- and demand-side of the power system. The key drivers for flexibility in the distribution systems are intermittent and uncontrollable generation, controllable and flexible consumers, intelligent and smart systems, bi-directional and decentralized systems, and multi-energy systems [10].

The primary source of flexibility for system operators has traditionally been power generators from larger industries. However, potential flexibility available to address local problems consists of small-scale demand-side resources. To enter such a market, new developments consisting of aggregation and automated solutions are required [13].

To further maximize the efficiency of the market and utilize all potential resources within it, regulatory barriers and rules set by marketplaces need to be challenged as the technology develops. Minimum bidding is such a challenge where one of the balancing market products of the Norwegian TSO, Statnett, has its current default minimum bidding size to 10 MW. However, recent studies propose that granularity of 1 MW can be feasible [14].

Regulation regarding independent aggregators should also be a focal point going forward. The aggregator will play a more prominent role in the future flexibility market, as the mobilization of small demand-side resources to a large extent will occur through them. As a consequence, it will be of high interest to implement clear regulations for independent aggregators [13].

2.2 Participants in the flexibility market

The flexibility market is a competitive trading platform between regulated buyers such as TSOs and DSOs, non-regulated buyers such as BRP and flexibility sellers being aggregators. This is accomplished under the supervision of an independent market operator to guarantee neutrality [15]. The DSOs operate on a local or regional market with low, medium and occasionally high voltage networks. They require flexibility for voltage control, congestion management, loss reduction and redundancy at lower voltage levels in a cost-efficient manner [13].

The BRP, on the other hand, buys flexibility for portfolio optimisation and is responsible for the imbalances of the power exchange market. It is, however, unclear how the BRP will be impacted and regulated in regard to flexibility trading [13]. The consumers and prosumers are the flexibility providers and are managed by the aggregator, who sells the flexibility for profits. The flexibility of these participants is reflected through bids and offers on the day-ahead, intraday or balancing market [10].

The aggregator is a link between the flexibility supplier and buyer, as seen in Figure 2.1. They are in charge of acquiring flexibility from consumers and prosumers, aggregating them into a portfolio, creating flexibility services and offering these to distinct markets, serving different market players. The aggregator can provide both up- and down-regulation of flexibility, depending on the bid of the marked actors. Up-regulation is provided by reducing the end-user consumption, while down-regulation entails an increase in consumption [16]. The aggregator receives value of the up- or down-regulated flexibility in return and shares this with the end-users as an incentive for shifting their loads [17]. Their role with respect to the traditional BRP and other retailers is yet to be defined as the market keeps on emerging [13].

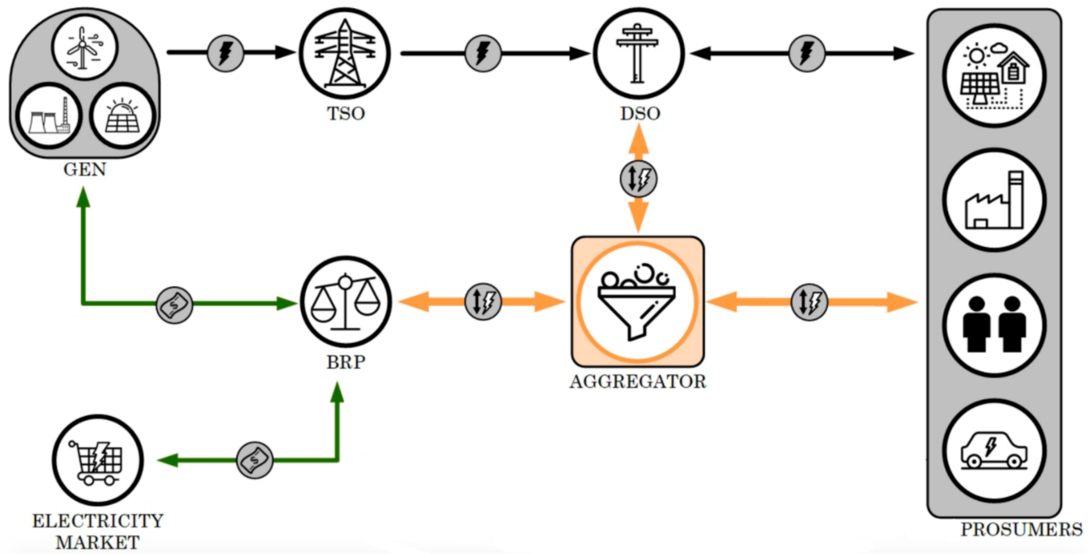


Figure 2.1: *Flexibility market overview showing the relations between the different participants, such as DSO, TSO, BRP, aggregator, consumers and prosumers [16].*

2.3 Action sequence of the flexibility market

The timeline and action sequence of the flexibility market is given by three processes: contracting and bidding, activation, and settlement and validation. During the contracting and bidding process, the participants of the flexibility market communicate in order to reach an agreement on the price and quantity of flexibility. The DSO provides a future analysis to investigate and predict future congestion or voltage violation problems. If the risk of a potential violation is detected, the DSO sends a flexibility request to the market operator. The BRP receives the portfolio forecast and estimates future imbalances simultaneously. If desired, they will also send a flexibility request [10].

In the activation process, the flexibility buyers activate their acquired flexibility by sending a flexibility activation request to the market operator. The aggregator provides activated flexibility demand by scheduling and controlling the supplier. An activation confirmation is sent back to the market operator after the aggregator receives an activation signal. The market operator further forwards the flexibility provision confirmation to the DSO and BRP, respectively [10].

The settlement agreement of a long term availability contract, the last step in the process, can easily be settled and paid in accordance with the terms and conditions among the participants. Settlement and validation for activation of the flexibility resources are however more complex. It is difficult to identify the exact amount of flexibility activated as a result of the flexibility trade, creating a baseline estimation problem. This problem refers to how challenging it is to predict scenarios in the absence of flexibility activation. However, flexibility trading is still based on volume, where the market actors bid on up- or down-regulation depending on such a baseline [13].

2.4 System architecture

There are several system architecture models for LFM. A four-layer flexibility market architecture is presented in [10]. This kind of system architecture is based on other models such as the smart grid architecture model in [18] and peer-to-peer (P2P) energy trading in [19], which both provide various outlooks on the system layering in smart grids. The four-layer interoperable flexibility market architecture is illustrated in Figure 2.2 and presents the key elements and technologies of flexibility markets.

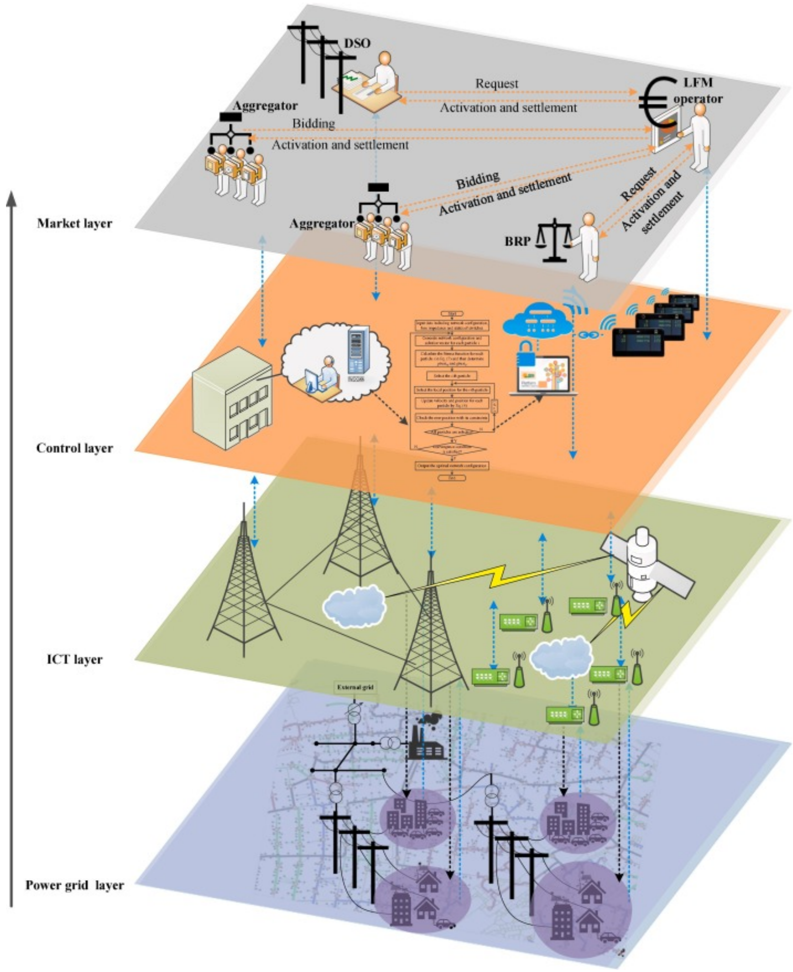


Figure 2.2: *Flexibility market architecture model with four layers top to down: market, control, ICT and power grid layers [10].*

The bottom layer is the power grid layer. This is composed of all hardware technologies and other physical components that can be installed to perform different electrical distribution system services. This includes flexibility resources and components such as transformers, smart meters, energy storage units, loads, DER, consumers, prosumers and aggregators, all providing supply-, grid- or demand-side flexibility. The power grid layer is where flexibility trading is conducted [10, 20].

The ICT layer supports flexibility trading with communication devices, protocols, applications and information flow, in order to monitor, control and manage systems [10]. Communication devices include sensors, wired or wireless communication connections, routers, switches, servers and various types of computers. The protocols include Transmission Control Protocol, Internet Protocol and Point-to-Point Protocol. Together with software components, this allows for fast bidirectional communication over the internet. Information transfer and file exchange are elements found in communication applications. The information flow refers to the content of each message transferred between the sender and receiver [19, 20].

The control functions of flexible resources are managed in the control layer. Quality and reliability preservation of power supply and power flow control are defined through several control strategies in this layer. Such control functions may include voltage, frequency and active power control. All the technical work such as baseline estimation and load forecasting is performed at the control layer with objectives, constraints and regulations of flexibility dictated by the ICT and market layer [10, 19].

The market layer regulates the flexibility trading between the four participants: DSO, BRP, aggregator and LFM operator. It is responsible for the management of flexibility transactions in the market. This layer accommodates use case business processes, services and organizations. Objectives, regulatory and economic constraints are included to facilitate various business models and use cases [10, 19].

2.5 Implementation of flexibility

DR services and LFM are set to be vital in bringing society to the core of the energy transition. DR is a resource defined as the temporary change in electricity usage patterns by the end-user during high wholesale market prices or when system reliability is at risk. It is a promising service offering flexibility to the market through supply and demand balancing in power systems, with a high share of renewable generation [21].

Several EU projects have been carried out to test and validate various flexibility concepts. InterFlex (2017-2019) is one of the EU projects, exploring the benefits of flexibility for the local distribution grid. LFM was one of the topics researched in the project, where flexibility mechanisms were established and the roles of stakeholders were defined. Further, the project included development of ICT tools which consisted of forecasting engines, market platforms and aggregator interfaces. The LFM had formats to match the request from the DSO with the offers of the aggregators, while identifying suitable time frames for the activation of flexibility [22].

The General Data Protection Regulation (GDPR) rules were mentioned as one of the challenges in the DR solutions. The project revealed that access to data was an essential input to forecasting and flexibility procurement. However, the complexity of the DR agreement forms were too high for residential customers and should be simplified without lowering the level of privacy, according to the conclusion of the project [22].

FLEXCoop (2017-2021) is another EU project. The main project assets included an introduction of a complete automated DR framework and tool suite for residential electricity consumers, which further would enable energy cooperatives to explore DR business models and take the aggregator role. The solution enabled aggregators to accurately forecast flexibility demand and DR potential of prosumers. Based on a properly adapted performance measurement and verification methodology, objective DR settlement and prosumer remuneration were ensured. The solution also included a marketplace were prosumers could choose among more than one aggregator [23].

eDream (2018-2021) is also an EU research program, aiming to develop innovative tools for DR optimal programs, including DR forecast, profiling, segmentation and load forecasting. The project investigated the development of a new blockchain application for a decentralized P2P network that directly connected consumers and sellers. The project also looked into a loop near real-time DR verification and secure data handling. Moreover, it provided flexibility on a medium or low voltage grid level, supporting the integration of renewable sources into the existing power system, rather than having to invest in new grid infrastructure. By following the proposed technologies in this program, the emission savings were estimated to be 10 %, and cost-saving and energy reduction for customers would be 25 % [24].

DRIMPAC (2017-2022) is an ongoing EU research program. According to the program, there are two main roadblocks preventing direct participation of residential buildings in LFMs: technology-related roadblocks and consumer-related roadblocks. Technology-related roadblocks are related to the communication of DR-signals still being a significant barrier in the rise of DR programs. Consumer-related roadblocks entail DR programs being intrusive to consumer privacy and can bear the risk of increased energy bills or other inconveniences. The project hence proposes a universal technological framework, a user-friendly DR program where the consumer can take ownership of their energy transition. This will include the necessary end-to-end communication for procurement and delivery of flexibility and competitive services and dynamic tariff schemes, representing both the consumer and energy systems interests faithfully [25].

The integration of a successful flexibility market provided by DR will rely on the redesign and adaption of current regulatory framework through subsided programs such as the EU programs. Nevertheless, challenges concerning baseline estimation, flexibility estimation and attaining data integrity while respecting consumer privacy are still to be resolved with the integration of ICT [26].

To provide DR in the residential sector, the DSO can obtain flexibility from end-users connected to their distribution network. Following a flexibility activation, the DR event needs to be settled. This can be accomplished by comparing the metered load to a baseline. Baseline estimation is an important tool for the DSO in order to validate and quantify the service delivered by an aggregator [26, 27].

2.6 Summary

- The flexibility market is a trading platform between buyers such as TSOs, DSOs and BRP, and flexibility sellers being aggregators.
- LFM architecture consists of four layers: market, control, ICT and power grid.
- The DSO is seeking new concepts and market tools to enable active system management and control using flexibility.
- Flexibility is activated in the activation process, followed by a settlement and validation process between the DSO and aggregator.
- Baseline estimation is an important tool for the DSO in order to validate and quantify the service delivered by an aggregator.

3 Baseline estimation

The procurement of flexibility aims to facilitate the DSO with a market-based instrument to mitigate congestion and overloading of the distribution grid. The DSO acquires this flexibility from market participants, such as aggregators and end-users. To participate in a DR system could be beneficial to the consumer, as they are offered financial incentives for reshaping their electricity consumption patterns. After the DR event has taken place and flexibility has been delivered, a settlement process is required. As the flexibility activation can not be measured physically, the buyer and seller of flexibility must agree on a settlement process. One way to solve this problem is to compare load measurements to an estimated baseline. The baseline represents the traditional load curve in the absence of a DR event, which needs to be estimated, creating a baseline estimation problem [13, 28].

Flexibility includes different types of consumption modifications. Curtailment is such a modification where demand reduction takes place during a DR event, as shown in Figure 3.1. The demand reduction requires evaluation, a challenge which is demanding to solve as it can not be directly measured. In order to quantify this curtailment, the baseline must be estimated, which is an estimate of what consumers would have consumed in the absence of a DR event. The activated flexibility would then be acquired by subtracting the metered load from the baseline. Developing a model that could accurately obtain the baseline, remunerate consumers and incorporate more volatile renewable resources is becoming a major issue for all DSOs [28].

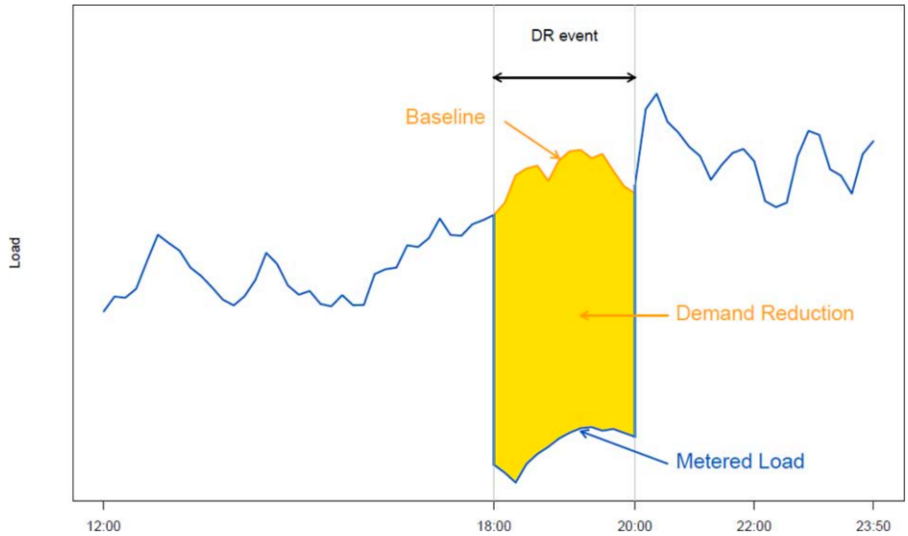


Figure 3.1: *A DR event in the form of demand reduction, where the blue line represents the metered load. During demand reduction, the metered load decreases. From the estimated baseline, it will be possible to approximate the demand reduction during this DR event [28].*

It is of importance for both the aggregator and DSO to obtain an accurate baseline estimation, as the difference between this and load measurements gives the basis for the economic compensation calculation. An incorrect baseline estimation can lead to over- or under-payment of the flexibility and further affect the interests of all DR participants [7]. If the end-user gets under-compensated for their flexibility, it could result in less participants willing to engage in the flexibility market. On the other hand, if end-users were to be over-compensated, the cost for the DSO would increase, reducing the potential impact and benefits of DR programs [29]. To achieve an accurate estimation, available measurements and historical baselines are required [6].

The load measurements compared to the baseline can be found on a household or substation level. In Norway, the aggregator will have available household level AMS data of DR participants, while the DSO will have data for whole substations. In addition, the DSO will have access to AMS data in 60-minute frequency within their grid area through Elhub. Elhub is a central information technology system which supports distribution and aggregation of metering values for consumption and production in Norway [30]. The AMS data at Elhub will, in the near future, change to 15-minute intervals in order to have better control of the inputs and outputs in the grid. Data access and handling at Elbhub is in accordance with GDPR [31].

In order to estimate baselines, load forecasting methods can be utilized. Load forecasting methods predict load demand and can thereby be used to estimate consumption in the absence of a DR event [8]. This chapter will further introduce low voltage load forecasting, including commonly used explanatory variables and performance metrics. In addition, different load forecasting methods and techniques will be presented. This part of the literature review was carried out in the project report, given in [12].

Additionally, a literature review on load forecasting for baseline estimation will be presented. Two of the methods presented in the literature review were of particular interest, and extensive theory regarding these two will therefore be presented. Lastly, different challenges regarding baseline estimation are included. As there is a lack of literature regarding baseline estimation on substation level, the focus will mainly be on baseline estimation on residential level or for aggregated residential consumption, which artificially creates substation data.

3.1 Low voltage load forecasting

Traditionally, load forecasting has been implemented at high voltage level. This level includes the aggregated demand of a considerable amount of consumers, typically hundreds of thousands [32]. Other than at smart meter level, there is a lack of analysis of low voltage load forecasting. Few papers are aimed at the secondary or primary substation level of the distribution network. Low voltage feeders, which carry energy to a substation, include on average about 50 households and are usually less volatile than smart meter data [8].

Low voltage networks include a wide variety of behaviours, and the number of consumers connected is one of the largest indicators of forecast accuracy. It, therefore, becomes more difficult to accurately forecast smaller substations as the relative error increases exponentially with the decrease of the substation. In addition to including the aggregation of individual households, low voltage networks consist of many different components such as streetlights, cameras and other street furniture [8].

As there is a lack of literature regarding forecasting at low voltage level, it is essential to highlight the differences between low voltage level demand and medium or high voltage levels. The volatility increases at low voltage level due to lower aggregation of demand. In addition, the variation of demand increases with various feeders, consisting of different numbers and types of consumers. Further, there is a limited understanding of the forecasting method inputs at this level. These differences will cause significant changes in the techniques and methods that are applied to forecasting low voltage demand in comparison with medium and high voltage level demand forecasting [8].

One of the published literature reviews on low voltage load forecasting is given in [8]. This paper aims to provide an overview of current methods and applications, in addition to presenting challenges and recommendations. The latter includes modelling uncertainty due to weather, as only a handful of the papers reviewed by the authors utilized weather predictions instead of weather actuals. By using weather actuals, the forecast errors are being under-reported.

In [33], a tutorial review on probabilistic load forecasting is provided, which includes notable techniques, methodologies, evaluation methods and common misunderstandings. The paper also includes a literature review of the subject, as well as future problems in the field of probabilistic load forecasting. Future challenges that the paper include are climate variability, EVs, wind and solar generation, energy efficiency and DR.

Short-term load forecasting at low voltage level is focused on in [32], in addition to the effect of temperature. Short-term is defined between one and four days ahead. The paper provides a detailed analysis of several state-of-the-art methods, including both point and probabilistic low voltage load forecasts. 100 real feeders are utilized to evaluate the accuracy of these methods, and the tested methods include kernel density estimation, simple seasonal linear regression, autoregressive (AR) and Holt-Winters-Taylor exponential smoothing method. The paper discovers a weakened correlation between demand and temperature, at least in the UK area where the feeders were located, with seasonality being a stronger driver for demand.

In [34], a review and outlook on energy forecasting are provided. The paper presents a review of influential energy forecasting papers, where research trends are summarized, and the importance of reproducible research is discussed. Machine learning and artificial intelligence (AI), forecast combinations, hierarchical forecasting and probabilistic forecasting are among the topics discussed. An outlook on the future of energy forecasting is also included.

Load forecasting can be classified into single-point prediction or multiple point prediction, as stated in [35], according to whether one or several points are being forecasted. Data used in forecasting can be minutely, hourly, daily or annually, based on the frequency. Single-point prediction will forecast one frequency step ahead, while multiple point prediction will forecast a given number of steps ahead. The paper mentions five strategies for multiple point prediction, including recursive and direct strategies.

Forecasting horizons can be very short-term, short-term, medium-term and long-term. Very short-term includes forecasting up to a few hours, while short-term is day-ahead and up to a few days. Medium-term includes up to weeks and months, and long-term is from a year and further [8].

This section will further include explanatory variables commonly used in low voltage load forecasting, followed by performance metrics. Additionally, different methods and techniques will be presented.

3.1.1 Explanatory variables

Explanatory variables are used as inputs in load forecasting methods, and the selection of appropriate inputs is of importance. Feature selection is the process of selecting a subset of the most relevant variables in the development of a forecasting model. This involves reducing the number of variables to improve the accuracy of the model, reduce the computational complexity and prevent overfitting [8, 36].

Overfitting can occur if the model consists of excessive variables and it becomes unnecessarily complicated. Therefore, irrelevant or redundant variables are discarded from the model in the feature selection process. Feature selection methods include filter, wrapper and embedded methods. Filter methods use measures such as correlation to score the variables. Wrapper methods are based on the results of testing subsets of variables in a preliminary model, and embedded methods search the space of the variables and the model parameters simultaneously [36].

Explanatory variables can be categorized into meteorological, seasonal and econometric, as shown in Table 3.1. Most papers include historical load data, in addition to one or several explanatory variables from these categories. Historical data is often included as measured consumption from one or several previous minutes or hours, depending on the data frequency, often called lagged consumption. Seasonal variables can be given as minute, hour, date, month and year, in addition to day of year or month, week of year and type of day. Type of day can include whether it is weekday, weekend or holiday [8].

Table 3.1: *Explanatory variables divided into the categories meteorological, seasonal and econometric [8].*

Meteorological	Seasonal	Econometric
Temperature	Minute	Number of consumers
Humidity	Hour	Social class
Solar irradiance	Date	Gender
Wind speed and direction	Month	Age group
Precipitation	Year	Data of neighboring substation
Pressure	Day of year	
	Day of month	
	Week of year	
	Type of day	

Meteorological variables, especially temperature, are the most common inputs in load forecasting methods, out of these categories. This variable can also be combined with other variables such as humidity, solar irradiance, wind speed, wind direction, percipitation and pressure. Although temperature is one of the most common explanatory variables, some papers have researched its effect in forecasting methods [8].

Several methods with and without temperature data were considered in [32], and the authors found no effect on the short-term forecast accuracy. On the other hand, the paper given in [37] established that temperature accounted for about half of the variation in the load when temperature and relative humidity were incorporated in day-ahead low voltage transformer level forecasts. However, this paper is from 2013, and the research within this field has made significant progress since then.

Both weather forecasts and actual weather data can be used in load forecasting methods. The authors in [8] state that predicted weather should be used when predicting load, though few papers reviewed by the authors utilized this as input. Most papers apply the actual weather observations, which can lead to the corresponding forecast accuracy being over-optimistic. According to the authors, papers considering low voltage level require more local weather inputs, though several papers utilize average country-based data instead.

Econometric variables are also common inputs in load forecasting methods. These variables can include the number of consumers, social class, gender, age group and data of neighboring substations and are mainly used for long-term forecast methods. Some of these variables require additional information about the consumers, which is not always available. Different papers have therefore conducted surveys to get information about social class, gender and age group [8].

3.1.2 Performance metrics

It is crucial to evaluate the performance of a model when forecasting loads. The performance of forecasting methods is assessed based on the error between the actual load and the predicted load. The authors in [38] state that the most common performance metric in load forecasting is root-mean-square error (RMSE), followed by mean absolute percentage error (MAPE). In addition, other performance metrics such as mean absolute error (MAE) and mean square error (MSE) were mentioned. Mean bias error is acknowledged as a standard performance metric in [39] and is used to describe whether a model is under- or over-predicted.

MSE is the average squared difference between predicted and actual load, while RMSE is the square root value of MSE, which represents the average difference between the predicted and actual values. RMSE can be calculated by equation 3.1, where N is the sample size, y is actual load and $\hat{y}$ is predicted load [40].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{3.1}$$

A low RMSE value suggests that the model is well-fitted for the data set. The acceptable range of RMSE is, however, dependent on the load profile used in the forecasting [40]. As stated in [39], RMSE represents the short-term performance of a model. A high RMSE score indicates a poorly performed feature selection process, according to [41].

MAPE is a statistical measure used to define the accuracy of a machine learning method on a particular data set. A disadvantage of this metric is that a few numbers close to zero can cause a large MAPE value, even though the absolute error is small. As residential consumption can be close to zero in periods, this performance metric may thereby not be suitable for this purpose [42]. However, MAE does not have this same disadvantage and can be calculated by equation 3.2. MAE is used to quantify the accuracy of load forecasting methods, where a lower value indicates a higher accuracy [43]. RMSE and MAE have the same unit as the dependent variable, which is power in this thesis [7].

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{3.2}$$

Another common performance metric is the coefficient of determination (R^2). This is a regression error metric used to evaluate the accuracy and efficiency of a model. R^2 values describe the performance of the model in percent, as well as the variation in the dependent variable after it has been predicted by explanatory variables. A higher value of R^2 represents a more suited model and better results. Equation 3.3 can be used to calculate R^2, where SS_{res} is the sum of squares of residual errors of the data model and SS_{tot} represents the total sum of errors [44].

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{3.3}$$

Both RMSE and R^2 indicate how well a regression model fits a data set, as stated in [40]. However, the RMSE value represents the typical distance between the predicted and actual value, while the R^2 value suggests how well the explanatory variables can explain the variation in the dependent variable. It can therefore be useful to calculate both metrics, as each gives valuable information about the performance of a model.

Other performance metrics include estimation error, which represents the absolute value of the difference between actual and forecasted load and is calculated using the mean of the relative RMSE [45]. Estimation bias is another metric, which plays an essential role in assessing the performance of a method. A positive bias value indicates that the method overestimated the actual consumption of the consumers, and vice-versa [43].

Another metric is variance, which describes the amount the prediction will change if trained with different data. A high variance indicates that the model will take into account the random noise in the data, rather than the actual dependent variable. Low variance indicates a robust model. Both low bias and variance are preferred, as the model will then be consistent and highly accurate. Achieving the right balance between these two metrics is often referred to as the bias-variance trade-off [46].

3.1.3 Methods and techniques

There are several different load forecasting methods and techniques, which can be divided into statistical and time-series approaches, machine learning and other artificial intelligence (AI) approaches, and probabilistic forecasting. Load forecasting methods can also be combined, in a hybrid approach, to increase the accuracy of the forecasts [8].

Statistical and time-series approaches are commonly used methods for short-term load forecasting [47–53]. The methods usually have linear parameters, including linear regression and traditional time-series approaches. In recent years, nonlinear approaches, such as nonlinear regression, have become more popular as well [8].

Linear regression estimates the relation between one or several explanatory variables and the dependent variable. These methods are easy to interpret and solve and are often coupled with meteorological explanatory variables. Simple linear regression predicts the dependent variable using only one explanatory variable. If one explanatory variable is insufficient to produce accurate forecasts, multiple linear regression (MLR) can be implemented instead. MLR assumes the relation between explanatory and dependent variables to be sufficiently modelled by utilizing a linear modelling approach. The parameters in this approach are usually estimated based on the minimization of the residual sum of squared errors [8, 54].

Time-series methods include AR, moving averages (MA) and autoregressive integrated moving average (ARIMA). These methods can also be used with exogenous inputs, resulting in methods such as ARX and ARIMAX. All of the mentioned methods are frequently utilized for load forecasting. There are also seasonal versions of these methods, though these appear to be less used. Nonlinear methods have the advantage of being adaptable, though they may be more exposed to overfitting. An example of a nonlinear regression method is nonlinear autoregressive exogenous (NARX) [8].

The development within machine learning and AI is also reflected in the field of low voltage load forecasting [35, 39, 55–59]. Methods within this load forecasting category have been common in the past decades. These methods are divided into nearest neighbors, classification and regression trees (CART), kernel-based learning approaches and deep neural networks [8].

K-nearest neighbor (kNN) is a nearest neighbor method and can be applied for nonparametric regression. The method is considered simple and can often be used as a robust reference point. It is therefore often included in many comparative studies. CART is one of the most popular methods used in machine learning. Regression trees (RT) is a form of CART, where the dependent variable takes in a continuous set of values [8].

Approaches based on kernel learning functions include Gaussian process (GP), support vector machine (SVM) and support vector regression (SVR). GP is a stochastic process built on just-in-time learning in addition to a kernel function. The kernel function can also be used in probabilistic methods based on Gaussian distribution. SVM is a nonparametric and robust method that has often been used for smaller machine learning problems. The application of SVM to regression problems is called SVR, which has several advantages for nonlinear, small sample and high dimensional data sets. A disadvantage of SVR is its high computation complexity [8, 58].

Deep neural networks have gained relevance in the field of low voltage load forecasting, with advancements in the deep learning area. Recurrent approaches utilizing long short-term memory (LSTM) were initially the most popular. The LSTM cells contain several parameters which are trainable, expanding the overall number of parameters in the training data. However, LSTM trained for specific households or substations may be sensitive to overfitting. An alternative to LSTM is a gated-recurrent unit, which has fewer parameters and is hence able to avoid problems regarding overfitting [8].

Artificial neural network (ANN) has become common in load forecasting after gaining much attention in machine learning. The most basic form of ANN is a feed-forward neural network. This network has one or several hidden layers, in addition to being trained using the back-propagation algorithm and stochastic gradient descent. Other approaches based on deep neural networks are convolutional neural network (CNN), recurrent neural network (RNN), auto-encoders for feature representation and transfer learning [8].

Probabilistic forecasting is also commonly used [60–64] and includes underlying uncertainty. With the increasing volatility at low voltage level, this category may be more suitable than time-series and machine learning forecasting methods in low voltage load forecasting. The uncertainty is often denoted as a full probability distribution, such as simple prediction intervals, quantile estimates, full continuous distributions and ensembles, allowing for more informed decision-making. To evaluate probabilistic forecasts, proper scoring functions such as continuous ranked probability score or pinball loss (PL) score are required [8].

Quantile regression is one of the most common methods of probabilistic forecasting. This approach is usually simpler regarding computations. Density estimates are another form of probabilistic forecasting, though less common than quantile regression, as the methods typically more complex. Kernel density estimates are the summation of kernel functions with specific bandwidths, which must be trained for each explanatory variable, and are commonly used for density forecasts. A disadvantage of this method is that only a few conditional variables can be included, as the parameters are time-consuming to train [8].

Table 3.2 summarizes the different approaches in load forecasting presented in this section, with corresponding methods and sub-methods.

Table 3.2: *Summary of the different approaches in load forecasting with corresponding methods and sub-methods.*

Approach	**Methods**	**Sub-methods**
Statistical and time-series	Linear regression	Simple linear regression, MLR
	Time-series	AR, MA, ARIMA, ARX, ARIMAX
	Nonlinear	NARX
Machine learning and other AI	Nearest neighbors	kNN
	CART	RT
	Kernel-based learning	GP, SVM, SVR
	Deep neural networks	LSTM, gated-recurrent unit, ANN, CNN, RNN, auto-encoders for feature representation, transfer learning
Probabilistic forecasting	Quantile regression	
	Density estimates	Kernel density estimates

3.2 Load forecasting for baseline estimation

Several load forecasting methods and techniques can be implemented for baseline estimation. This literature review categorizes load forecasting methods for baseline estimation into averaging, regression, machine learning and control group methods. Averaging methods have the advantage of being transparent and straightforward, compared to the other categories, according to [65]. The remaining categories have, however, the advantage of not being prone to baseline manipulation [45, 65].

3.2.1 Averaging methods

In averaging-based baseline estimation methods, load patterns of customers on adjacent days are assumed to be similar. HighXofY and exponential MA are typical averaging-based methods. When applied to residential customers, these methods can lead to significant errors, as residential load patterns can not be maintained at the same stable level as commercial and industrial loads. Aggregated residential load patterns are more stable than individual residential load patterns but tend to have stronger heterogeneity when compared to industrial and commercial loads. This is due to residential electricity consumption directly relating to human activities, and loads are thus more vulnerable to changes in natural and social factors [66].

Five different baseline methods were tested in [45] on real-world smart meter data from 66 residential customers. The different techniques included the averaging methods HighXofY and LastYdays, in addition to a nonlinear regression, a machine learning with neural network and a polynomial interpolation method. HighXofY consists of taking the average load of the X highest consumption days from a set of Y relevant days preceding the DR event. High5of10 was used in this paper. LastYdays takes the average of the last Y days, and not just the days with the highest consumption. This paper used the last ten days. Bias and estimation error were used to evaluate the performance of each method.

Neural network and polynomial interpolation outperformed the averaging and regression methods. The low accuracy of the HighXofY and LastYdays was expected, as these methods only had historical consumption as inputs. These methods are also prone to baseline manipulation, as customers can easily increase their load during the days prior to the DR event. The baseline is, however, harder to manipulate in the LastYdays method, as the customers must increase their consumption over a more extended period of time. Both methods are considered simple, though LastYdays is somewhat easier to implement.

In [43], the main objective was to evaluate the performance of different baseline estimation methods for residential customers. Among the tested methods were High5of10, Low4of5, Mid4of6 and exponential MA. LowXofY is the opposite of HighXofY, where the X lowest consumption days out of Y days are chosen. The low and high consumption days are rejected in the MidXofY method, and the remaining X consumption days are used to estimate the baseline. Exponential MA includes computing the initial average load of a customer and then continuously computing the exponential MA with the arrival of new observations. Historical consumption was used as an explanatory variable in all of the methods, and smart meter data from 6 435 residential and small industrial customers were utilized in the testing. MAE and estimation bias were included to evaluate the performance of each method.

The results showed that MAE was the lowest for LowXofY, subsequently being the most accurate method. However, this method had a negative bias, implying an underestimated baseline. Exponential MA had, in contrast, a bias relatively close to zero, though with lower accuracy compared to LowXofY. As a result, none of the methods could be described as the most efficient method when estimating baselines for individual customers. In addition, the efficiency of the methods was, to a great extent, dependent on the consumption patterns of the individual customer. However, when aggregating the consumption of several customers, the exponential MA method showed improved results.

High5of10, Low5of10 and Mid4of6 were used to estimate baselines in [67]. A smart metering data set from Ireland was used to test the methods, with bias and estimation error as the performance metrics. The methods excluded weekends due to limited space. Low5of10 had the best estimation error compared to the other averaging methods and a regression method. As stated by the authors, this may be due to some uncommonly high consumption days in the historical data, which the Low5of10 method could filter out in the data selection. In contrast, High5of10 had to take these high consumption days into account.

Table 3.3 summarizes all the papers mentioned in this section with respective methods, explanatory variables and performance metrics used in the different papers. All of the papers use historical load data as explanatory variables, as averaging methods require this to estimate a baseline. Bias and estimation error are the most common metrics in these papers.

Table 3.3: *Papers that utilize averaging methods to estimate baseline with corresponding method, explanatory variables and performance metrics.*

Paper	Method	Explanatory variables	Metrics
[43]	HighXofY, LowXofY, MidXofY and exponential MA	Historical load data	MAE and estimation bias
[45]	HighXofY and LastYdays	Historical load data	Bias and estimation error
[67]	HighXofY, LowXofY and MidXofY	Historical load data	Bias and estimation error

3.2.2 Regression methods

The paper given in [43] presented ANN as a regression method, in addition to the four averaging methods mentioned in section 3.2.1. ANN utilized multiple regression based on neural network to estimate the baseline with a feed-forward neural network. Historical consumption data from the past few days was implemented as the explanatory variable in the method, and MAE and estimation bias were the performance metrics. Neither ANN nor the averaging methods could be described as the most optimal method for estimating individual baselines. ANN showed superior results when aggregating the consumption, in addition to being the most efficient approach when considering bias. The authors stated that future work should focus on including meteorological variables to improve the baseline estimations further.

In [68], a cubic regression spline method is developed to improve forecasting accuracy for baseline estimation. The method was tested in an air conditioning cycling DR program in Southern California, where aggregated consumption from residential customers was used. Daily average temperature, humidity, hour and two-day lagged consumption data were among the explanatory variables in the method. As the relationship between consumption and hour was nonlinear, a cubic regression spline method was developed to model this relationship without any parametric assumptions. The novel method showed promising results, where MAPE and RMSE were used as performance metrics to evaluate the method.

A MLR method is implemented in [69] to estimate the baseline for residential customers using a smart meter data set from Ireland consisting of 5 000 customers. The performance metrics MAE, bias and overall performance index (OPI) were applied to carry out an error analysis. OPI is the weighted sum of the absolute value of MAE and bias, where both metrics had the same weight as they were considered equally important. The consumption data of different weekdays was used as input. In addition to a standard MLR method, the paper also presented an adjusted MLR method. This included calculating the difference between the actual load from the event day a couple of hours before the event started and the estimated baseline in the same time period. Further, the absolute value of the difference was added to the estimated baseline during the event.

The adjustment of MLR had a modest effect on the performance metrics, according to the results, though it somewhat improved the method. Both MLR methods were compared to unadjusted and adjusted methods of High5of10 and exponential MA. None of the MLR methods showed superior performance compared to the averaging methods. To improve the results, the authors clustered the customers into different groups of ten households each and aggregated the consumption. This resulted in a significantly improved MAE for the unadjusted methods, while the bias remained the same, hence improving the OPI. However, the clustering had a negative effect on the adjusted MLR.

Simple linear regression was used as baseline estimation method in [67]. Historical load data was used as the explanatory variable in the method, where only weekdays were included due to limited space. To test the method, smart metering data from residential customers was used. Bias and estimation error were the performance metrics presented to evaluate the method. Compared with two control group methods and three averaging methods, the regression method had the highest estimation error. As stated by the authors, this may be due to the lack of explanatory variables, especially meteorological variables such as temperature, humidity and wind speed.

In [45], a nonlinear regression method was proposed to estimate baselines for 66 residential customers in Australia, using real-world smart meter data. Temperature was the only explanatory variable in the method, and bias and estimation error were used to evaluate the performance. A nonlinear method was chosen as it could represent the relationship between load and temperature closely. Four other baseline estimation methods were also presented in the paper, including averaging, interpolation and machine learning methods. The nonlinear regression method had a low bias value, though higher estimation error compared to interpolation and machine learning. However, the regression method is considered robust against baseline manipulation, as it requires long historical data set as input. Consumers can thereby not change the outcome of this method by increasing their consumption in a smaller time period.

All papers mentioned in this section are summarized in Table 3.4 with corresponding method, explanatory variables and performance metrics. As presented in the table, historical load data is the most common explanatory variable, followed by temperature. [68] is the only paper reviewed which uses more than one variable. Bias and estimation error are commonly used performance metrics in these papers, and [69] is the only paper that presented a metric not presented in section 3.1.2.

Table 3.4: *Papers that utilize regression methods to estimate baseline with corresponding method, explanatory variables and performance metrics.*

Paper	Method	Explanatory variables	Metrics
[43]	ANN multiple regression	Historical load data	MAE and estimation bias
[45]	Nonlinear regression	Temperature	Bias and estimation error
[67]	Simple linear regression	Historical load data excluding weekends	Bias and estimation error
[68]	Cubic regression spline	Daily average temperature, humidity, hour and two-day lagged historical load data	MAPE and RMSE
[69]	MLR and adjusted MLR	Historical load data excluding weekends	MAE, bias and OPI

3.2.3 Machine learning methods

In [70], short-term load forecasting was implemented using a SVR to estimate the baseline for consumption of office buildings. SVR has the advantage of strong non-linear capabilities. The paper presents a new SVR forecasting method, which used the ambient temperature two hours before the DR event as explanatory variables. Electricity loads for four typical office buildings were used as sample data to test the method.

According to the authors, the new SVR method resulted in a higher degree of prediction accuracy and stability in short-term load forecasting compared to traditional forecasting methods. The traditional forecasting methods included historic data average methods, historic data average with morning adjustment methods and polynomial regression methods. Absolute error (AE), mean error (ME) and MAE were used as performance metrics to evaluate the forecasting methods. The authors stated that the new SVR method is suitable for real-time calculation. However, it was noted that the paper only predicted the baseline for eight hours on working days. As some DR events may be shorter, the prediction accuracy of the method may vary.

A case study for baseline estimation of a substation in India is presented in [71]. Different methods of baseline estimation calculated by forecasting techniques were analyzed. ANN is one of the methods presented, which utilized past load consumption, temperature and humidity data as explanatory variables. The method has the advantages of a high level of accuracy and the ability to respond to sudden changes, in addition to being reliable.

In the paper, the ANN method was built using a multilayer neural network and Lavenberg-Marquardt backpropagation learning algorithm. To evaluate the performance of the forecasting methods, root mean square deviation (RMSD), normalized root mean square error (NRMSE), average relative error (ARE) and normalized relative error ratio (NRER) were used as performance metrics. The ANN method gave improved results in terms of higher accuracy, improved bias and less variance compared to other baseline estimation methods, such as averaging, maximum value, adjustment and regression methods.

A novel probabilistic baseline estimation method that utilizes a deep learning-based clustering method is presented in [72]. The method consisted of four stages: a daily load profile pool construction stage, a deep learning-based clustering stage, an optimal cluster selection stage and a quantile regression forests method construction stage. The deep learning-based clustering stage was applied to handle a large number of daily patterns, improving the performance of the baseline estimation. Clusters with similar daily load patterns were found in the optimal cluster selection stage. The average consumption of the three most recent admissible days was one of the explanatory variables in the quantile regression. Real smart meter data was used to conduct the case studies in the paper.

According to the authors, the novel method showed superior performance compared to simple average methods and average daily load profile methods. The latter included different approaches of clustering and quantile regression. MSE and ARE were used as deterministic performance metrics, while PL, Winkler score (WS) and prediction interval coverage probability (PICP) were used as probabilistic metrics to evaluate the performance of the baseline estimation methods. To improve the method further, the authors suggested exploring more advanced estimation methods, such as deep neural network.

A novel stacked autoencoder (SAE) based baseline estimation method for residential customers is proposed in [66], which utilizes the data reconstruction capability of a SAE. Two SAEs are synchronously trained in the method, where one generates a pseudo-load pool, and the other is used to select a pseudo-load to reconstruct a residential baseline. To conduct the pseudo-load selection, a SVM classifier is self-trained. The paper uses a real data set consisting of 328 residential smart meter customer readings in order to validate the proposed strategy. Both MAPE and RMSE are used to evaluate the performance of the methods.

The novel method was compared with three other standard machine learning algorithms, including extreme learning machine, LSTM and stacked denoising autoencoder. In addition, the novel method was compared to existing methods, such as HighXofY and exponential MA. Results showed that the accuracy of residential baseline estimation is significantly improved in the proposed method compared to the existing baseline estimation methods. The proposed method had the advantage of handling the variation contained in the residential load compared to existing approaches. In addition, the method was robust in scenarios with high load diversity and required no additional information from customers except smart meter data. The high accuracy of the method was achieved by using real smart meter data, according to the authors.

A machine learning with neural network method is presented in [45] to estimate baselines for residential customers. The method had day of week, working day indicator and previous consumption as well as temperature as explanatory variables. The authors stated that this method was weather sensitive, making it suitable for residential baseline estimating. Real smart meter data from 66 residential customers in Australia were used to test the neural network method, in addition to four other methods, including averaging and regression. To evaluate the performance of the methods, bias and estimation error were used. Neural network outperformed the averaging and regression methods. The results showed that this method had the lowest bias, meaning it had the least tendency to over- or underestimate the baseline.

Table 3.5 summarizes the papers mentioned in this section with method, explanatory variables and performance metrics. All papers, except [70], have some form of historical load data as an explanatory variable. Temperature is also used as an explanatory variable in all papers except from [45] and [66]. ARE is the only performance metric presented in more than one of these papers. [72] is the only paper which utilizes probabilistic performance metrics, such as PL, WS and PICP, as the rest of the papers only have deterministic metrics.

Table 3.5: *Papers that utilize machine learning methods to estimate baseline with corresponding method, explanatory variables and performance metrics.*

Paper	Method	Explanatory variables	Metrics
[45]	Neural network	Day of week, working day indicator and historical load data	Bias and estimation error
[66]	SAE	Historical load data	MAPE and RMSE
[70]	SVR	Ambient temperature	AE, ME and MAE
[71]	ANN	Historical load data, temperature and humidity	RMSD, NRMSE, ARE and NRER
[72]	Deep learning quantile regression forests	Average of consumption three last admissible days from cluster, temperature and hour	MSE, ARE, PL, WS and PICP

3.2.4 Control group methods

In control group methods, residential customers are clustered based on their load patterns. Customers are further divided into a test group, and a control group within each cluster, where the test group are DR participants and the control group are non-DR participants. The baseline is estimated for the test group by utilizing the average consumption of the control group in the same cluster. Control group methods assume that the test and control group share similar load patterns within the same time period, which can be an optimistic assumption to make, especially if the group size is small [66].

Accurate baseline estimation is of great importance when evaluating the performance of DR in the settlement process. However, baseline errors are unavoidable due to the random electricity consumption behaviors of customers. Bias is often used to quantify the estimation error. As the actual baseline would never exist when flexibility is activated, the actual bias is impossible to meter and is hence rarely investigated, according to [7]. Therefore, a control group matching-based method was introduced to estimate the baseline bias in this paper.

The customers were divided into two groups: DR and control. Non-DR customers were in the control group. Bias information of the control customers was utilized to estimate the bias of the DR group during the DR event day. The control group customers had similar bias distribution with the DR group in days prior to the DR event day. The aim of this was to find the probability distribution of the baseline bias of the customers to sample from the control group. There are two main probability density estimation methods, including parametric and non-parametric. The authors chose a non-parametric approach, the kernel density estimation, as non-parametric methods narrow estimation errors caused by the assumption of a theoretical distribution.

The paper utilized a data set of more than 4 000 residential customers in a case study showing that the proposed method had overall better performance than an averaging bias method in terms of accuracy and robustness. The performance was evaluated by RMSE, MAE and Theil's inequality coefficient (TIC). TIC ranges between 0 and 1, where 0 indicates perfect estimation and 1 suggests perfect inequality or negative proportionality between the actual and estimated values.

Control groups were selected based on individual load curves from residential load curves in [28]. The residential load curves were available through smart meter data. The selected control group could adapt to the number of participants in the DR program at any time, which was one of the main advantages of this approach, as stated by the authors. Another advantage was that the method did not require a long historic to estimate the baseline, unlike regression methods. This control group selection method was compared with day and weather matching methods and a constrained regression method, where the new method gave more accurate results.

The control groups were used in three different approaches to estimate baselines: the control group load curve alone, the adjusted control group load curve and the control group load curve as an explanatory variable in a regression method. The adjustment included a multiplicative ratio, calculated as the ratio between the average load curve of the DR group and of the selected control group on the last non-DR day before the event. Experimental data from a DR program in France were used to test the three methods, and MAPE and standard deviation were chosen as performance metrics. The DR group consisted of 280 customers, while the control group had 433. Applying adjustments or a regression method to the control group did not significantly improve the accuracy of the baseline. Different sizes of the control group were also tested, where the results showed that all control groups with over 100 customers had satisfying results.

In [67], a residential baseline estimation approach based on load pattern clustering was proposed in order to improve the accuracy of baseline estimation. To extract typical load patterns of each individual customer, an adaptive density-based spatial clustering of applications with noise algorithm was proposed. This was to ensure not aggregating several dissimilar load patterns together, which may have disadvantageous effects on the clustering. Unusual load patterns were also detected and removed. Further, k-means clustering was used to divide the customers into several clusters on the basis of load pattern similarities. Baseline estimation was then performed based on the actual load of non-participants in the same cluster during periods with flexibility activation.

The baseline was calculated using both direct average of the control group and by averaging the 15 % of the control group customers with the lowest daily consumption, called the Low15% method. To test the proposed control group method, a smart metering data set from Ireland with 4 000 residential customers was utilized. Bias and estimation error were the performance metrics used to evaluate the method. Four traditional methods: one regression and three averaging methods, were also included for comparison. The Low15% method outperformed all of the other methods presented, while the direct average method could not provide a lower bias and estimation error than the averaging methods.

A cohort-based baseline method is presented in [73], where the cohorts were groups of consumers with a similar consumption pattern. The cohorts were established from correlations of non-event day load patterns using k-means clustering. Event-day smart meter data from non-DR cohort members were used to estimate baselines for DR participants. This smart meter data was used in both direct averaging and as an input in piecewise linear regression to estimate the baselines. Temperature was also used as an explanatory variable in the regression. Actual utility meter data from residents in a city in the Southern U.S. were used to test the methods. Estimation error, bias and variability were the performance metrics used to evaluate the performance.

In most cases, the cohort-based method with regression had less accurate results than the direct average method, mainly due to a lack of explanatory variables. To validate the results of the methods, they were also compared with traditional baseline estimation methods. The results showed the cohort-based baselines using direct average were more accurate compared to the traditional methods. Suggestions for further work included testing the method on data sets from other regions and developing a method to adjust cohorts over time, as the behaviour of customers may evolve.

In [74], virtual control groups were presented to improve baseline estimations of residential DR. Control groups typically require a careful selection process and exclusion of the selected customers. A virtual control group, on the other hand, provides the same benefits of control groups without having these main disadvantages. For each DR event, a new virtual control group was created by pre-collecting information on which customers wanted to participate in the event through a mobile app. The remaining customers that did not want to participate constituted the control group. However, this mobile-based participation caused a selection bias, which the authors solved by combining simple baseline estimation with the virtual control group.

Data from a residential DR program in Korea operated through a smartphone application were used in the paper, excluding weekends and holidays. Mean error (ME) and MAE were calculated to evaluate the performance. The results showed that the virtual control group with simple baseline estimation was robust. In addition, the method had better performance in terms of ME and MAE in comparison with more traditional control group-based methods.

Control group method, baseline estimation method and performance metrics implemented in each paper are summarized in Table 3.6. MAE, bias and estimation error are the only performance metrics used in more than one paper, as seen in the table. In addition, direct average is the only baseline estimation method implemented in more than one presented paper. All papers, except from [7] and [74], test more than one baseline estimation method.

Table 3.6: *Papers that utilize control group methods to estimate baseline with corresponding control group method, baseline estimation method and performance metrics.*

Paper	Control group method	Baseline estimation method	Metrics
[7]	Control group for bias	Kernel density estimation	RMSE, MAE and TIC
[28]	Individual load curves	Control group load curve, adjusted control group load curve and regression	MAPE and standard deviation
[67]	Density-based spatial clustering	Direct average and Low15%	Bias and estimation error
[73]	Load patterns using k-means clustering	Direct average and piecewise linear regression	Estimation error, bias and variability
[74]	Virtual	Simple baseline method	ME and MAE

3.2.5 Summary

Load forecasting methods for baseline estimation were divided into four categories in this literature review, which are summarized in Table 3.7 with corresponding methods. The control group category is divided into control group methods and baseline estimation methods.

Table 3.7: *Summary of the four categories of load forecasting methods for baseline estimation presented with corresponding methods.*

Category	Methods
Averaging	HighXofY, MidXofY, LowXofY, LastYdays, exponential MA
Regression	ANN multiple regression, cubic regression spline, MLR, adjusted MLR, simple linear regression, non-linear regression
Machine learning	SVR, ANN, deep learning quantile regression forests, SAE, neural network
Control group	*Control group:* control group for bias, individual load curves, density-based spatial clustering, load patterns using k-means clustering, virtual *Baseline estimation:* Kernel density estimation, control group load curve, adjusted control group load curve, regression, direct average, LowX%, piecewise linear regression, simple baseline method

The main findings of this literature review regarding load forecasting for baseline estimation can be summarized as follows:

- Averaging methods are simple, though can have high errors for residential baseline estimation. This can be improved by aggregating the consumption, as mentioned in [43]. These methods are prone to baseline manipulation, according to [45].
- Regression methods have more accurate results when aggregating consumption [43, 68, 69]. Accurate baseline estimation requires several explanatory variables [67]. These methods are robust against baseline manipulation, as mentioned in [45].
- Machine learning methods outperform other traditional forecasting methods, such as averaging and regression, in [45, 66, 70, 71]. High level of accuracy, ability to respond to sudden changes and reliability are mentioned as advantages in [70], which implemented ANN as method.
- Control group methods can give accurate baseline estimations, as shown in [7, 28, 67, 73, 74], though require large control group sizes. According to [28], only groups with over 100 customers had satisfying results.

3.3 MLR and ANN

Two of the methods introduced in the literature review in section 3.2 were of particular interest for this thesis. MLR and ANN were the two chosen methods as both are commonly utilized in load forecasting for baseline estimation and will hence be explained in further detail in this section.

3.3.1 MLR

MLR is commonly used to identify relationships between the dependent and explanatory variables. The regression and machine learning algorithm operates with several explanatory variables. The line equation for MLR is given in equation 3.4, where $\hat{y}$ is the dependent variable, and β_0 is the intercept. $\beta_1, \beta_2, ..., \beta_i$ are the regression coefficients for the explanatory variables, while $X_1, X_2, ..., X_i$ are the different explanatory variables. Adding an extensive amount of explanatory variables may not be beneficial in MLR, as the model can overfit and become over-complicated [54].

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i \tag{3.4}$$

The regression coefficient for each explanatory variable indicates the size of the effect a specific variable has on the dependent variable. In addition, the sign of the coefficient indicates whether the effect is negative or positive. The coefficient explains how much the dependent variable is expected to increase or decrease when the explanatory variable is increased by one unit if the remaining explanatory variables are constant [75].

Normally distributed error terms and linear relationships between dependent and explanatory variables are assumptions in MLR. In addition, there should be no collinearity between the explanatory variables, meaning they should not be correlated to each other. As MLR includes several explanatory variables, it can provide an understanding of the relationships between them. However, this also makes the method sensitive to outliers, which might have a disadvantageous effect on the estimation [76].

The most utilized training algorithm in MLR is ordinary least squares (OLS). The basic principle of OLS involves minimizing the sum of squared errors (SSE), where SSE is defined as the sum of the distance from each point to the optimal fitting of a line [77].

The summary of the OLS model typically includes temporary results and different probabilities. Mainly, there are three interesting factors included in the summary:

- The dependent and explanatory variables and their corresponding probability
- R^2
- Probability of F-statistic

Firstly, the dependent and explanatory variables and their corresponding probability are of importance, as the probability should always be less than the variable value in order to be significant. R^2 is also displayed in the summary, indicating the percentage of consumption variance explained by the explanatory variables. Lastly, the probability of F-statistic should be close to 0, as this implies that the model is statistically significant [54].

In addition to these three factors, a warning will be displayed if strong collinearity is indicated. Such a warning implies that the feature selection process was not carried out properly, and the model may overfit. If all three factors are in an acceptable range and there is no warning of strong collinearity displayed in the summary, the selection of explanatory variables is suitable, and the model is set for residual analysis [54].

An assumption in MLR is that the dependent variable and error terms are normally distributed [76]. Therefore, the error terms have to be calculated in order to verify this before using MLR to estimate baselines. Residual analysis can be implemented for this purpose, which examines the difference between actual and estimated baseline. Granted that the error terms are normally distributed, baseline estimations can be carried out using the model [54].

3.3.2 ANN

ANN is the building block of supervised deep learning, training the algorithm with input and output layers. Multilayer perceptrons (MLPs), also known as multilayer feed- forward neural networks, are a common type of neural network based on supervised learning. Supervised learning utilizes labelled data to generate a model, which allows the system to classify the output as known classes [78]. The general architecture of MLP includes three layers: the input layer, multiple hidden layers and output layers. The number of hidden layers relies on the data, where a few layers may result in underfitting, while several layers may make it overfitting [79].

The input and output layers are decided based on the data and the use case which is being solved. The input layer is dependent on the number of explanatory variables, while the number of outputs is equal to the number of dependent variables. Figure 3.2 illustrates an ANN with three hidden layers, an input and an output layer, where the information flows from the input layer to the output layer through the hidden layers [80].

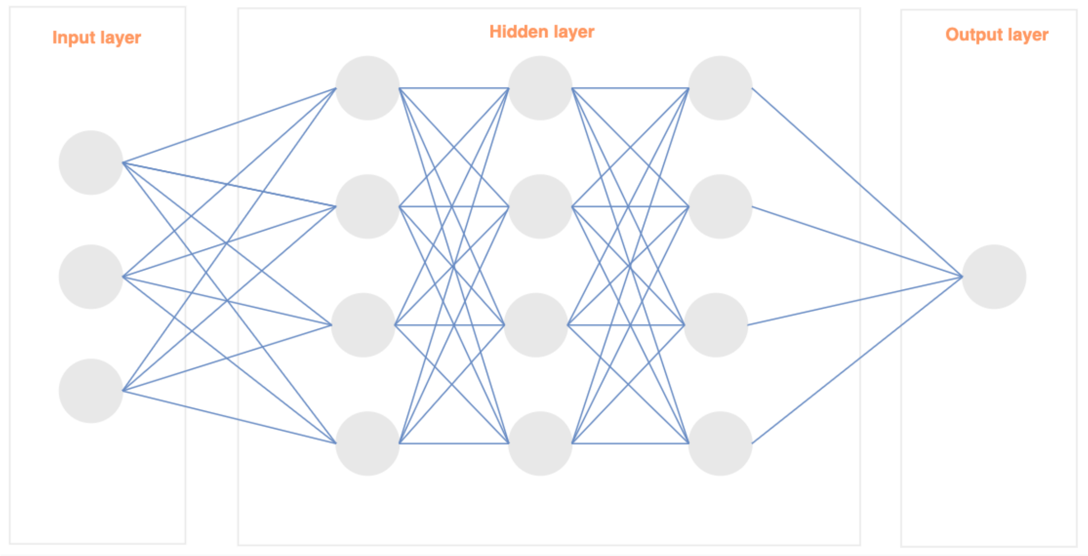


Figure 3.2: *ANN with three hidden layers, an input and an output layer [81].*

The different layers consist of nonlinear computational elements called neurons. These are the basic processing element in the network. The neurons in the input layer will be distributed to every neuron in the first hidden layer and further on to the remaining layers before ending up at the output layer. Figure 3.3 illustrates the input and output from a single neuron. x_1 to x_n in this figure represent the input values, while w_1 to w_n are the weights. Weights are adjustable coefficients able to determine the intensity of the input signal. b is the bias in the neural potential, while f is the chosen activation function, and U is the output. Activation functions convert linear outputs from a neuron into nonlinear outputs. This ensures that the ANN learns nonlinear behaviour. Rectified linear unit (ReLu) and Tangens hyperbolicus (tanh) are examples of different activation functions [82].

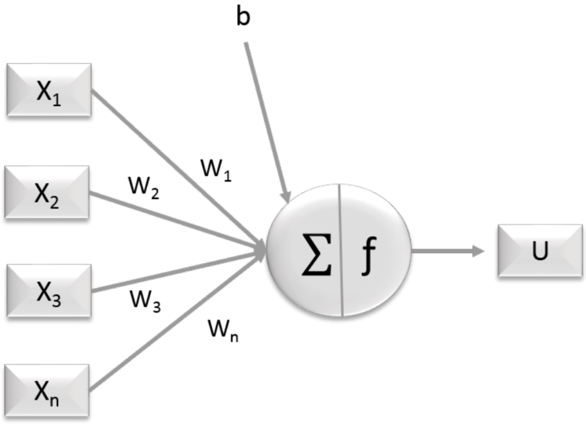


Figure 3.3: *One artificial neuron with several inputs, x_i and one neuron output, U. Bias, b, weights, w, and activation function, f, are also included [83].*

A mathematical model, as seen in equation 3.5 and 3.6, reflects the neural structure to learn and acquire knowledge from experiences and is effective in problem-dynamic and nonlinear solving such as estimation and pattern recognition. The denotations used for the equations are the same as for Figure 3.3, other than ne, which represents the neural potential [83].

$$ne = \sum_{i=1}^{n} x_i w_i + b_i \tag{3.5}$$

$$U = f(ne) \tag{3.6}$$

The more neurons or hidden layers chosen, the more computation the algorithm will require due to more calculations and optimizations. The aim is to find the minimum amount of hidden layers and neurons which can result in the highest accuracy. The number of neurons and layers can be increased step by step until a threshold or saturation point is reached, where the accuracy is not improving [80].

A MLP is trained to minimize the error between the dependent variable and the estimated values. The training algorithm of MLP can be divided into two steps: forward propagation and back propagation. In the first step, the data is passed to the input layer. From the input layer, it is multiplied by certain weightages. Every weight is assigned a very small random value, as the algorithm has yet to learn the data. It will then be passed on to every neuron present in the first hidden layer. The neuron will sum the input variables and apply it to an activation function, presented in Figure 3.3, to compute the output. The forward propagation step continuous to propagate through the hidden layers to the output layer. The output layer of the network produce the final value, i.e. the estimated value [78].

Further, back propagation is introduced. This is the tuning of weights and bias in the neural network based on error from the previous iteration, i.e. epoch. The error expresses the difference between actual and estimated values. The ideal error value is 0, as the neural network is expected to produce an output value as close as possible to the actual value. The chain rule is applied for the tuning, in turn calculating the gradient descent of each neural network parameter. If the error is larger than a given threshold, it will be propagated backwards to adjust the weights in a manner that minimizes the error. After the whole data has been passed through the neural network once, it is referred to as one epoch. After one epoch, the most optimal value for each weight is given [84–86].

Batch size is a hyperparameter that defines the number of samples to be fed through the MLP before updating the internal model parameters. There are three main configurations to compute the batch size for the training data set. When the batch size is equal to one, it is known as stochastic gradient descent. When the batch size equals the total rows in the data set, it is known as full batch gradient descent. If the batch size is less than the total number of rows, it is known as a mini-batch gradient decent [87].

For mini-batches, the data will go through a number of rows before computing an average SSE of these. The batch size is a critical hyperparameter tuned in accordance with the data in ANN. After completing this, back propagation is performed [87].

If stochastic gradient descent is applied, changing the weights of the neural network after every row, this could, in larger data sets, tend toward overfitting. This will not generate a set of equations which are generalized or able to represent the whole data. Hence, batch size equal to one is discouraged, as it leads to overfitting. For full batch gradient, the learning process of the data will not be as good, as the weightages are not changed until all the rows are passed. This will produce a suboptimal result, leading to underfitting [88].

In order to optimize the ANN, several parameters such as the batch size, epoch, kernel used, number of neurons, number of hidden layers and activation functions need to be tuned in order to find the most optimal model. Hence, ANN is a time-consuming algorithm which is only used when it seems to be worth it or in cases where the data is so large that it can not be solved efficiently in a traditional supervised machine learning algorithm, such as linear regression or decision tree random forest [80, 89].

3.4 Baseline estimation challenges

Baseline estimation is identified as one of the main challenges in the deployment and development of DR. In terms of the key challenges and barriers that pertain to residential baseline estimation, accuracy, integrity and simplicity are three critical characteristics. Through literature, error and accuracy calculations have been recognised as the most predominant challenges to baseline estimation [90].

A baseline has per definition an error, as it is formed from a synthetic and hypothesised load profile. However, in order to enable DR in Norway, accurate baseline estimations are required. The underlying error is one of the main challenges regarding baseline estimation, as the financial compensation from the flexibility depends on the accuracy of the baseline, weakening the consumer confidence in the DR program [9].

Moreover, flexibility utilization on a system level is well established with wholesale markets. Baseline estimations are applied for larger industrial and commercial loads, where the principle of power scheduling and balancing responsibility is used to operate. On a distribution level, it is argued to be more difficult to provide accurate baseline estimations, as the loads are smaller with more irregular consumption [9, 65].

An accurate baseline estimation is an essential component in evaluating the accuracy of DR programs. It should not be biased towards any parties, as this could affect the participation of the DR program. If the estimation is biased towards the DSO, actual reduction will be overestimated and the DSO has to pay more remuneration than what the consumer delivered. This could result in a less motivated DSO operating the DR program. In the case of underestimation, the consumers efforts to reduce their loads are not valued, as appropriate payment or incentives are not provided [71].

A persistent problem with baseline estimation approaches is related to the admissible days in a DR activation. A central approach in baseline estimation is the distinction of DR days and admissible days. DR days represent the days where a DR event takes place, and the consumption pattern changes. These measurements are thus not part of the baseline estimation inputs. Admissible days represent all non DR events, where flexibility is not activated, typically historical consumption patterns [65].

A customer can through the aggregator participate in the wholesale and ancillary service markets and sell services to the DSO through the LFM. The parallel participation in several markets raises an issue. From the perspective of a DSO, a day without flexibility transfer in the LFM would be considered an admissible day. However, the consumption patterns on those same days could be affected by aggregation-induced control, if they were to sell to other markets. If such days were to be considered admissible, it could be problematic in the training of baseline estimation. The baseline calculations would become more challenging as they would be dependent on prices, power system conditions and on bidding and control strategies of each aggregator [65].

Although the inaccurate estimation of customer baseline is claimed as the main barrier to DR program inefficiency, integrity is also a key factor in this issue. A DR program should not be affected by irregular consumption, which further influences the baseline calculations. A baseline estimation should hence have a high level of integrity and be robust, protecting it against the consumers attempts towards baseline manipulation [90].

A large part of the current literature is focused on the improvement of baseline estimation accuracy. Advanced approaches are expected to yield more accurate results. However, less intricate baseline methods are advocated. Simplicity and transparency in a baseline agreement between the aggregator and DSO are of the essence in order for all parties involved to understand and operate efficiently. With more advanced methods, follows greater difficulties regarding communication between involved parties, technological challenges and future disputes could arise. Hence, the majority of DR programs and pilot projects are currently using averaging methods, which have been widely accepted by system operators and third-party entities due to their simple and reliable approach [65].

The rebound effect is another challenge regarding baseline estimation that has been discovered in recent times. After a DR event has occurred, the DSO may observe a phenomenon known as the rebound effect. The conditioning loads of the customer may exceed the expected baseline level. The difference between the actual and estimated consumption following the activation of a DR event is known as the rebound effect, and is illustrated in Figure 3.4. Rebound effects that lead to peaks larger than expected may cause congestion for the DSO. This could further result in the DSO incentivizing or in this case, penalizing the aggregator based on the baseline estimations [68, 91].

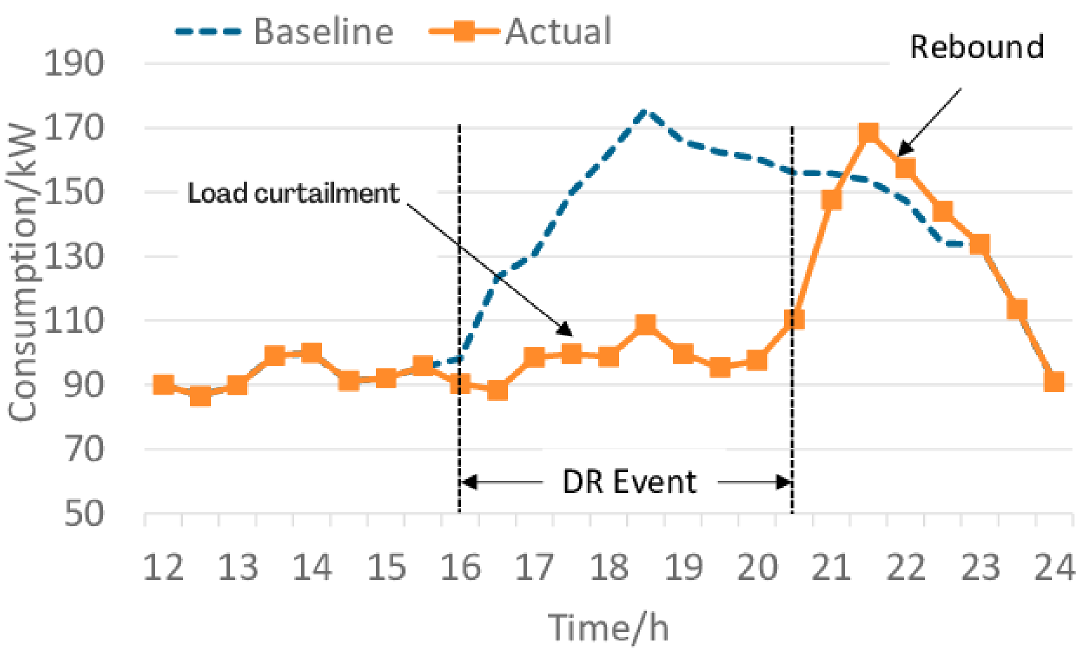


Figure 3.4: *A typical DR event with load curtailment and rebound effect [92].*

Appropriate methods for estimating the rebound effect can be developed through R&D activities and will depend on the DR resource in question and its operational characteristics. If such methods were to be created, they could be used to adjust the positions of the market participants and determining compensation measures [93].

Historical consumption is a standard explanatory variable used to estimate the baseline, as seen in Table 3.1. The actual consumption value without activated flexibility, such as the baseline, is however unknown during a DR event. One way to compute actual historical consumption is through recursive or direct strategies [94].

Recursive forecasting estimates a single time-series strategy at a time, where each forecast is determined based on the previous forecast. Direct strategy is another forecasting method, where a separate time-series strategy for each forecasting horizon is predicted, and forecasts are only computed for the observed data. Recursive forecasting is biased when the underlying method is nonlinear, whereas direct forecasting uses fewer observations when estimating the method, giving it a higher variance. Hence, there is a trade-off between the bias and estimation variance when looking at the two strategies [94].

3.5 Summary

To summarize this chapter:

- As the flexibility in a DR event can not be measured physically, the buyer and seller of flexibility must agree on a settlement process. Comparing load measurements to a baseline estimation is one way to solve this problem. Baseline is an estimate of what consumers would have consumed in the absence of a DR event.
- Low voltage load forecasting methods, which predict load demand, can be used to estimate baselines. These methods require explanatory variables as inputs, such as lagged historical demand. Performance metrics are implemented to evaluate the performance of a model. Load forecasting methods can be divided into statistical and time-series approaches, machine learning and other AI approaches, and probabilistic forecasting.
- Load forecasting methods for baseline estimation were categorized into averaging, regression, machine learning and control group methods in this chapter. The literature review revealed that aggregated consumption gives more accurate baseline estimations than residential consumption, as this is more stable. ANN and MLR were methods mentioned in the literature review, which were of particular interest.
- Accuracy, integrity and simplicity are three critical characteristics of residential baseline estimation.

4 Data processing

Data analysis is the first step of every load forecasting method, as it is crucial to understand and analyze any given data before estimating a baseline. Prior to the data being analyzed and further used in baseline estimation, it has to be prepared. This chapter will include the methodology for the data preparation. In addition, the data will be analyzed and discussed in order to get a solid understanding of it. An overview of the steps regarding data preparation and analysis for this thesis and their connection is presented in Figure 4.1. As some of the data preparation was the same as for the specialization project [12], the methodology remains somewhat similar.

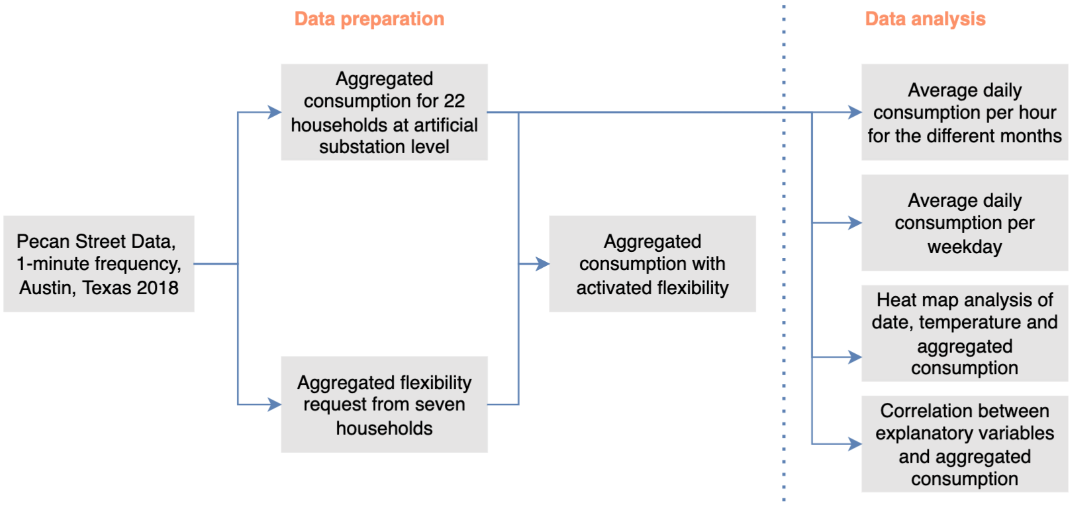


Figure 4.1: *Block diagram explaining the major steps involved in the data preparation and analysis.*

The data preparation in this thesis was conducted in order to aggregate load measurements from several households, creating consumption data for an artificial substation, as seen in Figure 4.1. Synthetic data for available flexibility from seven households were also aggregated, creating aggregated consumption with activated flexibility, as shown in the figure. The aggregated consumption needed to be further analyzed in order to evaluate the available measurements before implementing them into the baseline estimation methods. Consumption patterns based on time and temperature and the correlation between the variables were inspected. The object of the analysis was to get a deeper understanding of the data.

4.1 Data preparation

As mentioned in section 1.3, the main research question of this thesis includes how the DSO can validate demand-side flexibility at substation level. In order to achieve validation at substation level, consumption data at this level was required. This was artificially created by aggregating consumption data from several individual households in Austin, Texas. Therefore, components such as streetlights, cameras and other street furniture were not included in the artificially created substation data.

Times-series consumption data for 25 households with 1-minute frequency from 2018 was retrieved from Pecan Street Dataport [95]. The data for each household was given in kW. Several loads were included in the individual house consumption data, such as wall outlets in bedrooms, kitchen and living rooms, air compressor and washing machine. Rather than adding up all the individual loads for each household, the total consumption for each house was extracted from the "grid" column. This column represented the measured power drawn from or fed to the electrical grid. As measurements can include error and uncertainty, adding up several measurements will increase the total error in the consumption values. Single measurements provided a smaller chance of inaccurate values and were therefore preferred.

As the data was rather large, with some possible timestamps missing, linear interpolation for each house was performed in order to get more accurate results. When plotting the consumption for each house, House 4, 12 and 14 revealed negative consumption for some timestamps. The consumption values denoted what the end-user was consuming, values that, in reality, could not be negative.

There might be several reasons why these consumption values showed negative on the measuring device. These measurements are, for instance, dependent on real-world data, which can be somewhat cluttered from time to time due to some sensor malfunction, noises or mishappening during data collection. All households with negative consumption were hence removed from the artificial substation to ensure realistic data. The remaining 22 houses were added up and used for further analysis and simulation experiments. The lowest value of the aggregated consumption was 0.000 kW, the peak value was 107.9 kW and the mean value was 29.63 kW.

In order for the DSO to validate flexibility activation, consumption with activated flexibility is required. The difference between consumption with activated flexibility and the estimated baseline will be the resulting flexibility activation. This flexibility activation will be the basis of which the DSO will pay the aggregator for their flexibility services. In this thesis, it is assumed that all requested flexibility was activated. Values for flexibility request were hence required in regard to obtaining consumption with flexibility activated.

The data from Pecan Street included values for EV chargers for seven households, House 1, 2, 10, 11, 13, 15 and 19, which were extracted as values for flexibility request. Other individual load measurements could have been chosen instead. EV chargers were used as it was measured for several houses. By including more houses, several flexibility patterns could be obtained, resulting in varying flexibility request values.

EV chargers had, however, low values and did not include measurements at every timestamp of 2018. A constant value of 5 kW was therefore added at every timestamp. This was conducted to better mimic the balancing market, where flexibility is activated continuously throughout the day to balance the grid. The 5 kW increase also helped differentiate between the consumption with flexibility activation and consumption without activated flexibility, i.e. the baseline.

Consumption with activated flexibility was created by extracting the aggregated flexibility request from the aggregated consumption, assuming that all requested flexibility was activated. The aggregated consumption is the actual baseline, which is consumption without activated flexibility. All three curves are shown in Figure 4.2 for 03-08-2018.

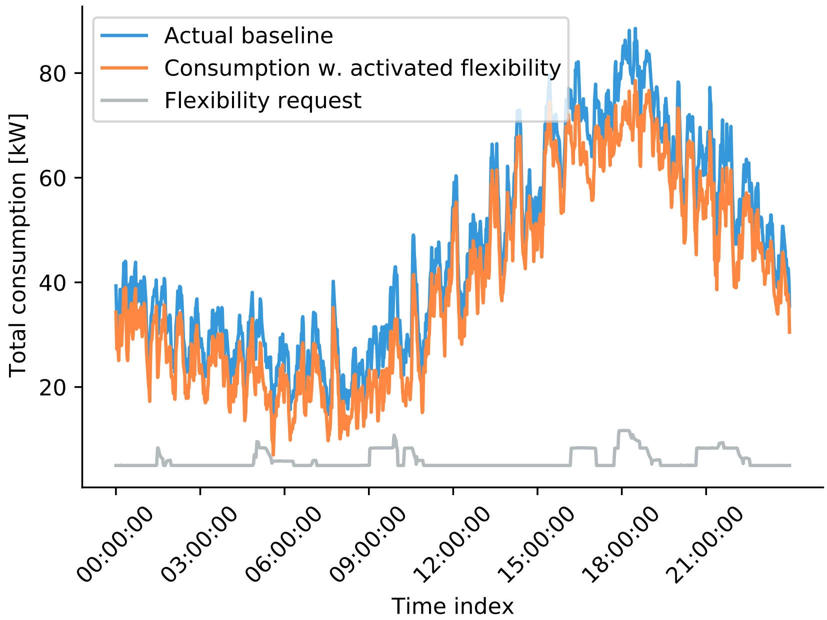


Figure 4.2: *The actual baseline, consumption with flexibility activation and flexibility request for 03-08-2018.*

After the consumption was aggregated, it was prepared for simulation experiments. This included creating explanatory variables to be used as inputs in the different load forecasting methods for baseline estimation. The timestamps were divided into several explanatory variables: hour, minute, month, day of month and weekday. Temperature data from Austin was gathered from World Weather Online [96] and added as a meteorological explanatory variable. As the temperature was given in 1-hour frequency and did not change much minute-wise, the values were interpolated with the nearest neighbor method to get a frequency of 1 minute. Nearest neighbor was chosen as this interpolation method was simple to implement.

Historical load data was also included by implementing two new variables: consumption m-1 and m-2. These variables represented the consumption in the two prior time steps. Since the data was given in a 1-minute frequency, the prior time step corresponded to the previous minute. None of the variables was categorical, and there was hence no need for dummy encoding to convert them into numeric variables.

After the data was aggregated and prepared for simulation experiments, it was analyzed in order to find specific patterns which could further lead to a greater understanding of the baseline estimation results. To analyze the artificially created substation data, it was grouped by date, hour and weekday. Power was also converted to energy for the data analysis.

4.2 Data analysis

To gain a fundamental understanding of the data utilized in the simulations, an analysis of the data was performed. It was interesting to analyze how the consumption per day in each month varied throughout the day. As the consumption data implemented in the simulation experiments contained values for a whole year, it was essential to understand the patterns of each month. The different patterns may affect how the baseline estimation methods are trained and how accurately they estimate baselines. Figure 4.3 shows the average daily consumption per hour for each month in 2018.

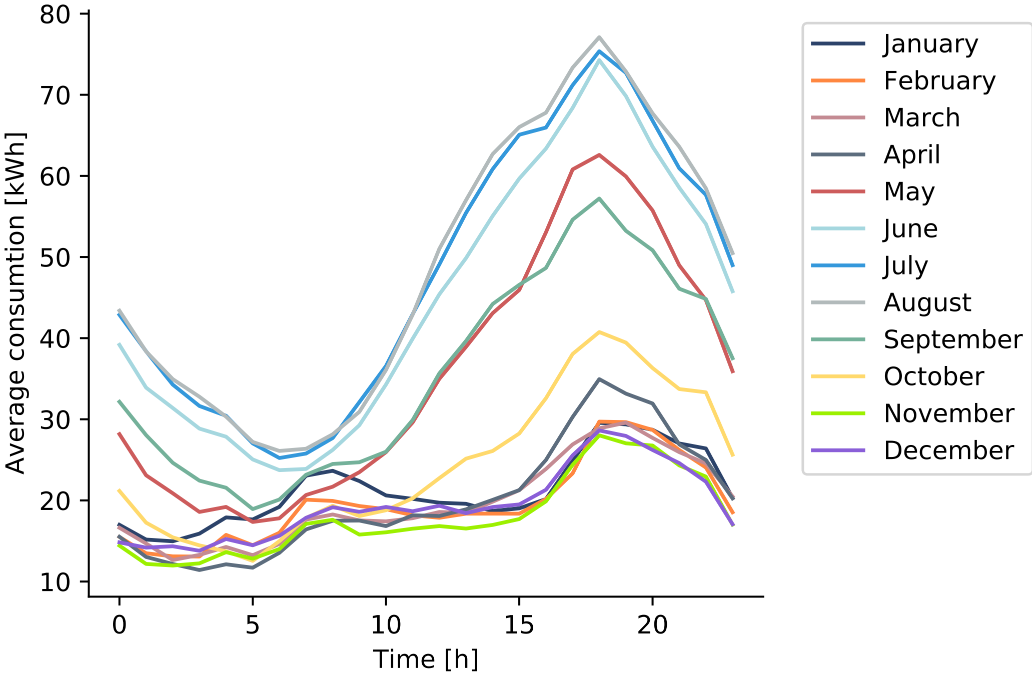


Figure 4.3: *Average daily consumption per hour for each month in 2018 for all 22 households.*

As seen in the figure, August had the highest average daily consumption per hour, followed by July and June. One reason these months are dominant might be the increased need for cooling during these months. According to [97], 87 % of homes in the U.S. utilize air-conditioning equipment, and this will account for a large portion of residential consumption during the summer. It was also stated that consumption in the summer increases rapidly throughout the day along with the temperature, with a larger peak in the afternoon or evening. This corresponds well with the consumption pattern illustrated for June, July and August in the figure. May and September also appear to have a similar pattern, with overall higher consumption throughout the day with only one significant peak.

All of the months have a peak between 6 and 7 PM, as illustrated in Figure 4.3. In addition, January, February, March, April, October, November and December have a smaller peak around 7 and 8 AM. All of these months have similar consumption patterns, with overall lower consumption per hour and two peaks during the day, as seen in Figure 4.4. This figure shows the average daily consumption for the different months in 2018, excluding the months with the highest consumption: May, June, July, August and September.

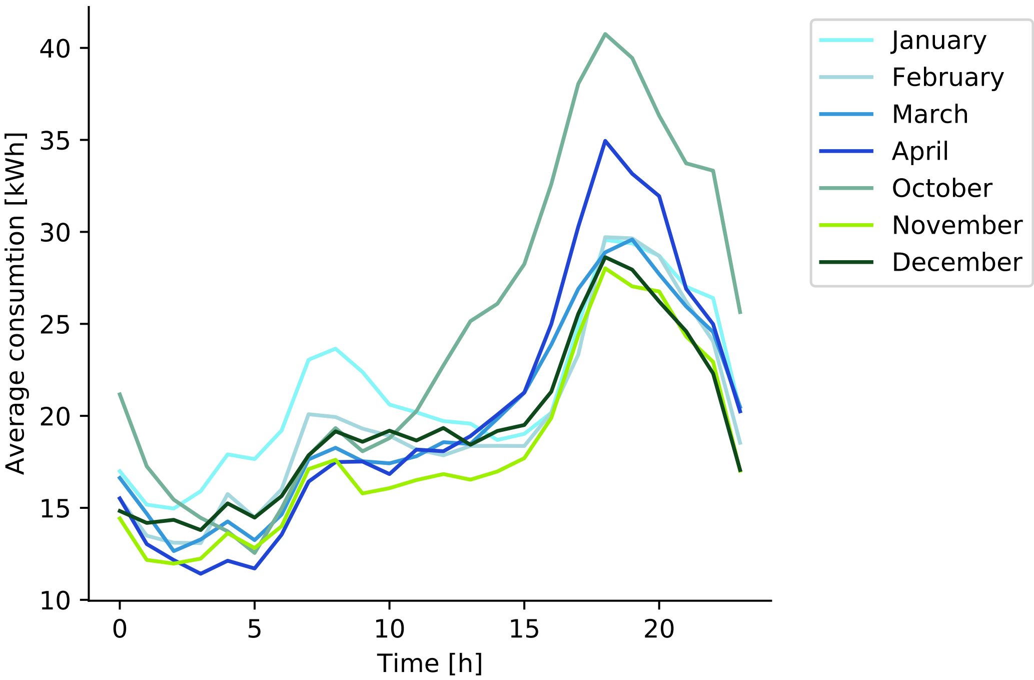


Figure 4.4: *Average consumption per hour during a day for January, February, March, April, October, November and December in 2018 for all 22 households.*

According to [97], it is common during winter months in the U.S. to have a morning and evening peak, as one-third of U.S. households use electric furnaces or heat pumps for space heating. Given the daily consumption patterns of each month in 2018, both hour and month can be important explanatory variables when estimating baselines with this consumption data.

Further, it was of interest to analyze the consumption variation throughout the week. To compare the consumption of different weekdays, the average consumption per weekday was calculated. Figure 4.5 shows the average consumption per weekday in 2018 for all of the 22 households.

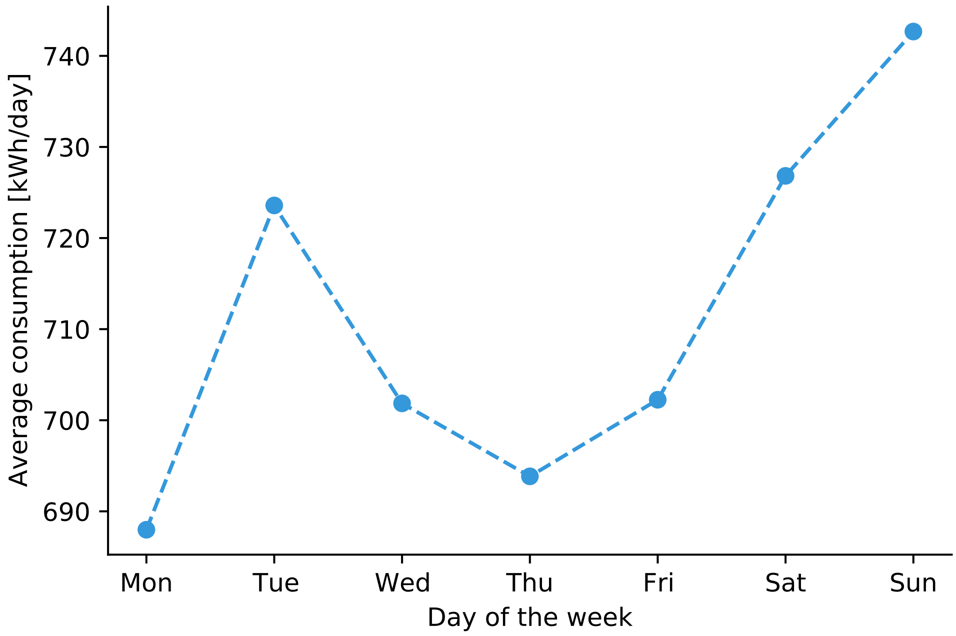


Figure 4.5: *Average consumption per weekday in 2018 for all 22 households.*

The figure portrays a variation in consumption depending on the day of the week. Weekday could therefore be an important explanatory variable in baseline estimation methods, as this may affect the accuracy of the results. However, the difference between the day with the lowest and highest average consumption is only around 55 kWh, equivalent to an increase of approximately 8 %. Monday has, on average, the lowest energy consumption, followed by Thursday. Saturday and Sunday are the days of the week with the highest average consumption, with Sunday being the highest. Tuesday has, however, an unexpected high average consumption compared to the rest of the weekdays, though the variation between the different weekdays is not particularly large measured in kWh.

A heat map was created to analyze the correlation between time of year, temperature and consumption. A heat map is a graphical representation of data and uses a color-coding system to represent different values [98]. Figure 4.6 shows how the daily consumption over time varies with change in average daily temperature, where the color axis represents consumption. As seen in the figure, the data points have a lighter color further to the right in the graph, which indicates a correlation between the increase in consumption and temperature.

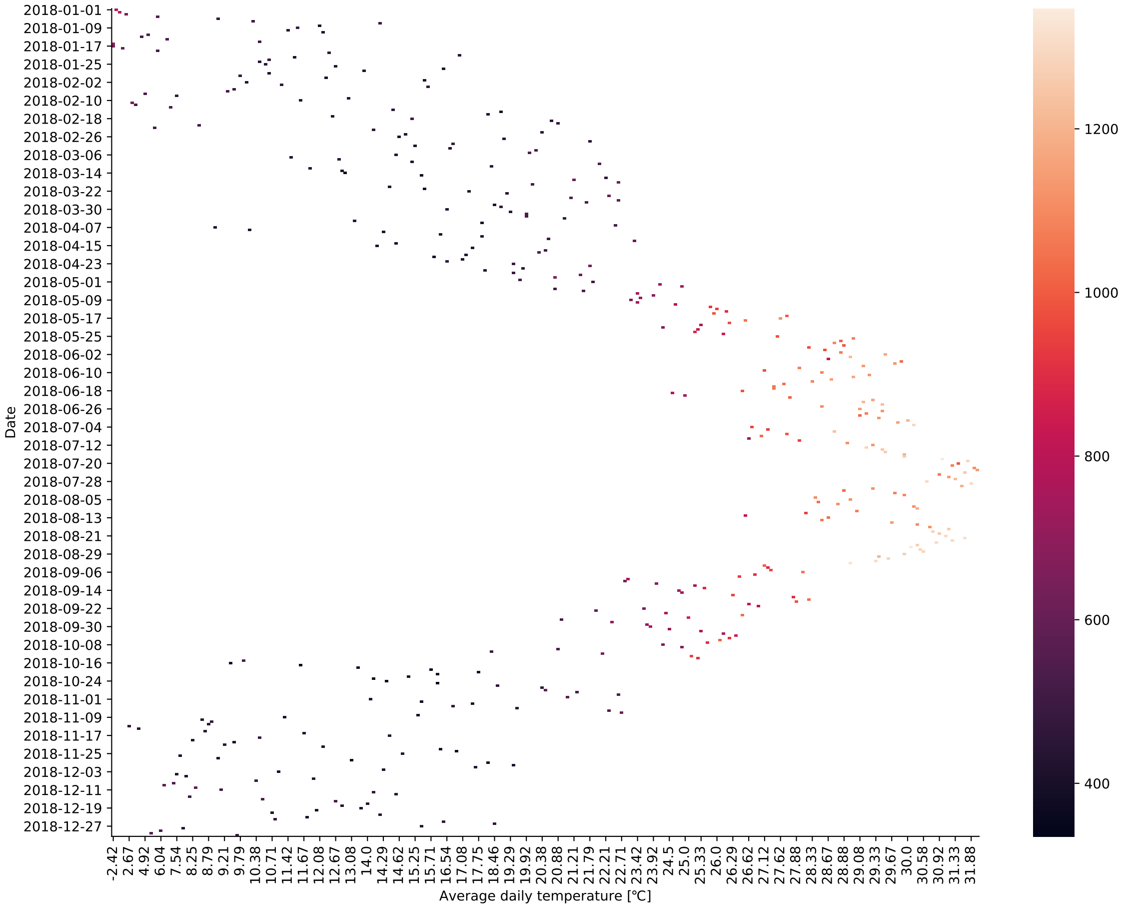


Figure 4.6: *Heat map for Austin, Texas from January to December 2018 showing daily consumption in kWh and average temperature in °C.*

The lightest points, illustrating high consumption, appear during the summer months in the heat map. This corresponds well with Figure 4.3, which shows higher average consumption per hour in these months. There are also a few lighter points towards the very start of the x-axis, where the temperature is at its lowest. In the interval between 5 and 20°C, the points are darker, indicating a lower consumption.

As mentioned in section 4.1, a set of explanatory variables were selected as inputs to the baseline estimation methods. It can be beneficial in the simulation experiments to understand the relationship between these variables and the dependent variable. The dependent variable is the baseline in this thesis, representing the aggregated consumption without flexibility activation. Correlation describes the statistical relationship between two variables and can be calculated to understand the importance of each explanatory variable. The calculated correlation value can be positive, negative or neutral. When both variables increase and decrease simultaneously, the correlation is positive. Oppositely, when one variable decreases while the other increases, the correlation will be negative. The correlation is neutral when it has a value of zero, indicating that the variables are independent of each other [99].

The resulting correlations are shown in Table 4.1. As seen in the table, the variables that correlate the most with the baseline are consumption m-1 and m-2. However, this is expected, as these two variables represent the consumption in the two prior minutes. Temperature has a correlation of 64.13 % and can hence be viewed as an important variable as well. This corresponds well with the relationship between temperature and consumption shown in Figure 4.6. Hour has the highest correlation out of the seasonal variables, with a value of 43.04 %. Minute has the lowest correlation among the variables, in addition to being the only variable with a negative correlation. However, as this was the frequency of the consumption data used in the simulation experiments, it was included as an explanatory variable regardless of the negative correlation.

Table 4.1: *Correlation between consumption without flexibility activated, i.e. the baseline, and the explanatory variables for 1-minute frequency.*

Explanatory variable	Correlation [%]
Consumption m-1	99.20
Consumption m-2	98.46
Day of month	0.226
Hour	43.04
Minute	-0.196
Month	7.552
Temperature	64.13
Weekday	2.837

4.3 Summary

In this chapter, both data preparation and analysis have been presented. To summarize the data preparation:

- Actual baseline represents artificially created substation data, aggregated from 22 households. This is consumption without activated flexibility.
- Actual flexibility activation values were extracted from seven houses and aggregated.
- Consumption with activated flexibility was created by extracting the actual flexibility activation from the actual baseline.

In the data analysis, the main findings were:

- As the average daily consumption per hour varied for each month, both hour and month can be important explanatory variables when estimating baselines with this consumption data.
- There was a variation in consumption per weekday, making weekday a relevant explanatory variable as well.
- Temperature can be an influential explanatory variable in the baseline estimation methods, as there was a correlation between the increase in consumption and temperature.
- The lagging explanatory variables, consumption m-1 and m-2, had the strongest correlation with the consumption.

5 Baseline estimation methodology

For a given baseline estimation with load forecasting, different scenarios and strategies are required with different regression methods. In the experimental set-up of this thesis, one linear, MLR, and one non-linear regression method, ANN, were the chosen machine learning algorithms. The data preparation from section 4.1 will be used as the base for this methodology chapter.

Linear regression was one of the chosen methods, as these are easy to solve and interpret. Simple linear regression only requires one explanatory variable, which may not always be sufficient to estimate baselines accurately. Therefore, MLR was chosen as the first method, as several explanatory variables can be used [54].

ANN was chosen as the second regression method. This non-linear series of algorithms endeavours to recognize underlying relationships and patterns in large data sets. Traditional forecasting methods have some limitations considering complex, non-linear relationships. Given the ability of ANN to model and extract underlying features and relationships, it was expected to provide a robust alternative to MLR. The method can in addition perform tasks that are not applicable for linear methods [87, 89].

In this thesis, two scenarios and two strategies will be implemented in the given order:

- Scenario 1: Train and test data with the same explanatory variables. m-1 and m-2 represent the consumption without activated flexibility.
- Scenario 2: Train and test data with different explanatory variables. m-1 and m-2 represent consumption without activated flexibility in the training set and consumption with activated flexibility in the testing set.
- Recursive: Use one model for recursive strategy. Based on scenario 2.
- Rectify: Use a recursive model and a second model for rectifying strategy.

Two different scenarios were created, implementing both regression methods, respectively. Scenario 1 estimated the baseline in a simple manner, using explanatory variables that would not be accessible to the DSO during a DR event in a real-world scenario. Hence, scenario 2 was introduced to account for this, implementing the actual baseline as lagging explanatory variables for the training set and consumption with activated flexibility for the testing set. The remaining explanatory variables were the same for both scenarios. Scenario 2 gave inaccurate results as the simulation was trained and tested on different m-1 and m-2 data. To further decrease the error and improve the baseline estimation, recursive strategy, followed by rectifying strategy were introduced.

Figure 5.1 showcases a block diagram of the simulation experiments conducted for the baseline estimations. The two scenarios use 1-minute frequency with MLR or ANN. The recursive strategy was carried out with one model, model 1, including either ANN or MLR. The same model was used further in the rectifying strategy, in addition to adding a second model, model 2. The two models implemented in the rectifying strategy gave a total of four different combinations: ANN+ANN, ANN+MLR, MLR+ANN and MLR+MLR. The data for these strategies were either given in a 1-minute, 5-minute or 60-minute frequency. All of these simulations were carried out to investigate and improve the performance of a baseline estimation, to complement the research question regarding the DSO verification of demand-side flexibility at substation level.

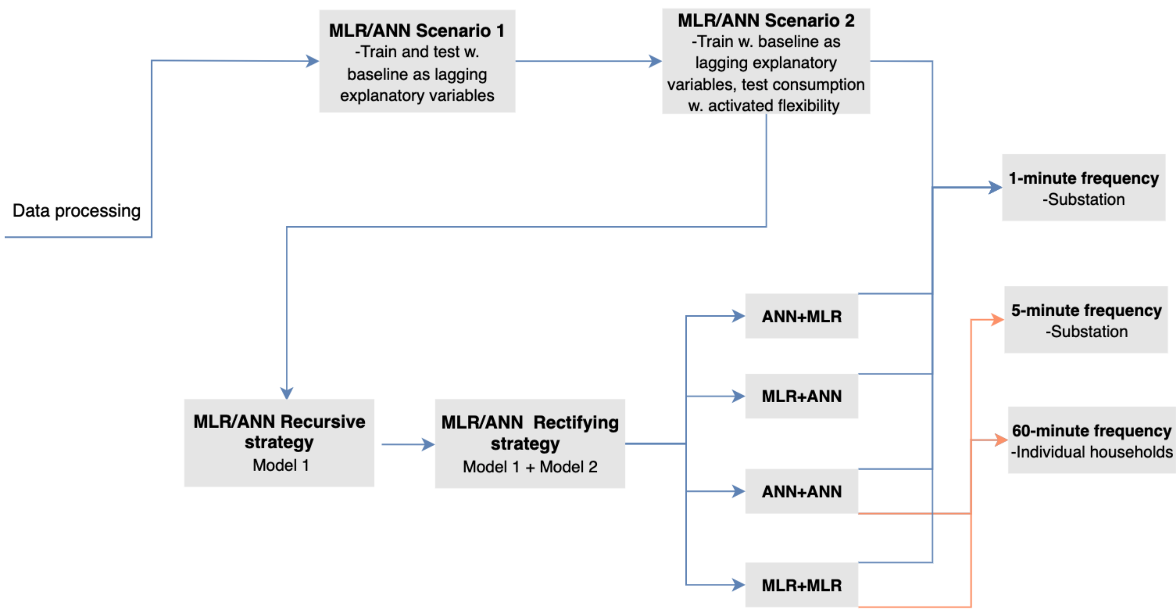


Figure 5.1: *Block diagram with data processing as the input. The rest of the diagram showcases an overview of the different scenarios and strategies used to obtain the final results.*

In this chapter, experimental results for scenario 1 and scenario 2 will be presented. Some parts regarding MLR were disclosed in the specialization project [12], and hence remain the same. Further, the methodology for the recursive and rectifying strategies for both regression methods will be explained. Different frequencies of the generated data were tested out for both regression methods. Baseline estimation was in addition to substation level, performed on individual houses. The results from the scenarios are preliminary results for the thesis and are hence introduced in the methodology chapter. Results for the recursive and rectifying strategy will be presented later on in chapter 6, as these strategies produce the final results.

5.1 Scenario 1

Scenario 1 was the most fundamental simulation implemented for the thesis and laid the foundation for the later experiments. This scenario used the actual baseline data as lagged consumption for two of the explanatory variables. In a real-world scenario, the DSO would have these measurements available before a DR event takes place on substation level, with a 1-minute frequency. Both the DSO and aggregator have the historical baseline consumption data for single households. For the DSO, this would, however, be on a 60-minute frequency, as explained in chapter 3, while the aggregator would have a frequency depending on their individual contract with the household.

1-minute frequency consumption was used for the baseline estimations, as this is the most relevant data available for the DSO at substation level. From this, seasonal explanatory variables were assembled and are presented in Table 4.1. Hour and month were included in the variables as the average daily consumption per hour varied for each month, as shown in Figure 4.3. Weekday was chosen as another variable due to the fluctuating consumption per weekday, as seen in Figure 4.5. Even though minute had a negative correlation with the dependent variable, i.e. the actual baseline, it was selected as a variable due to this being the frequency of the data. Lastly, day of month was selected as the final seasonal explanatory variable since it was mentioned in section 3.1.1 to be another standard variable.

Temperature is another form of data available to the DSO and was included as the only meteorological explanatory variable. A correlation between consumption and temperature was also discovered in Figure 4.6 and Table 4.1, validating the choice of this variable. Most papers include historical load data, as stated in [8], often as the consumption from one or several previous time steps. Lagged consumption was therefore included as explanatory variables in the form of m-1 and m-2. In addition, these two variables showed the strongest correlation with the dependent variable, as seen in Table 4.1, making them important inputs to the methods. Flexibility request and consumption with activated flexibility were not included as explanatory variables since these did not affect the actual baseline and thereby not the baseline estimation either.

The dependent variable and explanatory variables got split in the processed data. Explanatory variables are often denoted as X, while the dependent variable is known as y. The data got further separated into test and train. The training data set was used to build the methods, while the testing set was used to estimate the baselines. A total of 70 % of the data was kept for training, and the remaining 30 % was held for the test set, as suggested by [94].

Appropriate explanatory variables must be chosen in the feature selection process to build the best simulation [54]. There are several ways to perform feature selection, as mentioned in section 3.1.1. In the specialization project [12], this was carried out by selecting and dropping variables manually. This selection process resulted in dropping day of year and week of year. Hence, these nonessential variables were not included in order to prevent overfitting for any of the simulations in the thesis.

Single-step ahead time-series prediction was used for the regression methods. As stated in [35], single-point prediction only predicts the next step for a time series. For the simulations, this indicated that one prediction could only forecast the next minute, 5-minute or hour, depending on the frequency of the data. This prediction method was preferred to multi-step ahead, as it was a simpler method to implement into MLR and ANN.

To evaluate and compare the performance in a manner that fit all the simulations, calculation of performance metrics was included. MAE, R^2 and RMSE were the chosen metrics, as these are some of the most common metrics in the literature, as stated in section 3.1.2. MAE indicated the accuracy of the methods. RMSE was included to understand the distance between estimated and actual values. The metric can penalize large errors, hence being a good predictor when tuning hyperparameters or batch training a deep neural network. R^2 was included to explain to what degree the explanatory variables can account for the variance in the data. The MAE and RMSE values were desired to be low, while the R^2 should be close to 100 % to get the most accurate baseline estimation.

Both scaled and unscaled explanatory and dependent variables were tested in order to achieve the best results. Scaling is a method in machine learning performed to normalize the range of explanatory variables [100]. In theory, scaling the data should have benefited the optimization with faster convergence and higher accuracy. However, for both MLR and ANN, this was not the case, as both methods performed poorly with scaled values. Unscaling was hence preferred for all the simulations.

5.1.1 MLR

After building the MLR method, which included splitting the data into train, test, X and y, the summary of the OLS model was displayed. The summary included three interesting factors: probability of the dependent and explanatory variables, R^2 and probability of F-statistic. All three factors were within an acceptable range after the linear regression with OLS. In addition, there was no warning of strong collinearity displayed in the summary. This indicated a suitable selection of explanatory variables, and further feature selection was hence not required.

Before estimating baselines, the error terms were calculated to determine whether they were normally distributed, as this was an assumption made in linear regression [54]. The plotted error terms closely resembled normal distribution, as seen in Figure 5.2, implying that baseline estimations could be conducted in the testing data set.

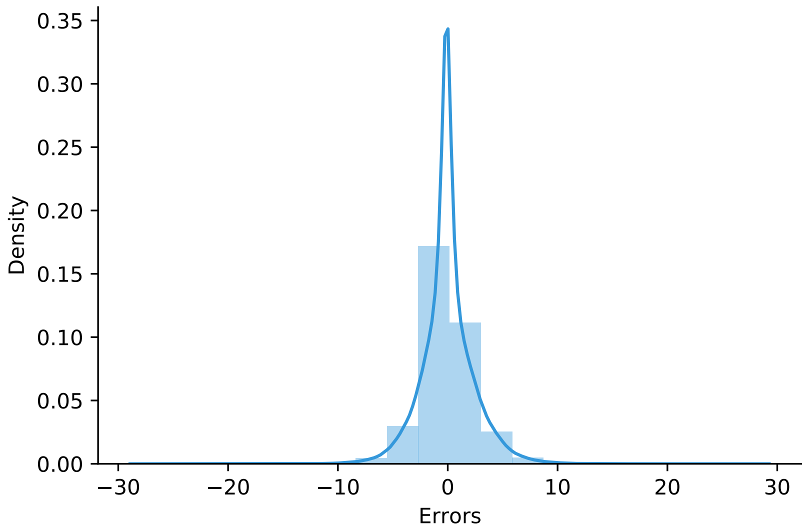


Figure 5.2: *The calculated error terms of the method, which resembled normal distribution.*

Figure 5.3 illustrates actual and estimated baseline between 7 and 9 PM, 01-11-2018 for scenario 1. As seen in the figure, the estimated baseline follows the actual baseline to a degree with a small lag. High accuracy of this estimation was expected, as this scenario had the actual baseline values for the m-1 and m-2 inputs.

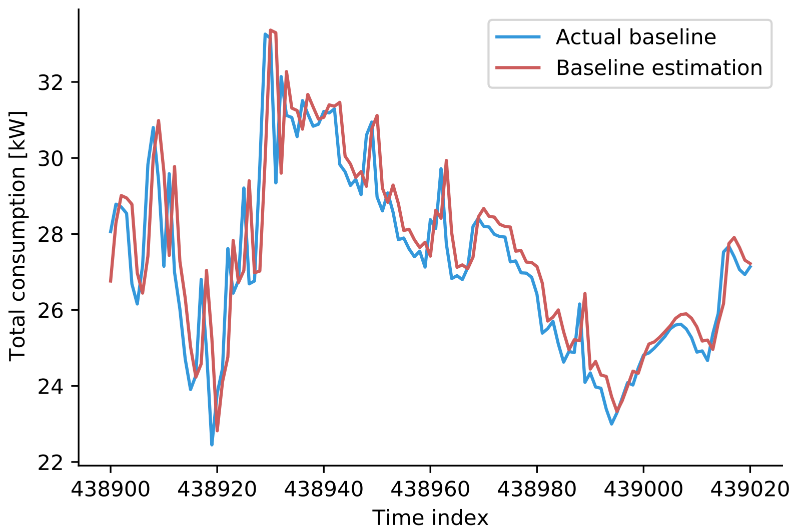


Figure 5.3: *Baseline estimation and actual baseline in scenario 1, using the regression method MLR, in time period between 7 and 9 PM, 01-11-2018.*

The actual baseline represents the consumption for the artificially created substation at a given time. This baseline will not be available in real life but was used in this scenario to help train the method and to compare with the estimated baseline, the red line. Both curves are computed from the test set. The estimated baseline should be as close as possible to the actual baseline for the DSO and aggregator to both be satisfied with the validation method.

To further interpret the performance of the MLR method, regression coefficients for the explanatory variables were calculated and are given in Table 5.1. m-1 had the highest effect on the dependent variable, as seen in the table. The coefficient value implied that when m-1 was increased by one, the dependent variable would increase by 0.9707. This could explain the slight lag between the estimated and actual baseline showcased in Figure 5.3, as the dependent variable and m-1 did not increase with the same amount. Month was the variable with the second-highest effect, followed by temperature and hour. Minute was the only variable with a negative regression coefficient, indicating that when minute increased with one, the dependent variable would decrease by 0.0004.

Table 5.1: *Calculated regression coefficients for the explanatory variables.*

Explanatory variable	Coefficient
Consumption m-1	0.9707
Consumption m-2	0.0112
Day of month	0.0002
Hour	0.0177
Minute	-0.0004
Month	0.0287
Temperature	0.0178
Weekday	0.0011

5.1.2 ANN

The first ANN scenario was a simple regression method where the test and train sets had the same types of explanatory variables. Several hyperparameters with different combinations were tried out for this scenario. The most optimal output included two hidden layers with units equal to 20 and 10, ReLu and tanh as activation functions, epochs set to 80 and a batch size of 1 000. A fraction of the baseline estimation results from this simple scenario is shown in Figure 5.4, with time stamps from 7 to 9 PM, 01-11-2018.

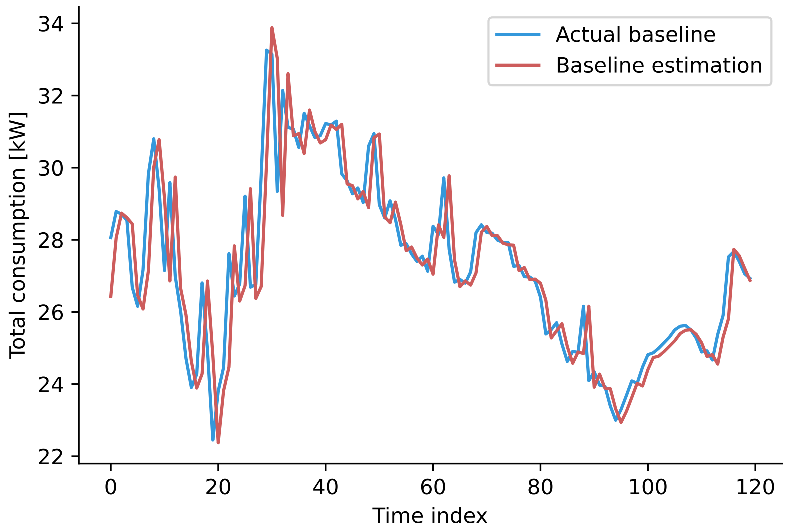


Figure 5.4: *Actual baseline and baseline estimation for scenario 1, using ANN as a regression method, with time stamps from 7 to 9 PM, 01-11-2018.*

The calculated MAE, R^2 and RMSE for both MLR and ANN are shown in Table 5.2, with results for testing and training. The testing set has a lower overall accuracy compared to the training set, as the methods are optimized for the latter. The calculated R^2 value for testing indicated that 97.60 and 97.46 % of the variation in consumption could be explained for MLR and ANN, respectively. A higher R^2 value illustrates a more accurate method and results, as mentioned in section 3.1.2. As the calculated R^2 values for both training and testing were close to 100 %, it may indicate well-fitted methods. MAE and RMSE had low values for both training and testing, which were preferred. Overall, the MLR has a slightly higher error in training but does test marginally better than ANN.

Table 5.2: *Calculated MAE, R^2 and RMSE for the train and test data sets for scenario 1. The performance metrics are calculated after MLR or ANN have been carried out.*

	Train			**Test**		
Method	**MAE [kW]**	**R^2 [%]**	**RMSE [kW]**	**MAE [kW]**	**R^2 [%]**	**RMSE [kW]**
MLR	1.646	98.43	2.410	1.260	97.60	1.883
ANN	1.595	98.43	2.306	1.336	97.46	1.891

Due to the simplistic characters of this scenario, the outputs seem quite reasonable. However, in a real-life scenario, the train and test values would not be the same, and thus, this type of scenario would not succeed.

5.2 Scenario 2

In order to make the simulation closer to a real-life scenario, the second scenario got trained with the actual baseline and tested with data for consumption with activated flexibility. This type of simulation would be closer to a real-world scenario, as the DSO only has the actual consumption without activated flexibility as available data when training a forecasting method and data with activated flexibility available when testing the data in a flexibility market.

As the values for consumption without flexibility activation are not available during a DR event, scenario 2 had to include different input values for some of the explanatory variables from scenario 1. In the test set, m-1 and m-2 represented consumption with activated flexibility rather than the baseline for scenario 2. The remaining explanatory variables stayed the same.

5.2.1 MLR

Figure 5.5 shows actual baseline, consumption with activated flexibility and estimated baseline between 7 and 9 PM, 01-11-2018, for scenario 2. The estimated baseline follows consumption with activated flexibility instead of the actual baseline. In this scenario, the difference between the actual and estimated baseline was greater than in scenario 1. This could also be observed by comparing the calculated metrics given in Table 5.2 and 5.3. As the training set was the same for both scenarios, the performance metrics calculated for this set were similar. However, the testing set metrics indicated a greater inaccuracy in the baseline estimation of scenario 2.

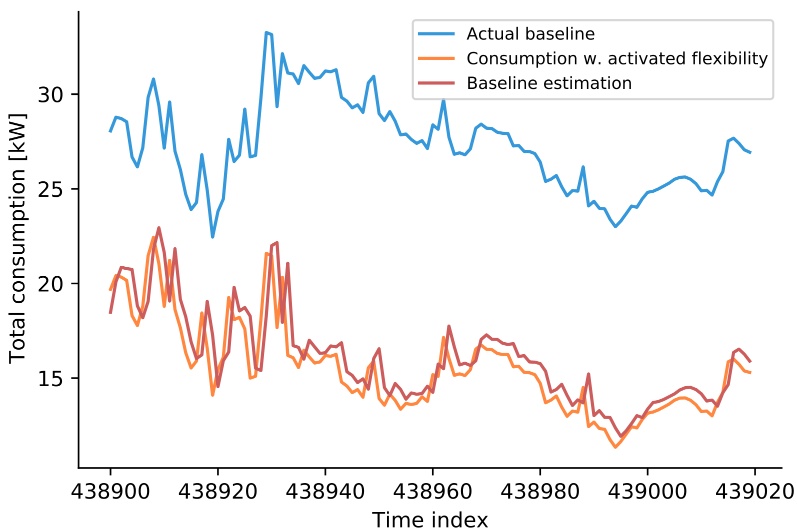


Figure 5.5: *Actual baseline, consumption with activated flexibility and baseline estimation in scenario 2 for MLR in the time period from 7 to 9 PM, 01-11-2018.*

The difference between estimated baseline and consumption with activated flexibility is the estimated flexibility activation. Figure 5.6 illustrates the estimated flexibility activation and the actual flexibility activation for scenario 2, in the time period 01-11-2018 from 7 to 9 PM. As seen in the figure, the estimated flexibility activation does not resemble the actual flexibility activation regarding the amount of flexibility or pattern. For this specific time period, the DSO would have underestimated the flexibility activation and further under-paid the aggregator.

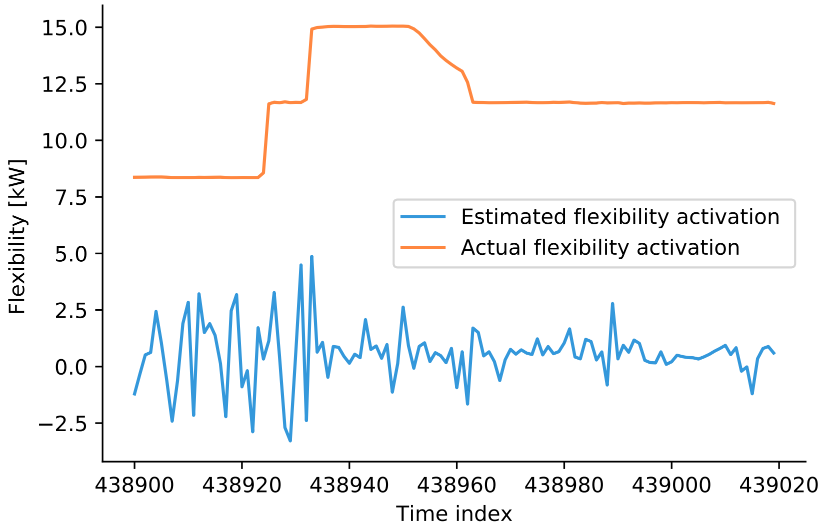


Figure 5.6: *Estimated and actual flexibility activation for scenario 2 with MLR, in the time period from 7 to 9 PM, 01-11-2018.*

The regression coefficients in Table 5.1 showed that m-1 had the most significant effect on the dependent variable. When this variable was replaced by m-1 values for consumption with activated flexibility, the baseline estimation followed the consumption with activated flexibility rather than the actual baseline.

5.2.2 ANN

In this scenario, the baseline estimation is closer to the consumption with activated flexibility rather than the actual baseline, as seen in Figure 5.7. This is due to the testing data using m-1 and m-2 values for consumption with activated flexibility. However, compared to the consumption with activated flexibility, the baseline estimation seems to lag with one time step. The baseline estimation is somewhat aligned with the actual baseline in the x-axis direction. This figure shows a fraction of the baseline estimation results, with time stamps from 7 to 9 PM, 01-11-2018. The hyperparameters used in this scenario were optimized for two hidden layers with units equal to 10 and 5, ReLu and tanh as activation functions, epochs set to 80 and a batch size of 1 000.

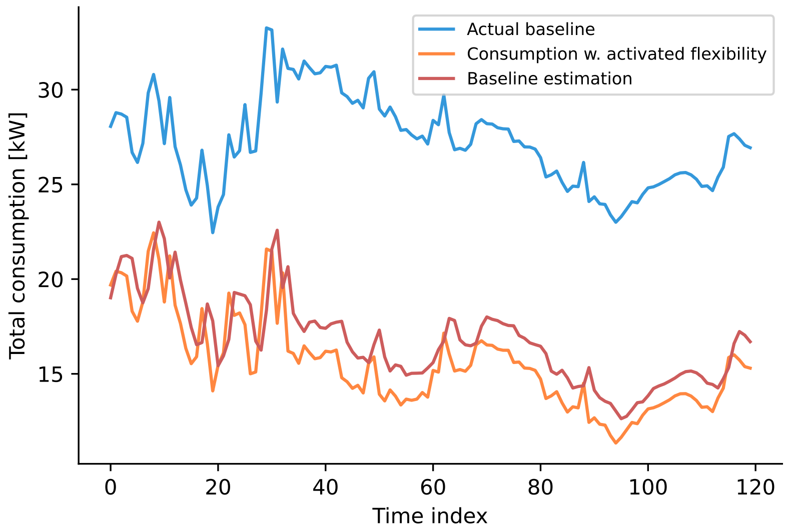


Figure 5.7: *Actual baseline, consumption with activated flexibility and baseline estimation for scenario 2, using ANN, from 7 to 9 PM, 01-11-2018.*

Figure 5.8 represents a selected time period of the estimated and actual flexibility activated in scenario 2 with ANN. In this period, the estimated flexibility activation is underestimated compared to the actual flexibility activation. The estimated flexibility activation is, in addition, negative at several time indices throughout the time period. This is due to the baseline estimation being lower than the actual baseline, resulting in lower estimated flexibility activation than what was actually activated.

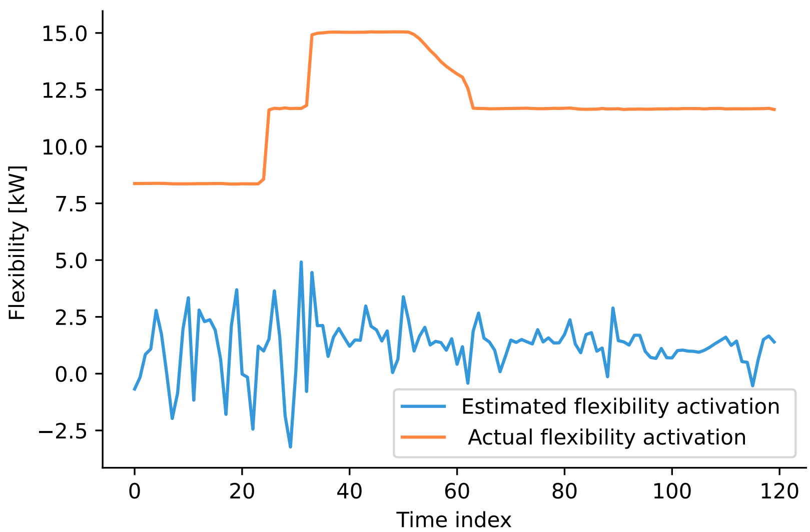


Figure 5.8: *Estimated flexibility activation and actual flexibility activation for scenario 2, using ANN, from 7 to 9 PM, 01-11-2018.*

To get a more holistic view of scenario 2, three different performance metrics are presented in Table 5.3, comparing both regression methods used. A common denominator for both methods is the train set yielding better results than the test set for the performance metrics. For the train values, the accuracy is higher for ANN, as both MAE and RMSE indicate a lower error than MLR. The R^2 is also marginally higher for ANN. For the test sets, all three performance metrics were significantly better for the ANN.

Table 5.3: *Calculated MAE, R^2 and RMSE for the train and test data sets for scenario 2. The performance metrics are calculated after MLR or ANN have been carried out.*

	Train			**Test**		
Method	**MAE [kW]**	**R^2 [%]**	**RMSE [kW]**	**MAE [kW]**	**R^2 [%]**	**RMSE [kW]**
MLR	1.646	98.43	2.410	6.253	67.08	7.030
ANN	1.584	98.45	2.288	4.887	76.70	5.706

5.3 Recursive strategy

Despite scenario 2 for both methods being more realistic than scenario 1, training with actual baseline values and testing with consumption with activated flexibility led to inaccurate results, as seen in Table 5.3. The inaccuracy mainly occurred due to the m-1 and m-2 values being replaced from training to testing. As seen in Table 4.1, consumption had the strongest correlation to these explanatory variables, and they, therefore, had to be as realistic as possible.

To improve the results from scenario 2 and make the simulations reflect a real-world scenario to a higher degree compared to scenario 1, a recursive strategy was implemented. The strategy continuously replaced the m-1 and m-2 values to make these more similar to the actual baseline values. Recursive strategy is defined as using one model, i.e. model 1, multiple times, where the prediction from the prior time step is used as input in the prediction of the following time step [101], as shown in Figure 5.9.

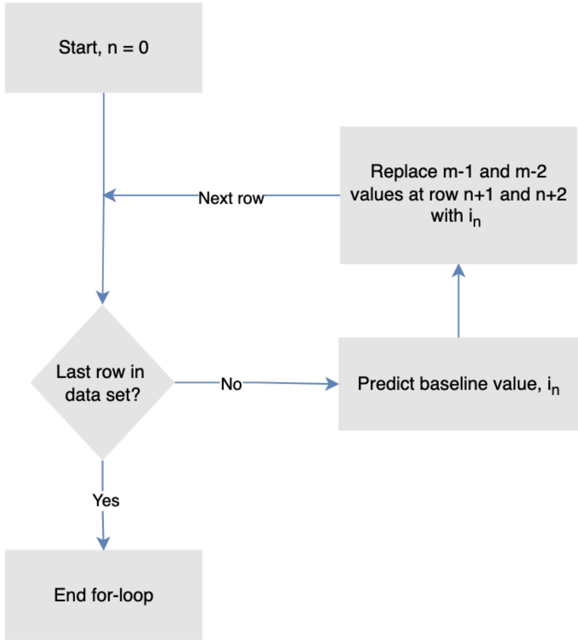


Figure 5.9: *Flowchart of the recursive strategy, where the prediction from the prior time step is used as inputs in the predictions of the following time steps.*

The recursive strategy was implemented by looping through all the data, predicting the baseline one row at a time. Each new prediction, i_n at row n, was inserted as the m-1 and m-2 values in the subsequent rows, n+1 and n+2. A for-loop was created to implement the recursive strategy. Each step was hence predicted with results from previous steps as inputs, making the m-1 and m-2 values resemble actual baseline values. The ANN method used in the recursive strategy had 20 and 25 units and 90 epochs. The rest of the hyperparameters were the same as in the two previous scenarios.

The recursive strategy was implemented on both training and testing sets. Performance metrics were calculated in both sets, before and after implementing the strategy, in order to evaluate the performance. In a real-world scenario, the recursive strategy will only be necessary for the testing set, as the training data set will have actual baseline measurements available to implement as m-1 and m-2 inputs. For the purpose of this thesis, it was, however, relevant to analyze how the recursive strategy performed in the training data set as well.

The starting points of m-1 and m-2 were set to the actual baseline values. Before implementing the recursive strategy and replacing these variables, the methods were therefore similar to scenario 1. In a real-world scenario after flexibility is activated, these values will not be available, as the actual baseline does not exist during a DR event. These values will only be available before a DR event when training the methods. However, as each prediction step in the recursive strategy depended on the previous steps, the first step needed to be predicted accurately.

If there was a significant error in the initial predictions, the error was assumed to be further accumulated in the following prediction steps. It was hence of importance to select accurate starting points for m-1 and m-2. For this thesis, the actual baseline values were seen as appropriate initial values for the recursive strategy.

Consumption with flexibility values, which can be measured in a real-world scenario, could have been potential starting points in the recursive strategy. However, as seen in the second scenario for MLR and ANN, replacing actual baseline values with consumption with flexibility values did not result in accurate estimations. With the assumed accumulating error in each prediction step, it was decided that these values were not fit as starting points.

5.4 Rectifying strategy

The results from the recursive strategy were not satisfactory, and hence a rectifying strategy was proposed. A rectifying strategy is a hybrid of a recursive and direct strategy. It seeks to combine the best properties of both strategies into one. The principle behind this strategy is to use a biased recursive forecast as input and adjust this to an unbiased forecast with smaller errors. If the mean of the residuals from the recursive strategy differs from 0, the rectifier should learn this bias and be able to correct it. This is accomplished with the help of residual error. The residual error is known as the difference between actual and estimated values. [94].

Both linear and nonlinear simulated time series from the recursive strategy, with MLR or ANN, were used to investigate the performance of the rectifying strategy. The residual error from recursive time series can include temporal structures like trends and bias. An ideal strategy would not permit any structure in the residual error, only random fluctuations that are not able to be modelled.

A second model, model 2, was created to estimate the residual errors from the recursive strategy. Model 2 was tested on both MLR and ANN, giving a total of four model combinations, ANN+ANN, ANN+MLR, MLR+ANN and MLR+MLR, when looking at model 1 and 2 together. Several hyperparameter combinations were tested out for the four combinations, and the most optimal combinations are shown in Table 5.4.

Table 5.4: *The most optimal hyperparameters for the four different rectifying strategies. All models had two hidden layers, ReLu as the first and tanh as the second activation function. The first unit is for the first hidden layer and the second unit is for the second hidden layer. Only the models with ANN have hyperparameters.*

Combination	Hyperparameter	Model 1	Model 2
MLR + ANN	Units	-	10/5
	Epochs	-	80
	Batch size	-	5 000
ANN + ANN	Units	25/20	15/10
	Epochs	90	80
	Batch size	1 000	5 000
MLR + MLR	Units	-	-
	Epochs	-	-
	Batch size	-	-
ANN + MLR	Units	25/20	-
	Epochs	90	-
	Batch size	1 000	-

The residual error was calculated by subtracting the recursive strategy estimated baseline from the actual baseline. The explanatory variables for model 2 remained the same as for model 1, but the dependent variable in the new model was the residual error rather than the actual baseline. The rectifying strategy trained model 2 on the residual errors calculated from the training data set in model 1. Before predicting the error values, the train values were either run through an ANN or MLR method.

For the rectifying strategy, the final estimated baseline with adjustment for error was calculated, providing an additional lift to the model performance. This was accomplished by adding the estimated baseline values from the recursive strategy with the estimated error from model 2. Lastly, a final MAE, R^2 and RMSE were calculated for both the train and test set.

5.5 Different frequency

The data frequency for all the simulations was 1-minute. To understand the effect of data frequency on the baseline estimation performance, data with 5-minute frequency was used for both regression methods. This entailed that values from every five minutes were extracted instead of every minute. Consumption from all 22 households and actual flexibility activation were extracted in this manner.

With the frequency adjustment, the correlation between dependent and explanatory variables changed as well. The calculated correlation between the variables is shown in Table 5.5. Similarly to the correlation for 1-minute frequency, given in Table 4.1, consumption m-1 and m-2 have the strongest correlation with the dependent variable with 5-minute frequency before implementing any of the strategies. All the correlation values remained relatively similar to those calculated for 1-minute frequency.

Table 5.5: *Correlation between consumption without flexibility activated, i.e. the baseline, and the explanatory variables with 5-minute frequency.*

Explanatory variable	**Correlation [%]**
Consumption m-1	96.26
Consumption m-2	93.65
Day of month	0.254
Hour	43.13
Minute	-0.185
Month	7.569
Temperature	64.25
Weekday	2.827

MLR and ANN were trained and tested in the same manner as with 1-minute data frequency, including recursive and rectifying strategies. Similarly to the MLR method with 1-minute frequency, the error terms had to be calculated before making any estimations in the MLR method with 5-minute frequency. This was conducted in order to determine whether the error terms were normally distributed or not. The plotted error terms closely resembled normal distribution, meaning that baseline estimations could be carried out using this method.

The rectifying strategy for MLR was implemented with MLR as the second model and for ANN with ANN as the second model. After all the results were obtained, the 5-minute simulations were compared to the 1-minute simulations, to examine whether the frequency affected the performance of the two baseline estimation methods.

5.6 Individual houses

A simulation of the baseline at individual households was also conducted to examine whether more information regarding the validation of flexibility activation could be determined at a lower level. In a real-life scenario, the DSO could potentially use the information directly from an individual household level and the aggregator to validate the amount of activated flexibility. If the individual household was to show activated flexibility of 4 kW, while the aggregator reported 5 kW, the DSO could use weighted average or a simple averaging method to decide on the flexibility activated.

The frequency on the individual household level was adjusted to 60 minutes, as this is in coherence with the momentarily AMS data submission for Elhub. The frequency decrease was accomplished in the same manner as the 5-minute frequency. The input data hence decreased from the initial 525 600 to 8 760 rows.

For the flexibility, the 5 kW that was artificially added to the flexibility on substation level was removed in this scenario. As the consumption data on household level reaches 0 kW at given time stamps, the flexibility could not include an artificial constant that exceeded this.

Three of seven houses which provided flexibility in the Pecan Street data set were chosen for the individual household simulation. House 1, 10 and 19 were the chosen houses. Both MLR and ANN methods were used in the recursive and rectifying strategies in the same manner as for the 1- and 5-minutes frequency simulations. However, only two combinations, ANN+ANN and MLR+MLR, were used in these simulations.

5.7 Summary

To summarize the simulation experiments conducted in this thesis:

- Scenario 1 was a fundamental simulation which laid the foundation for the later experiments.
- Scenario 2 was a more realistic simulation compared to scenario 1, as only data available to the DSO was implemented.
- Recursive strategy was implemented to improve the results from scenario 2, without making the methods unrealistic like scenario 1.
- Rectifying strategy was implemented to improve the results from the recursive strategy.
- 5-minute frequency and individual households were tested to see if they provided any additional information for baseline estimation compared to 1-minute frequency and substation level.

6 Results

In this chapter, the results for the recursive and rectifying strategies for both MLR and ANN will be presented. Results regarding different frequencies are displayed afterwards. Lastly, results from baseline estimations of individual houses are included.

Each result will include figures of actual baseline, consumption with activated flexibility and baseline estimation, in addition to figures illustrating estimated and actual flexibility activation for a certain time period in the testing set. As the figures only display a fragment of the testing time period, they do not represent the whole testing period. Hence, each result will include tables with calculated performance metrics to get a more holistic view of the results.

6.1 Recursive strategy

Figure 6.1a illustrates the actual baseline, consumption with activated flexibility and baseline estimation for ANN after implementation of the recursive strategy. The figure displays values for 01-11-2018. As seen in the figure, the estimated baseline does not follow the pattern of the actual baseline. While the actual baseline constantly fluctuates with a high frequency, the estimated baseline appears to have a more stable tendency and does not capture the more intense fluctuations of the actual baseline. In some periods, the estimated baseline has lower total consumption values compared to the actual baseline and is closer to consumption with activated flexibility. Figure 6.5b shows estimated and actual flexibility activation for ANN in the same time period. Similar to the estimated baseline, the estimated flexibility activation does not follow the same pattern as the actual flexibility activation.

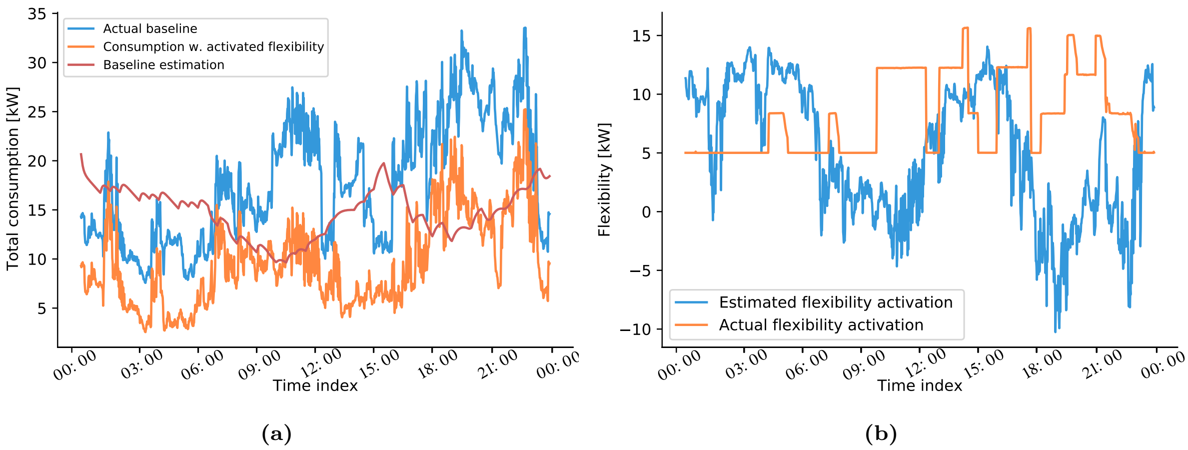


Figure 6.1: *Actual baseline, consumption with activated flexibility and baseline estimation (a) and estimated and actual flexibility activation (b) for ANN after implementing the recursive strategy for 01-11-2018.*

Figure 6.2 shows actual and estimated baseline from 01-11-2018 to 04-11-2018, in order to get a more comprehensive overview. The estimated baseline appears to follow the overall trend of the actual baseline, even though it does not seem to be accurate.

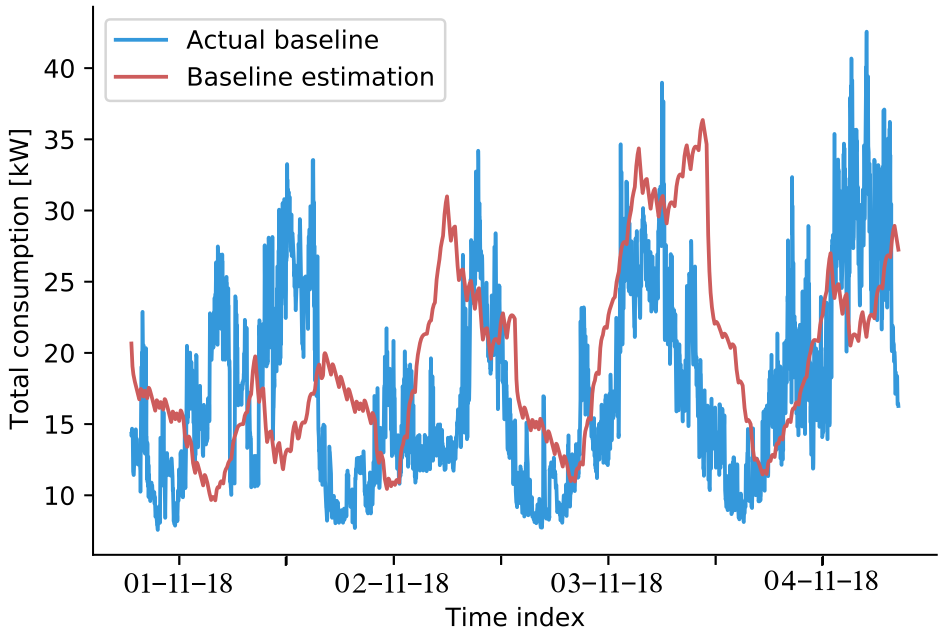
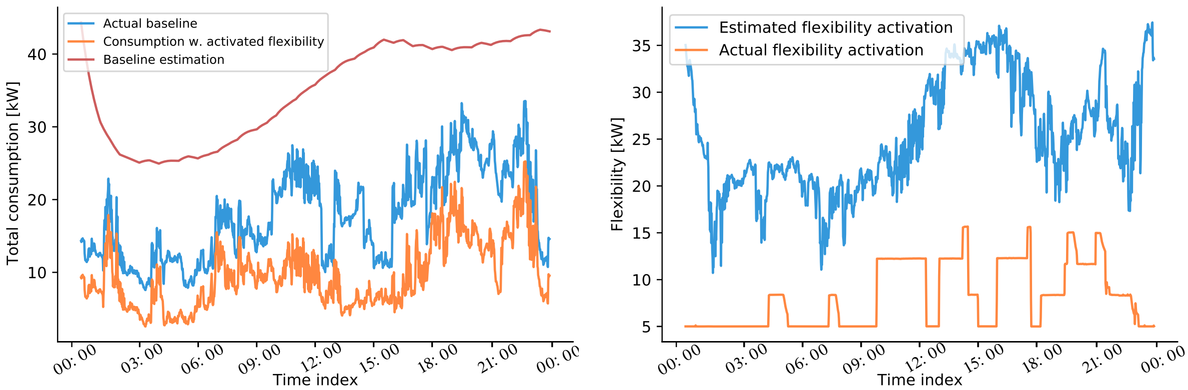


Figure 6.2: *Actual and estimated baseline for ANN after implementing the recursive strategy from 01-11-2018 to 04-11-2018.*

In Figure 6.3a, the actual baseline, consumption with activated flexibility and baseline estimation for MLR after implementing the recursive strategy are illustrated for 01-11-2018. While the actual baseline fluctuates, the estimated baseline curve is more smoothed. In addition, the estimated baseline has higher total consumption values than the actual baseline in this period. Estimated and actual flexibility activation for MLR after recursive in the same time period are shown in Figure 6.3b. The two flexibility curves are not similar to one another, as the estimated flexibility activation continuously fluctuates while the actual flexibility activation has a more stable pattern.



(a) **(b)**

Figure 6.3: *Actual baseline, consumption with activated flexibility and baseline estimation (a) and estimated and actual flexibility activation (b) for MLR after implementing the recursive strategy for 01-11-2018.*

A larger time period, from 01-11-2018 to 04-11-2018, of the actual and estimated baseline is displayed in Figure 6.4. Regardless of the two curves not being particularly similar, they do appear to follow the same overall trend to some extent, as they peak around the same time indices.

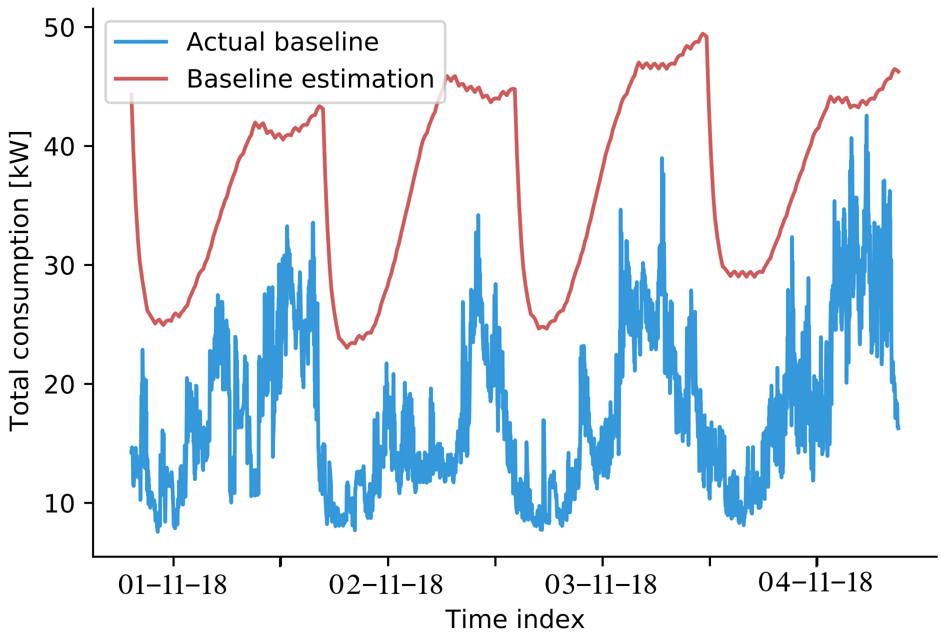


Figure 6.4: *Actual and estimated baseline for MLR after implementation of the recursive strategy from 01-11-2018 to 04-11-2018.*

Table 6.1 presents the calculated MAE, R^2 and RMSE for the train and test data sets for ANN and MLR. Values both before and after implementation of the recursive strategy are included in the table, where the values before recursive are equal to scenario 1. The two methods used have similar performance metric results before the recursive strategy, with R^2 close to 100 %, and MAE and RMSE relatively low. Both methods appear to have more inaccurate results after implementing the recursive strategy. In particular, the testing set for MLR has a poor performance after including the strategy, with a R^2 value of -92.34 %. MAE and RMSE for ANN after recursive are almost the same in training and testing.

Table 6.1: *Calculated MAE, R^2 and RMSE for the train and test data sets for ANN and MLR both before and after implementation of the recursive strategy.*

		ANN			**MLR**		
		MAE [kW]	**R^2 [%]**	**RMSE [kW]**	**MAE [kW]**	**R^2 [%]**	**RMSE [kW]**
Before recursive	**Train**	1.649	98.49	2.364	1.646	98.43	2.410
	Test	1.337	97.42	1.969	1.260	97.60	1.898
After recursive	**Train**	7.594	74.15	9.772	9.794	58.44	12.39
	Test	7.706	40.59	9.445	15.17	-92.34	16.99

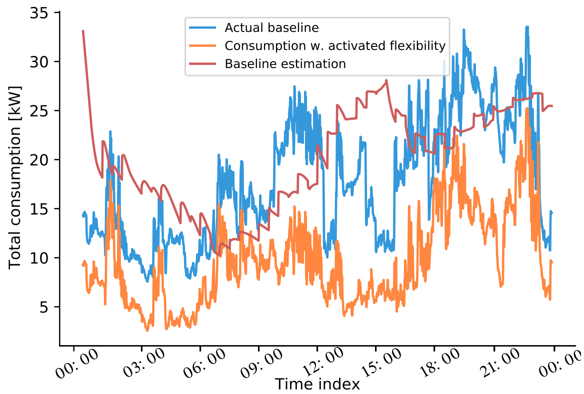
To further understand the effect of the recursive strategy, the difference between the estimated and actual baselines was calculated for both regression methods. The resulting error given in percent for row 1, 100, 1 000, 10 000 and 100 000 are shown in Table 6.2. For ANN, the error increases from the first four rows examined. However, the error decreases in the 100 000th row. A similar pattern is observed for MLR, though there is a decrease in error from row 1 to 100.

Table 6.2: *Percent error for row 1, 100, 1 000, 10 000 and 100 000 for both regression methods after implementing the recursive strategy.*

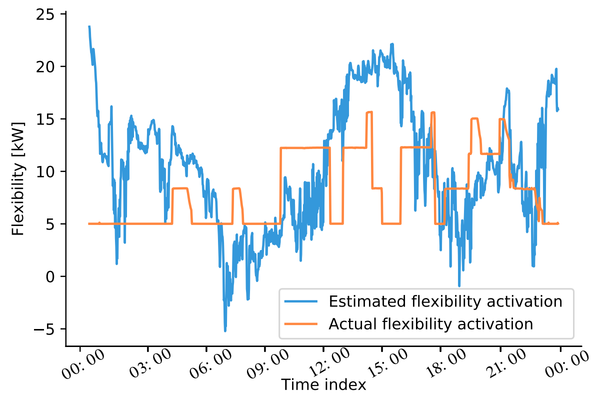
Row	**Error ANN [%]**	**Error MLR [%]**
Row 1	8.954	10.29
Row 100	10.36	7.458
Row 1 000	23.57	29.10
Row 10 000	84.65	145.8
Row 100 000	11.36	92.24

6.2 Rectifying strategy

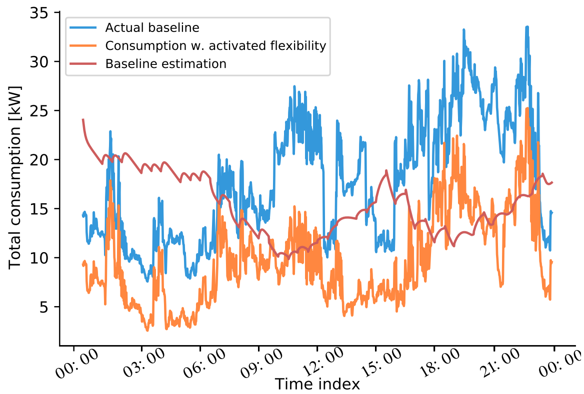
The rectifying strategy for ANN was implemented with both ANN and MLR as the second model. Figure 6.5a and 6.5c show actual baseline, consumption with activated flexibility and baseline estimation for 01-11-2018 after implementation of the rectifying strategy for ANN, with ANN and MLR as model 2, respectively. The actual and estimated baselines do not have the same fluctuating pattern in this time period. Whereas the estimated baselines have a somewhat more steady curve, the actual baseline varies greatly. Estimated and actual flexibility activation for ANN with rectifying strategy in the same period are shown in Figure 6.5b and 6.5d, with ANN and MLR as model 2, respectively. In the same manner as the estimated baselines, the two estimated flexibility activation curves do not follow the pattern of the actual flexibility activation.



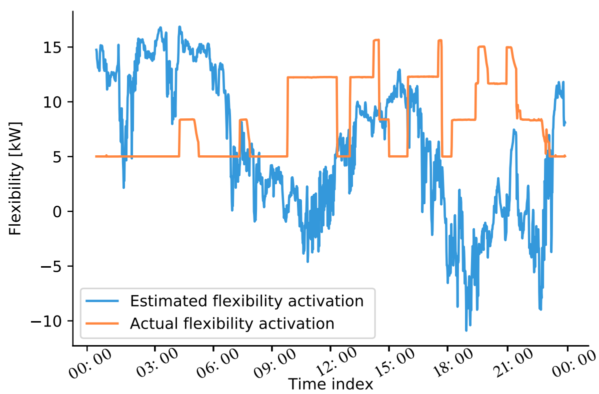
(a)



(b)



(c)



(d)

Figure 6.5: *Actual baseline, consumption with activated flexibility and baseline estimation (a) (c) and estimated and actual flexibility activation (b) (d) for ANN after implementing the rectifying strategy for 01-11-2018. ANN was used as both the first and second model in (a) and (b), while MLR was the second model in (c) and (d).*

The rectifying strategy for MLR was also implemented with both MLR and ANN as model 2. Figure 6.6a and 6.6c illustrate actual baseline, consumption with activated flexibility and baseline estimation for the rectifying strategy with MLR for 01-11-2018, with MLR and ANN as model 2, respectively. While the estimated baselines are more stable during this day, the actual baseline consists of a higher fluctuating frequency. Though the estimated and actual baselines are not similar, the estimated baselines are closer to the actual baseline than consumption with activated flexibility. Estimated and actual flexibility activation in the same period are shown in Figure 6.6b and 6.6d, with MLR and ANN as model 2, respectively. The flexibility curves do not follow the same patterns during this day.

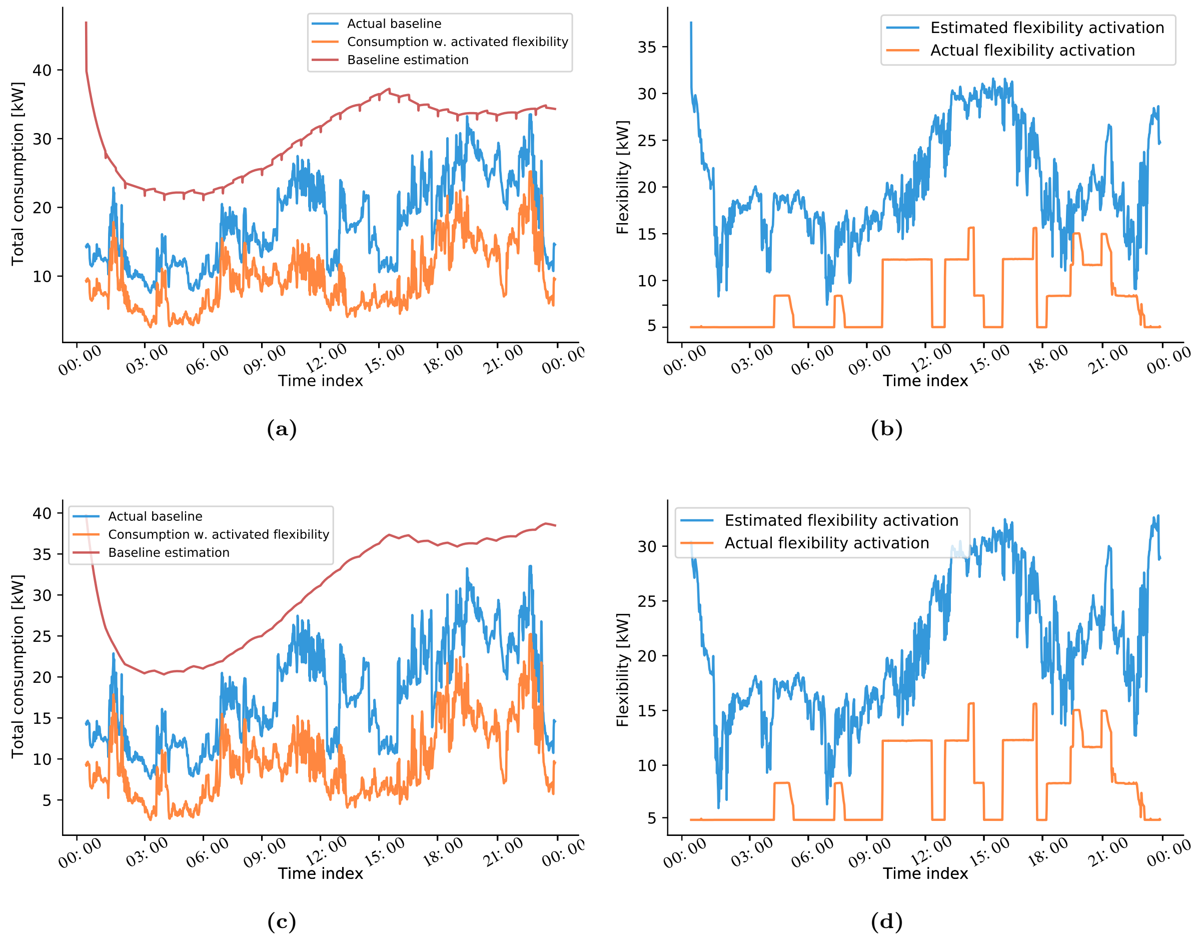


Figure 6.6: *Actual baseline, consumption with activated flexibility and baseline estimation (a) (c) and estimated and actual flexibility activation (b) (d) for MLR after implementing the rectifying strategy for 01-11-2018. MLR was used as both the first and second model in (a) and (b), while ANN was the second model in (c) and (d).*

The calculated performance metrics for all four combinations of the rectifying strategy are given in Table 6.3. The performance metrics include MAE, R^2 and RMSE for the training and testing data sets. The R^2 value of all four combinations decreases from training to testing, with MLR+ANN having the lowest value of -19.20 %. Similarly, the MAE and RMSE values increase in all four combinations from training to testing. ANN+ANN has overall the most accurate results according to the performance metrics, with the highest R^2 value and lowest MAE and RMSE values.

Table 6.3: *Calculated MAE, R^2 and RMSE for the train and test data sets for all four combinations of the rectifying strategy.*

		MAE [kW]	R^2 [%]	RMSE [kW]
ANN + ANN	**Train**	5.689	84.22	7.635
	Test	7.093	43.61	9.201
ANN + MLR	**Train**	6.488	80.24	8.543
	Test	8.445	28.98	10.33
MLR + ANN	**Train**	8.432	63.44	11.62
	Test	11.47	-19.20	13.38
MLR + MLR	**Train**	10.18	55.52	12.82
	Test	11.41	-18.36	13.33

To analyze the accuracy improvement from the recursive to rectifying strategy, the MAE decrease before and after implementing the rectifying strategy was calculated. As MAE is desired to be close to zero, a decrease in this metric before and after rectifying implies a more accurate method and thereby a performance improvement. The MAE decrease for all four combinations given in percent are shown in Table 6.4. ANN+ANN has the largest decrease in MAE for the training set with a decrease of 25.09 %. The testing set for ANN+MLR and the training set for MLR+MLR experience an increase in MAE after implementing the rectifying strategy. MLR+MLR has the greatest decrease in the testing set with a decrease of 24.79 %.

Table 6.4: *Decrease in MAE before and after implementing the rectifying strategy given in percent.*

		MAE decrease [%]
ANN + ANN	**Train**	25.09
	Test	7.955
ANN + MLR	**Train**	14.56
	Test	-9.590
MLR + ANN	**Train**	13.91
	Test	24.39
MLR + MLR	**Train**	-3.941
	Test	24.79

The correlation between the dependent and explanatory variables for the second model in the rectifying strategy was calculated to understand the relationship between the variables better. The dependent variable of the second model in the rectifying strategy is the residual error between the actual and estimated baseline from the recursive strategy. As the estimated baseline after the recursive strategy is different for MLR and ANN, the dependent variable and the lagging historical explanatory variables will differ. The resulting correlation values are given in Table 6.5 for both ANN and MLR. Month is the variable that correlates the most with the error for ANN, followed by day of month. Hour and weekday also have positive correlations with the error, while the remaining variables have a negative correlation. Consumption m-1 has the strongest negative correlation with the error for ANN.

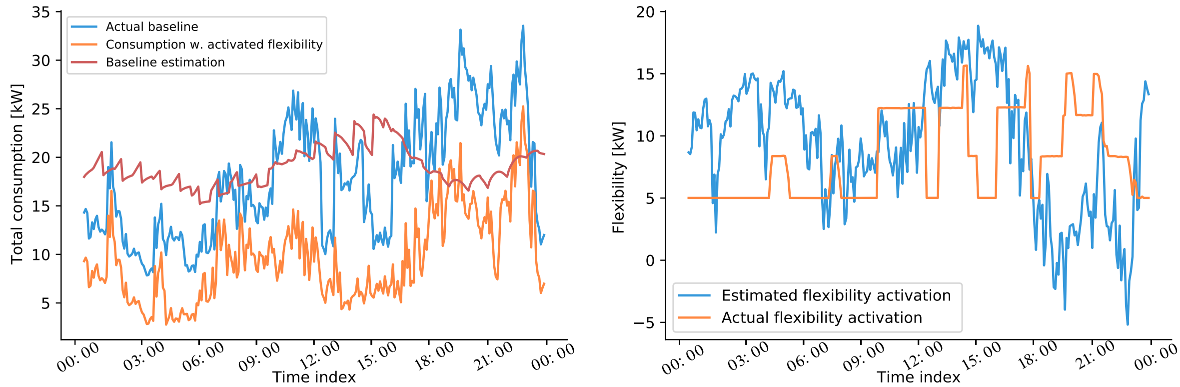
Table 6.5: *Correlation between error, which is the output of the second model in the rectifying strategy, and the explanatory variables for both ANN and MLR.*

Explanatory variable	ANN correlation [%]	MLR correlation [%]
Consumption m-1	-18.61	-6.561
Consumption m-2	-18.59	-6.568
Day of month	2.962	-4.975
Hour	1.518	6.272
Minute	-0.670	-0.209
Month	14.36	-39.46
Temperature	-15.58	10.84
Weekday	1.762	0.277

As seen in the table, temperature is the variable that correlates the most with the error for MLR, followed by hour. Apart from these two variables, weekday is the only remaining variable with a positive correlation to the error. Month has the most significant negative correlation, followed by consumption m-2. Compared to the correlations for ANN, MLR has fewer positive correlating variables in the second model.

6.3 Different frequency

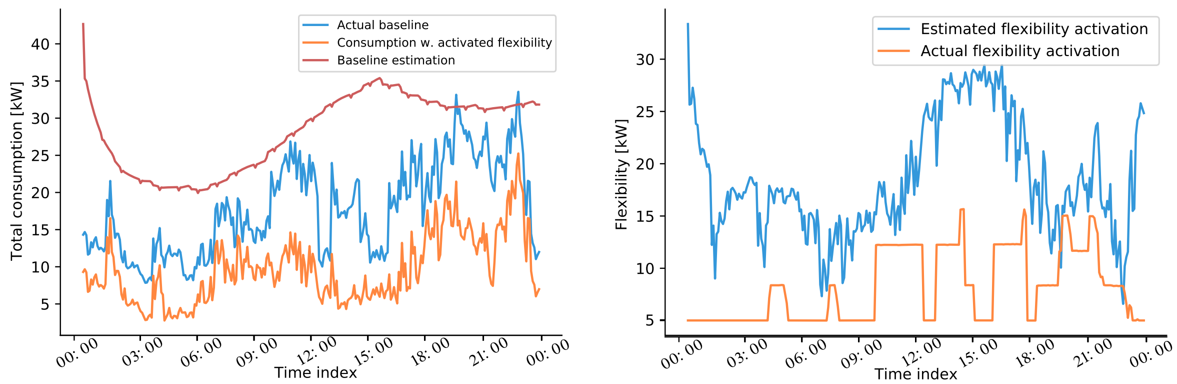
Figure 6.7a illustrates the actual baseline, consumption with activated flexibility and baseline estimation for ANN with a 5-minute frequency after implementing the rectifying strategy. ANN was both the first and second models. The figure shows values for 01-11-2018. The actual baseline appears to have a higher fluctuating frequency compared to the estimated baseline. Estimated and actual flexibility activation in the same time period are shown in Figure 6.7b. Though the two flexibility curves intersect at times, the overall pattern of the two curves is not similar.



(a) **(b)**

Figure 6.7: *Actual baseline, consumption with activated flexibility and baseline estimation (a) and estimated and actual flexibility activation (b) for ANN after implementing the rectifying strategy with a 5-minute frequency for 01-11-2018.*

Actual baseline, consumption with activated flexibility and baseline estimation for MLR with 5-minute frequency in the same time period are illustrated in Figure 6.8a. MLR was both the first and second model. The estimated baseline generally has higher consumption values during this day compared to the actual baseline. However, the estimated baseline is closer to the actual baseline than consumption with activated flexibility. Figure 6.8b shows estimated and actual flexibility activation during the same day. While the actual flexibility activation has a smoother pattern, the estimated flexibility activation fluctuates more.



(a) **(b)**

Figure 6.8: *Actual baseline, consumption with activated flexibility and baseline estimation (a) and estimated and actual flexibility activation (b) for MLR after implementing the rectifying strategy with a 5-minute frequency for 01-11-2018.*

The calculated performance metrics before recursive, after recursive and after rectifying for both ANN and MLR are given in Table 6.6. MAE, R^2 and RMSE are the metrics included, calculated for both the training and testing data sets. The results for MLR start relatively accurate before implementing the recursive strategy, with high R^2 value and low MAE and RMSE values. After the recursive strategy was implemented for MLR, the R^2 value decreased, and both MAE and RMSE increased. There is some improvement after rectifying for MLR, though the results do not appear as accurate as before the recursive strategy was implemented. The ANN method has overall better performance considering the performance metrics compared to MLR. Also, this method appears to have a decrease in performance after recursive. A slightly better result can be seen after rectifying for ANN compared to after recursive.

Table 6.6: *Calculated MAE, R^2 and RMSE for the train and test data sets for the two regression methods before recursive, after recursive and after rectifying for 5-minute frequency.*

		ANN			**MLR**		
		MAE [kW]	**R^2 [%]**	**RMSE [kW]**	**MAE [kW]**	**R^2 [%]**	**RMSE [kW]**
Before recursive	**Train**	3.708	93.07	5.060	3.765	92.79	5.161
	Test	2.894	89.94	3.888	2.847	89.98	3.881
After recursive	**Train**	6.925	76.82	9.255	9.767	58.67	12.36
	Test	7.769	40.20	9.480	14.11	-69.22	15.95
After rectifying	**Train**	5.611	84.48	7.574	10.25	54.98	12.90
	Test	7.511	43.68	9.120	10.35	-0.560	12.29

The correlation between the dependent variable and the explanatory variables for the second model in the rectifying strategy was calculated to better understand the relationship between these variables. The error between the actual and estimated baseline after implementing the recursive strategy is the dependent variable in the second model. The correlation was calculated for both ANN and MLR as these will have different results after the recursive strategy and hence different dependent and explanatory variables in the second model. For ANN, consumption m-1 has the strongest correlation with the error, closely followed by consumption m-2. Month is the only variable with a negative correlation with the error. This variable has, however, the strongest correlation with the error for MLR, followed by temperature. None of the variables has a negative correlation with the error for MLR.

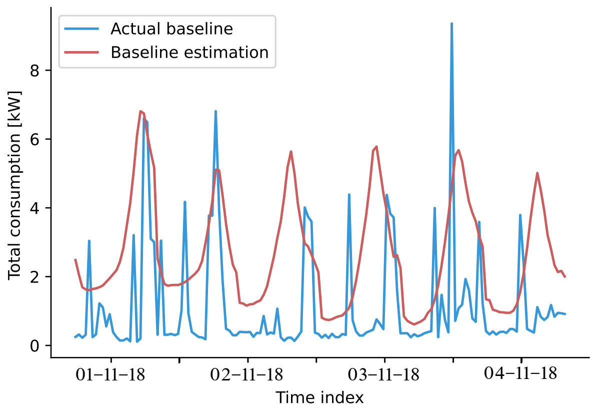
Table 6.7: *Correlation between error, which is the output in the second model in the rectifying strategy, and the explanatory variables for both ANN and MLR with 5-minute frequency.*

Explanatory variable	ANN correlation [%]	MLR correlation [%]
Consumption m-1	28.64	4.218
Consumption m-2	28.57	4.242
Day of month	8.986	4.581
Hour	17.50	7.875
Minute	0.400	0.251
Month	-2.658	36.70
Temperature	22.99	8.899
Weekday	6.906	0.441

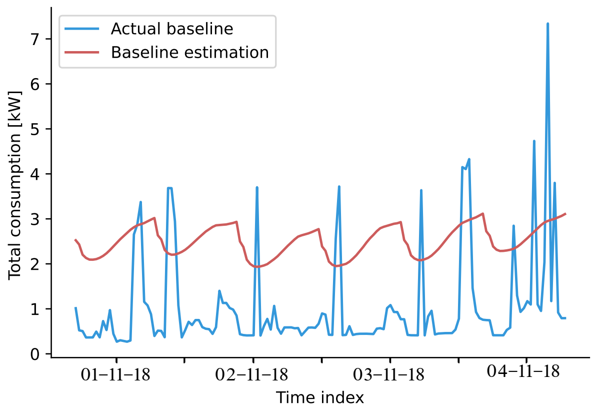
6.4 Individual houses

For the baseline estimation of the individual households, the figures in the following section will show results after the implementation of the rectifying strategy in the time period from 01-11-18 to 04-11-18. There will only be presented two model combinations for individual houses being ANN+ANN and MLR+MLR.

Figure 6.9 illustrates the actual and estimated baseline with ANN as the regression method for model 1 and 2 in Figure 6.9a and MLR for both models in Figure 6.9b. Both the MLR and ANN methods resulted in baseline estimations with smaller peaks than the actual baseline for House 1. The baseline estimation with ANN has more peaks within the given time period, in comparison with the baseline estimation for MLR. The ANN estimation has in addition, an overall steeper upward trajectory. MLR has more of a repeating pattern.



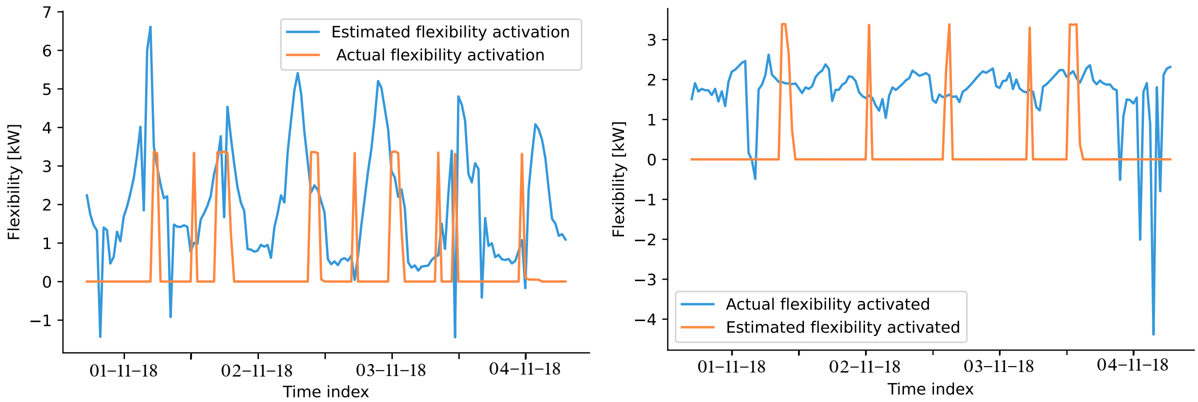
(a)



(b)

Figure 6.9: *Actual baseline and baseline estimation for ANN (a) and MLR (b) for House 1, with 60-minute frequency, after implementing the rectifying strategy in the time period from 01-11-2018 to 04-11-2018.*

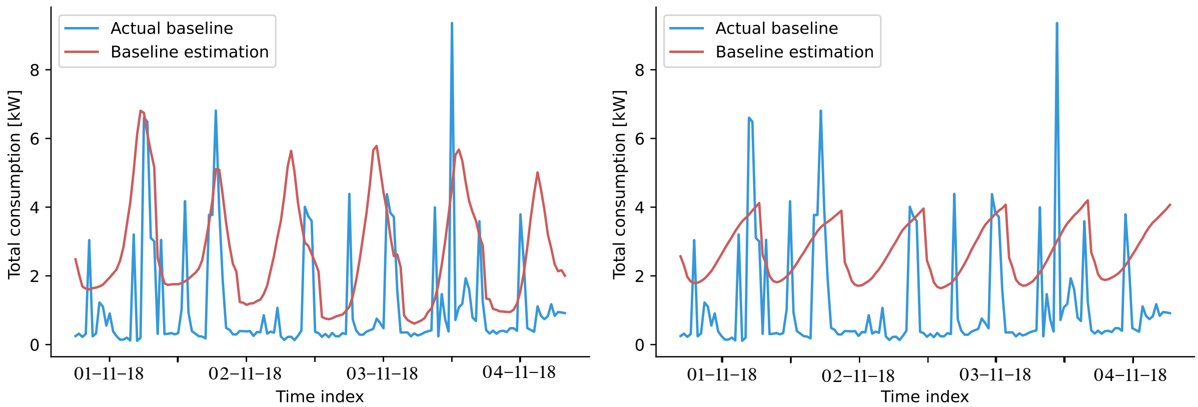
The estimated and actual flexibility activation for House 1 are presented in Figure 6.10. For ANN, Figure 6.10a, the estimated flexibility activation is somewhat fluctuating compared to the MLR method, Figure 6.10b. Baseline estimation with ANN have some peaks close to the actual flexibility activation.



(a) **(b)**

Figure 6.10: *Estimated flexibility activation and actual flexibility activation for ANN (a) and MLR (b) for House 1, with 60-minute frequency, after implementing the rectifying strategy in the time period from 04-11-2018 to 01-11-2018.*

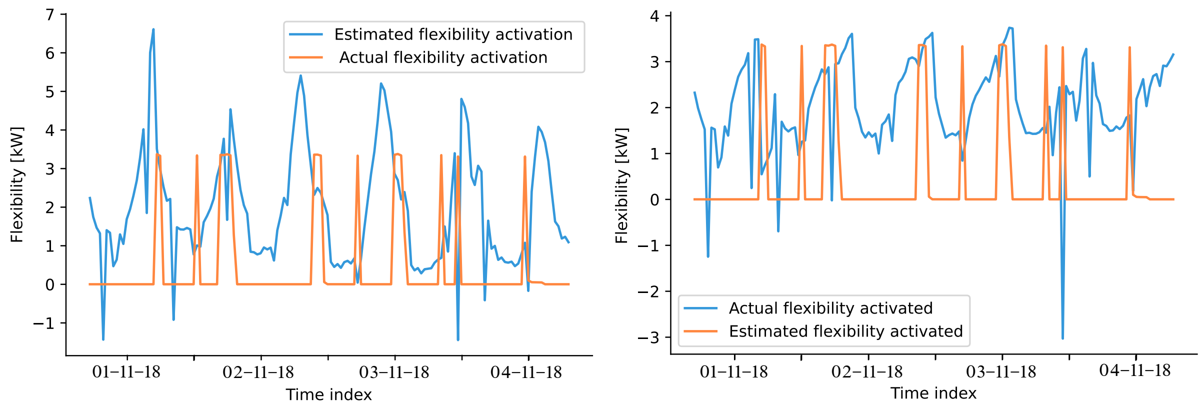
For House 10, the baseline estimations with ANN, Figure 6.11a and MLR, Figure 6.11b are presented in Figure 6.11. The baseline estimation for MLR a repeating pattern for the baseline estimation, while ANN is able to reach some of the peaks of the actual baseline. None of the methods reaches any of the troughs for the actual baseline.



(a) **(b)**

Figure 6.11: *Actual baseline and baseline estimation for ANN (a) and MLR (b) for House 10, with 60-minute frequency, after implementing the rectifying strategy in the time period from 01-11-2018 to 04-11-2018.*

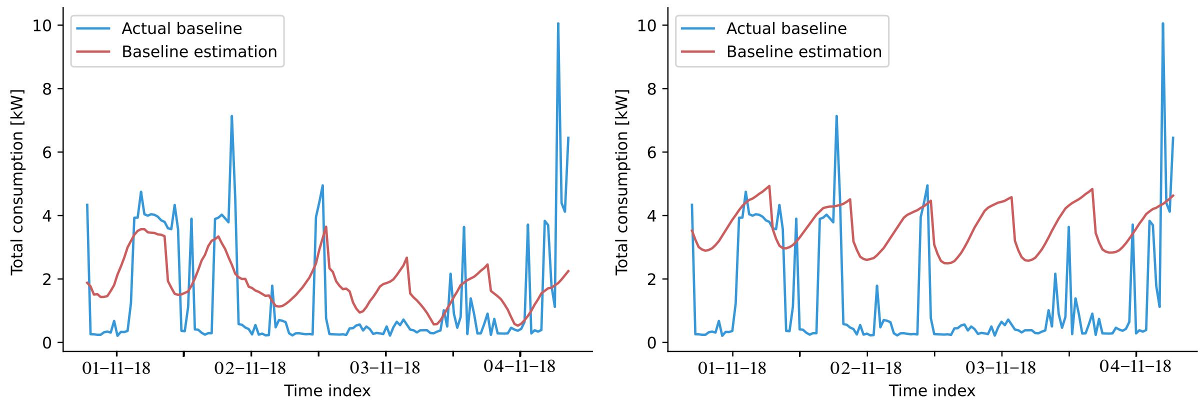
Estimated and actual flexibility activation using ANN, Figure 6.12a, and MLR, Figure 6.12b, is presented in Figure 6.12. The ANN method estimates a higher flexibility activation than the actual flexibility activation, while MLR has estimated activated flexibility peaks around the same values as the actual flexibility activation. For both methods, the estimated flexibility activation curves have several more peaks and lows compared to the actual flexibility activation.



(a) (b)

Figure 6.12: *Estimated flexibility activation and actual flexibility activation for ANN (a) and MLR (b) for House 10, with 60-minute frequency, after implementing the rectifying strategy in the time period from 01-11-2018 to 04-11-2018.*

A fragment of the estimated and actual baseline using ANN, Figure 6.13a and MLR, Figure 6.13b for House 19 is shown in Figure 6.13. Both the ANN and MLR methods result in baseline estimations with lower amplitudes than the actual baseline. Both the ANN and MLR baseline estimations have a similar drop at the start of the given time period. MLR estimates a higher baseline with a peak around 5 kW, while ANN has the highest point right below 4 kW.



(a) (b)

Figure 6.13: *Actual baseline and baseline estimation for ANN (a) and MLR (b) for House 19, with 60-minute frequency, after implementing the rectifying strategy in the time period from 01-11-2018 to 04-11-2018.*

Figure 6.14 present the estimated and actual flexibility activation for House 19, using ANN, Figure 6.14a, and MLR, Figure 6.14b, as methods. The first couple of time stamps gives a rather similar estimated flexibility activation trend for both methods. From the first actual flexibility activation peak, the trend for both methods differs more. The estimated flexibility activation for ANN has in general lower values than MLR.

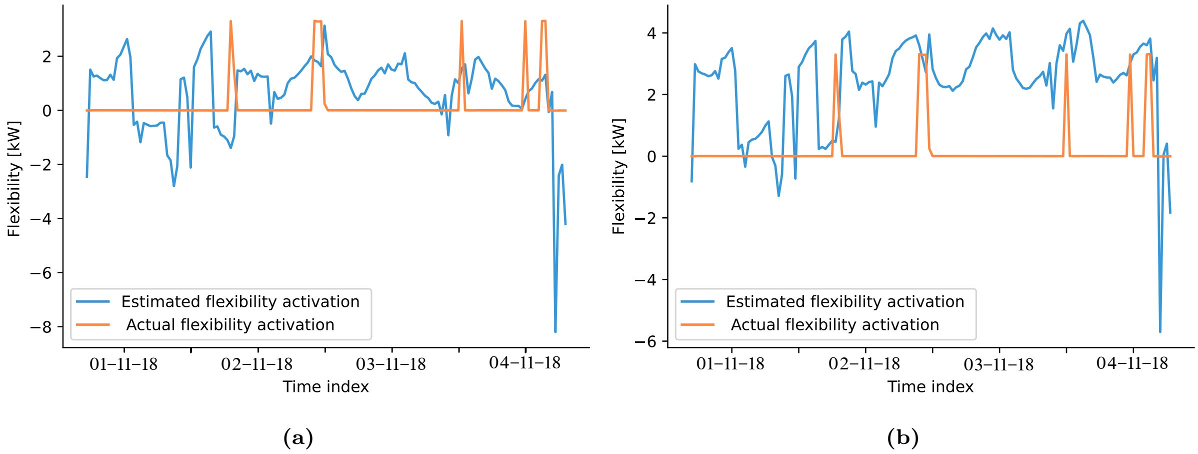


Figure 6.14: *Estimated flexibility activated and actual flexibility activated for ANN (a) and MLR (b) for House 19, with 60-minute frequency, after implementing the rectifying strategy in the time period from 01-11-2018 to 04-11-2018.*

The resulting MAE, R^2 and RMSE for all three individual houses after implementing the rectifying strategy are presented in Table 6.8. MAE and RMSE can not be directly compared for each house, as they all have different consumption patterns with different peaks. R^2 for the training data set is the highest for House 19 with ANN, and this is the only house with a positive R^2 value after testing the data. Overall, MAE and RMSE are lower for ANN, with some exceptions where the results are about the same. House 1 has the lowest R^2 values of all the houses, with the exception of the training value, which is slightly higher than the training value in House 10.

Table 6.8: *Calculated MAE, R^2 and RMSE for the train and test data sets for ANN and MLR, for House 1, 10 and 19 after implementing the rectifying strategy.*

		ANN			**MLR**		
		MAE [kW]	**R^2 [%]**	**RMSE [kW]**	**MAE [kW]**	**R^2 [%]**	**RMSE [kW]**
House 1	**Train**	0.888	36.41	1.199	1.040	22.47	1.325
	Test	1.169	-48.77	1.427	1.716	-141.9	1.820
House 10	**Train**	1.190	36.31	1.625	1.341	28.28	1.724
	Test	1.625	-42.71	2.073	1.833	-42.21	2.069
House 19	**Train**	1.177	43.87	1.593	2.790	38.33	1.670
	Test	1.279	22.29	1.632	2.208	-83.45	2.508

To gain a better understanding of the performance of the second model, the correlation between the dependent and explanatory variables was calculated. The dependent variable for model 2 is the residual error between the actual and estimated consumption from the recursive strategy. Table 6.9 presents the calculated correlation for all the individual houses. More than half of the values are negative. This indicates that the residual error will move in the opposite direction of the explanatory variables. The temperature has a positive correlation for all houses except ANN in House 1, making it the explanatory variable with the most number of positive correlations. However, these values are rather small, as the optimal correlation is 100 %. Month is the explanatory variable that has the lowest correlating values.

Table 6.9: *Correlation between error, which is the output in the second model in the rectifying strategy, and the explanatory variables for House 1, 10 and 19. The correlations are both for the MLR and ANN methods.*

Explanatory variable	**Correlation [%]**					
	House 1		**House 10**		**House 19**	
	MLR	ANN	MLR	ANN	MLR	ANN
Consumption m-1	-18.16	19.58	1.528	-17.43	-10.82	-6.86
Consumption m-2	-17.26	27.12	1.48	-16.22	-10.89	-6.19
Day of month	-3.77	-37.92	-1.63	3.92	-5.58	-11.86
Hour	-3.04	49.48	8.52	-1.35	2.289	1.58
Month	-30.71	-55.77	-14.14	-24.71	-34.18	-22.92
Temperature	6.55	-54.85	3.18	8.92	16.03	7.481
Weekday	-0.32	37.92	-1.37	7.22	-0.09	-1.59

7 Discussion

In this chapter, each result from the simulation experiments will be discussed and compared. Firstly, the results from the implementation of the recursive strategy are discussed, followed by the results from the rectifying strategy. Further, the different frequencies tested and baseline estimation of individual houses are discussed. Lastly, factors that may have affected the results and other subjects relevant to the DSO in the settlement process are mentioned and discussed.

7.1 Recursive strategy

The estimated baseline with the recursive strategy for ANN does not appear to be accurate, according to Figure 6.1a, as it does not capture the intense fluctuations of the actual baseline. In addition, the estimated baseline is more similar to consumption with activated flexibility in periods. A similar observation is made for estimated flexibility activation in Figure 6.1b, which does not illustrate the same pattern as actual flexibility activation. During this day, the DSO would both have over- and under-paid the aggregator for the activation of flexibility, as the estimated flexibility activation is both over and under the actual flexibility activation. However, these figures only display a fragment of the testing period and do, therefore, not illustrate the overall performance of the strategy.

The calculated performance metrics given in Table 6.1 demonstrate a more realistic view of the method performance. Before recursive, the ANN method appears to be relatively accurate, with R^2 values close to 100 % for both training and testing. The MAE and RMSE values are relatively small, which indicates an accurate method according to section 3.1.2. However, the performance of the method decreases after implementing the recursive strategy. The average difference between the estimated and actual baseline is 9.445 kW in the testing set, as implied by the RMSE value. This is a rather significant increase of 379.7 % compared to before recursive, where the average difference was 1.969 kW in testing. The same development can be seen for MAE, which increases by 476.4 % from 1.337 kW to 7.706 kW in the testing set, indicating an inaccurate baseline estimation.

R^2 also suffers from a decrease in performance. While the values before recursive were close to 100 %, the value for training after recursive decreases to 74.15 %. In the testing set, this metric decreases further to 40.59 %, signifying that the explanatory variables can only explain 40.59 % of the variations in the actual baseline. However, the performance metrics before recursive are equal to scenario 1, as the m-1 and m-2 values were the actual baseline values. These variables had not yet been replaced in the recursive strategy. This scenario was the most accurate scenario of the two presented in chapter 5, in addition to it using data not available to the DSO in the settlement process. Therefore, these metrics were expected to be more accurate than the metrics calculated after implementing the recursive strategy.

If the DSO utilized this baseline estimation method in the validation process, they would have considerably over- or under-paid the aggregator, as the average difference between the estimated and actual baseline is close to 10 kW. Therefore, the DSO would most likely not be choosing ANN with the recursive strategy to validate the flexibility. If the DSO chose this method, the aggregator would probably not accept it, as it may negatively affect both them and the end-users to validate the activated flexibility inaccurately.

As seen in Figure 6.3a, the estimated baseline for MLR with the recursive strategy does not resemble the actual baseline. Neither pattern nor amount is similar to the actual baseline. Similar behaviour can be seen for the estimated flexibility activation when compared to actual flexibility activation in Figure 6.3b. This baseline estimation would have resulted in the DSO over-paying the aggregator during this day, as the estimated flexibility activation is higher than the actual. The overall performance of the method can, however, not be determined by the fraction of the testing period illustrated in the figures.

The performance metrics given in Table 6.1 were thereby calculated to assess the overall performance of the method. Before the recursive strategy was implemented, MLR appeared to have satisfactory results with high R^2 values and low MAE and RMSE values. Implementing the recursive strategy had, however, a negative effect on all the performance metrics. In the training set, R^2 decreased from 98.43 % before recursive to 58.44 % after recursive, implying that only 58.44 % of the variation in the actual baseline could be described with the chosen explanatory variables. The testing set R^2 value decreased from 97.60 % to -92.34 % before and after implementing the recursive strategy. However, the performance metrics calculated before recursive consider the actual baseline values in the same manner as ANN before recursive and are therefore expected to be more accurate.

In addition, the high increase in MAE and RMSE after implementing the strategy compared to before indicates that this strategy may not be suited for MLR. In the test set after recursive, MAE has a value of 15.17 kW, indicating a rather high inaccuracy. Compared to before recursive, where MAE had a value of 1.260 kW in the testing set, there is an increase of 1 103 %.

RMSE has a value of 16.99 kW in testing after recursive, suggesting that the average difference between actual and estimated baseline is around 17 kW. This metric had a value of 1.898 kW before implementing the strategy, resulting in an increase of 795.2 % after recursive. Given that the average actual baseline value is 29.63 kW, the RMSE value after recursive could be considered relatively high. Therefore, it may be unlikely that any DSO would prefer this method to validate flexibility in the settlement process. This inaccurate baseline estimation would lead to an inaccurate validation of flexibility, subsequently making the DSO either over- or under-pay the aggregator significantly in the settlement process. An inaccurate validation may negatively affect all DR event participants, which is not desirable.

The poor performance of the MLR method with recursive strategy can be explained by the regression coefficients shown in Table 5.1. As seen in the table, consumption m-1 is the most significant variable in the method. When this variable is replaced with inaccurate predictions in the recursive strategy, the performance will be negatively affected, resulting in inaccurate baseline estimations. The strategy does hence not appear to perform well in MLR.

Out of the two regression methods, ANN appears to have the most accurate estimation when implementing the recursive strategy. This method has lower values of both MAE and RMSE in training and testing compared to MLR. In addition, the method has higher R^2 values. The performance metrics suggest that this method might be more suited for implementing the recursive strategy. MLR also have higher percent error in all the examined rows in Table 6.2, expect from row 100. On the other hand, neither ANN nor MLR show promising results after implementing the strategy. Even though ANN has better results than MLR, it may not be accurate enough for the DSO in a settlement process.

The recursive strategy was implemented to improve the results from scenario 2 and make the simulations reflect a real-world scenario to a higher degree compared to scenario 1. However, after implementing the strategy, the results do not appear to be improved when comparing the performance metrics given for scenario 2 in Table 5.3 and after recursive in Table 6.1. The R^2 values in training and testing decrease, while MAE and RMSE increase for both ANN and MLR. For instance, R^2 decreases by nearly 50 % in the testing set for ANN, which could be considered a rather remarkable decrease. Therefore, the recursive strategy does not seem to improve the results from scenario 2. On the other hand, the implementation of the strategy does make the simulations reflect a real-world scenario to a higher degree than scenario 1, as only data available to the DSO is used.

The recursive strategy was implemented on the whole data set, containing values for a whole year. However, a DR event will not last continuously over a whole year. It may be reasonable to assume that the strategy could perform better in a smaller time period. On the other hand, the percent errors calculated for ANN in Table 6.2 imply otherwise. As row 100 000 has a lower percent error than row 1 000, it is indicated that the errors with the recursive strategy may not accumulate over time.

Even though neither methods resulted in accurate baseline estimations with recursive strategy, it can be observed from Figure 6.2 and 6.4 that both baseline estimations follow the overall trend of the actual baseline. This may imply that the recursive strategy does not accumulate errors as initially assumed in section 5.3. The calculated percent errors, given in Table 6.2, also demonstrate this, as the errors for both methods are decreasing in the 100 000th row. Both the figures and the table indicate that regardless of the methods with recursive strategy being inaccurate, the baseline estimations follow a similar trend to the actual baseline as the errors do not accumulate.

A possible explanation for the error not accumulating might be that the only explanatory variables replaced in the recursive strategy were m-1 and m-2. The remaining variables, which also correlate with the dependent variable given in Table 4.1, stayed the same. Explanatory variables such as temperature and hour, which have high correlations with the dependent variable, may therefore affect the methods to keep the overall trend of the baseline. The correlation between consumption and temperature can also be seen in the heat map illustrated in Figure 4.6, which illustrates a parallel between the increase in consumption and temperature.

However, the correlation between the dependent variable and the remaining explanatory variables does not appear to be strong enough to make the baseline estimation more accurate when replacing m-1 and m-2 in the recursive strategy. For MLR, it can also be seen in Table 5.1 that the remaining explanatory variables do not have large enough regression coefficients compared to m-1 to impact the accuracy greatly.

As the strategy continuously replaced lagged consumption values with estimated baseline values, the methods became more realistic. However, an assumption was made when implementing the strategy, making the starting points the values of the actual baseline. These are values that the DSO will not have in a real-world scenario, making this implementation of the strategy not completely realistic. On the other hand, the starting points of the recursive strategy need to be accurate enough to avoid large errors from the start, as it was assumed that these errors would accumulate throughout the strategy.

Alternative starting points in a real-world scenario could be consumption values right before a DR event or values from a control group. The DSO will have access to substation data and may utilize the consumption data right before the start of a DR event as the starting points for the recursive strategy. However, it is of importance that the start of the DR event occurs when planned with this alternative. For example, if the aggregator activates the flexibility before the scheduled time, the DSO will start the recursive strategy with values for consumption with activated flexibility. This would have led to an imprecise start of the strategy, possibly leading to inaccurate baseline estimations. A solution could be for the DSO to start the recursive strategy with values a certain time period before the DR event was scheduled to avoid using consumption with activated flexibility values as starting points.

Values from a control group may also be alternative starting points in the recursive strategy. This will, however, require that the DSO has access to consumption data for a substation with similar consumption patterns, which may be challenging to obtain as substations have varying patterns. Regardless of the DSO getting access to such control group data, it will contain some errors as substations are rarely similar in terms of consumption patterns. Granted that control group consumption is available to the DSO, the DSO could prefer to rather use a control group method to estimate the baseline, as presented in section 3.2.4, instead of using these values as starting points in the recursive strategy.

7.2 Rectifying strategy

By comparing the performance metrics tables for recursive and rectifying, Table 6.1 and 6.3, it can be observed that the rectifying strategy, when using the same method for both models, is always better than the recursive strategy. This was expected as the rectifying strategy was implemented to decrease the error. The accuracy of the baseline estimation does, however not improve to the desired extent

The exact decrease in MAE from the recursive to rectifying strategy is given in Table 6.4. Despite the performance metrics for the rectifying results not giving promising results, there does seem to be an increase in accuracy after the rectifying strategy is implemented, especially for the test sets. By implementing the rectifying strategy, the MLR+MLR combination has the most significant MAE decrease at 24.79 %, followed by MLR+ANN with a 24.39 % decrease in MAE. However, the results are not as pleasing for the test sets with ANN as the first model. For ANN+ANN, the MAE decrease by only 7.955 %, and for ANN+MLR the MAE is increasing by 9.590 % from recursive to rectifying. This could indicate that MLR is a better fit for a recursive strategy when a rectifying strategy is being implemented. However, the recursive results with ANN already had higher accuracy than MLR, and hence the room for improvement with residual error was smaller.

From the four different rectifying combinations, ANN as the regression method for both model 1 and 2 gave the highest accuracy for all performance metrics. This could be expected as ANN have been predicting the most accurate or at least equivalent to the most accurate results for the former simulations. The second most accurate rectifying strategy was with ANN+MLR, followed by MLR+MLR and MLR+ANN. The results appear to have the highest accuracy when ANN is used as the first model. This could be related to model 1 having a greater impact on the strategy compared to model 2. When MLR is used as model 1, the estimations are less accurate.

Figure 6.5 presents two of the four combinations with ANN used in model 1, while Figure 6.6 shows the two remaining combinations with MLR used in model 1. It can be observed that the baseline estimation intersects with the actual baseline several times when using ANN for model 1. In contrast, the baseline does not intersect with the actual baseline when using MLR for model 1. This could further stress the impact of the first model. In the same manner as the baseline estimation for both Figure 6.5 and Figure 6.6, the estimated flexibility activation does also intersect with the actual baseline at several points when using ANN for model 1, compared to when using MLR for model 1.

When comparing Figure 6.5 and Figure 6.6, a parallel between the baseline estimation and the dominance of the first model could be drawn. Examining the subplots in each figure could reveal that the estimated curves with the same first model have similar trends, regardless of the regression method in the second model. This dominance of the first model is demonstrated graphically in these figures, compared to Table 6.3, where the same observation could be obtained from a numerical perspective.

The small difference in the results before and after rectifying could be related to Table 6.5, showing the correlation between the dependent and explanatory variables for model 2. The explanatory variables do not have a strong correlation with the dependent variable. A weak correlation leads to a poorly trained model, which further would not estimate as well. This could be a reason for the rectifying strategy not performing as well as assumed.

The calculated performance metrics in Table 6.3 are decent at best. The lowest MAE for the training set for ANN is at 7.1 kW per minute. If this was to be used by a DSO, they would have estimated the baseline wrong by 3 731 760 kW for a whole year. This value does however not take into account when the error is over- or underestimated. From the annual kW error, it can be assumed that the rectifying baseline estimations would not be good enough for a DSO to use in a real-world scenario.

By looking at the correlation values for model 1, Table 4.1, and model 2, Table 6.5, it can be observed that the values are lower for model 2. This indicates that the explanatory variables do not correlate strongly with the error, but rather with the consumption. A potential improvement could possibly have been to implement other explanatory variables for model 2. This could have increased the correlation between the explanatory and dependent variables. These new explanatory variables could, however only consist of values that a DSO has available in a real-world scenario.

Other methods than the ones implemented in this thesis could have been used to examine any difference in the results. It could have been possible to use a method with low variance and a high bias for the first model. A low variance would provide a robust method. For the second model, a new method with a high variance and low bias could have been introduced. A low bias would give a highly flexible method. This could optimize the rectifying strategy further and create an ideal compromise of the bias-variance trade-off. This could have potentially led to a more accurately estimated baseline, being more favourable to the DSO.

On the other hand, as MLR and ANN are both well-recognised regression methods within the field of machine learning, it could be argued that another strategy should have been implemented rather than focusing on a new method. A new strategy could have been implemented in order to reach all the high-frequency points of the baseline. This could have improved the current baseline estimations from the rectifying strategy.

7.3 Different frequency

To understand the effect of data frequency on the accuracy of baseline estimations, data with 5-minute frequency was tested for both regression methods. When comparing the estimated baseline and estimated flexibility activation for ANN with 5-minute frequency, illustrated in Figure 6.7, to ANN with 1-minute, shown in Figure 6.5, the change in frequency does not appear to have affected the accuracy greatly. A similar effect can be observed for MLR with different frequencies, shown for 5-minute in Figure 6.8 and 1-minute in Figure 6.6. However, these figures only display a fragment of the testing period, and the effect of different frequencies can not be determined from these.

The calculated performance metrics for 5-minute frequency, given in Table 6.6, give a more holistic view of the results of each regression method. The table includes results for before recursive, after recursive and after rectifying, with test and train values. Before implementing the recursive strategy, the ANN method with 5-minute frequency appears to be less accurate than ANN with 1-minute frequency. Performance metrics for ANN with 1-minute frequency are given in Table 6.1. The R^2 values are lower for 5-minute, in addition to the MAE and RMSE values being higher, indicating more inaccurate results. This could be due to m-1 and m-2 having a slightly weaker correlation to the dependent variable with 5-minute frequency, as shown in Table 5.5, compared to the correlation for 1-minute frequency, given in Table 4.1.

After implementing the recursive strategy in the training set, ANN with 5-minute frequency seems to have a slightly better result, according to the performance metrics, compared to 1-minute ANN. The two simulations are, however, quite similar for the testing set after implementing the recursive strategy and for both training and testing after implementing the rectifying strategy.

The correlation values calculated for the second model of ANN with 5-minute frequency, presented in Table 6.7, show that all explanatory variables, except for month, have a positive correlation with the dependent variable, i.e. error. This is a high number of positive correlating variables compared to the correlation values for ANN with 1-minute frequency, given in Table 6.5, where only half of the correlations are positive. Positive correlations lead to more accurate estimations as the dependent and explanatory variables decrease and increase simultaneously. Therefore, it would have been reasonable to assume a greater improvement after rectifying for the 5-minute ANN. However, as the results after rectifying for the two frequencies are almost similar, this does not seem to be the case.

Before implementing the recursive strategy, the MLR method with 5-minute frequency appears to be less accurate than MLR with 1-minute. This can be seen from the performance metrics for 5-minute frequency, given in Table 6.6, and for 1-minute frequency, given in Table 6.1. After implementing the recursive strategy, the performance metrics for training are similar for the two frequency simulations. However, MLR with 5-minute frequency seems to be more accurate in the testing set compared with 1-minute MLR. The R^2 value is negative for both frequencies, with a value of -69.22 % with 5-minute and -92.34 % with 1-minute, implying some improvement. MAE decreases by 6.988 % in testing from 1-minute to 5-minute, and RMSE decreases by 6.121 % from 1-minute to 5-minute, indicating a more accurate method.

After implementing the rectifying strategy with MLR, the two frequency experiments have relatively similar performance metric values in the training set. The simulation with 5-minute frequency does, however, appear to be somewhat more accurate in the testing set. With 1-minute frequency, R^2 has a value of -18.36 %, as seen in Table 6.3, and R^2 has a value of -0.560 % with 5-minute frequency. Even though neither is a desired value of R^2, the simulation with 5-minute frequency suggests a more accurate result. In addition, MAE and RMSE have lower values in the 5-minute simulation.

The improvement in performance from 1 to 5-minute frequency might be explained with the correlation values for model 2 in the rectifying strategy with 5-minute, given in Table 6.7. All of the explanatory variables have a positive correlation with the dependent variable in MLR, the error in this case. This could have affected the second model to estimate the error more accurately, leading to a more accurate baseline estimation after implementing this strategy. In comparison, the correlation values calculated for MLR with 1-minute frequency were mostly negative, as shown in Table 6.5. This indicates that when the dependent variable increases, most of the explanatory variables will decrease, which could have affected the second model to have less accurate estimations in the simulation with 1-minute frequency.

The frequency appears to have little effect on the accuracy of baseline estimation with ANN after implementing the rectifying strategy, according to the performance metrics. Therefore, neither of the two frequencies might be more favorable for the DSO, as the results are almost similar. As both frequencies resulted in an average difference between the estimated and actual baseline of around 9 kW, a DSO might not prefer either of the simulations to validate flexibility in the settlement process. A MAE value around 7 kW for both simulations also suggests that neither can be considered to be extremely accurate.

As the results improved somewhat from 1-minute to 5-minute frequency for MLR after implementing the rectifying strategy, a DSO might prefer the 5-minute frequency for this regression method. Even though the simulation with 5-minute improved the estimation, the results may not be considered very accurate for this simulation either. The RMSE value of around 12 kW suggests that the average difference between the actual and estimated baseline is quite high, considering that the average actual baseline value is 29.63 kW. Such a high average difference might not be preferable for a DSO in the settlement process.

The MAE value of around 10 kW also indicates that the simulation can not be considered accurate. In a settlement process, where accuracy is an essential factor for the DSO in order to not over- or under-pay, neither of the frequencies included in this thesis seems to be favorable. Out of the two regression methods, ANN both with 1 and 5-minute frequency appears to give more accurate results compared to MLR.

7.4 Individual houses

The MAE and RMSE for individual houses, given in Table 6.8, are lower compared to the results after rectifying at substation level in Table 6.3 for 1-minute and Table 6.6 for 5-minute frequency. The RMSE and MAE for the individual houses range between 0.888 and 2.508 kW. For the 1- and 5-minute frequency after rectifying, the MAE range between 5.611 and 10.35 kW and RMSE between 7.574 and 13.33 kW. The low values for the individual households could be misleading as they have a lower consumption compared to the consumption on a substation level, where the residential consumption of several households is aggregated. The low values result in a smaller difference between the estimated and actual baseline, the two main input variables to calculate the MAE and RMSE.

This could however have been avoided if the estimated and actual baseline had been scaled. By scaling all values, the data would have been transformed to fit within a specific scale such as 0 to 1. This could have made the performance metrics for the different frequencies easier to compare, as all results would have the same weightage when calculating the difference between the data points.

The input data for the individual households are 1/60th of the input data for a 1-minute frequency. The reduced input data is due to the frequency reduction from 1-minute to 60-minute, as this is the resolution available for the DSO at individual households through Elhub. This could have been a factoring element for the low R^2 value in the individual households, ranging from 43.87 to -48.77 % for ANN and 38.33 to -141.9 % for MLR. In comparison, the R^2 value range from 43.61 to 84.22 % for ANN and -18.36 to 55.52 % with MLR for the 1-minute frequency.

Both MLR and ANN are dependent on training the data with a reliable amount of input data in order to estimate a satisfactory baseline. However, with a training data of only 2 622 input rows for the individual households, it could have been expected that the MLR would have given more accurate results, as the ANN is built for larger and more complex data, as mentioned in chapter 5. On the contrary, ANN is characterized as a more complex and robust method compared to MLR. This could have been the reasoning behind higher R^2 results with the ANN as the regression method for both model 1 and 2 in the rectifying strategy.

To get a more complete view of the baseline estimation accuracy for the individual households, Figure 6.9, Figure 6.11 and Figure 6.13 can be examined. From the figures, it could be observed that the baseline estimations are not following the actual baselines to a large degree. Further, by analyzing Figure 6.10, 6.12 and 6.14, similar observations could be attained, where the estimated and actual flexibility activation curves do not coincide. These graphical representations could be of help in order to not miss-understand the low MAE and RMSE values as adequate performance metrics.

A factor that could have played a role in the low accuracy of the baseline estimation for the individual household is the consumption profile for the individual households compared to a substation level. At a residential level, the consumption pattern would fluctuate to a greater extent from one hour to the next compared to the aggregated load, which looks at the consumption on a higher level. Here, a consumer action would not affect the substation aggregated consumption to the same extent as it does on household level. With such different consumption patterns, the household baseline estimation has a more challenging time training the data than on substation level, where the training and testing data is more stable.

Individual houses could possibly require more explanatory variables custom-made to their specific daily routine. The number of residents, work pattern, vacation, transportation, etc. could be included at the household level in order to increase the performance metrics, further resulting in a baseline estimation with higher accuracy.

The correlation between the explanatory variables and the dependent variable for model 2 is low, as seen in Table 6.9, which provides statistical evidence of a negative relationship for several of the variables. Only 16 of 42 values presented in the table are positive. With more than half of the correlation variables being negative, the table indicates an inaccurate error estimation, as there is an inverse relationship between the parameters. Most of the explanatory variables are making the dependent variable function in the opposite direction of the desired results.

Compared to the forecasting at substation level, it was however not necessary to implement the recursive strategy during all time stamps on the individual household level. As the individual households do not have a constant flexibility activation, the recursive strategy could have only been applied during the DR-event, where flexibility is activated. This could potentially decrease the computation time of the individual houses further. In addition, it may have resulted in more accurate results after the recursive strategy, as each new loop would have actual baseline values as starting points. The same could have been done at substation level if flexibility was not constantly activated.

As suggested in section 5.6, the DSO could have used the information received from the aggregator and directly from the individual household to validate the amount of flexibility. The baseline estimation at both substation and household level could also have been used to validate flexibility, giving the DSO valuable information in the settlement process. However, as the baseline estimation of the individual household was somewhat inaccurate, it is questionable whether this information would be helpful to the DSO in a real-world scenario.

7.5 Relevant factors and subjects

As discussed in the sections above, the simulation experiments might not have given as accurate results as a DSO may prefer in a settlement process when validating flexibility. However, several factors may have contributed to these insufficiently estimated baselines. The amount of data used in the simulation experiments is one example. The size of the artificially created substation is another factor that could have affected the results. Factors such as these will be discussed in this section, in addition to other subjects that may be relevant for the DSO in the settlement process when validating activated flexibility.

When estimating the baselines, 70 % of the data were used for training, while the remaining 30 % were testing data. January until mid-September thus constituted the training set, while the rest of September to end of December was the testing set. As seen in Figure 4.3, there are two main consumption patterns for the months. May, June, July, August and September have higher overall consumption with only one significant peak throughout the day, while the remaining months have overall lower consumption with two peaks. A larger portion of the testing months has consumption patterns with two peaks and lower consumption.

However, only half of the training months have a similar pattern to the testing months. Figure 4.4 shows the average daily consumption per hour of all the months with two peaks and lower consumption. The months with blue shades are included in the training set, while the curves with green shades represent the months for the testing set. As seen in the figure, there are four months with blue shades and three months with green shades. Therefore, there might not have been enough months with a similar pattern to the testing months when training the methods, which may be one of the reasons for the inaccurately estimated baselines. A possible solution could have been to include a whole year of data for training and testing the simulations the following year. This would ensure training of all the months and their different patterns.

Another factor that may have affected the baseline estimation results is the size of the artificially created substation. The data used in the experiments contained consumption data for 22 households. In section 3.1, it is however mentioned that a low voltage feeder, which carries energy to a substation, includes on average about 50 households. As the data used in this thesis is under half the size, it may have been more volatile than a regular low voltage substation. More volatile data is more difficult to accurately estimate, as discovered in the literature review in section 3.2, which may have affected the baseline estimation results in this thesis.

The literature review found that the number of consumers connected to a substation is one of the most prominent indicators of forecast accuracy, as the relative error increases exponentially with the decrease of the substation size. Hence, it becomes more challenging to estimate baselines for smaller substations accurately. Also, the artificially created substation did not include different components such as streetlights and other street furniture. These components, in addition to more households, could have contributed to more stable data.

A significant subject to discuss when testing different baseline estimation methods is the importance of implementing a simple and agreeable method for the DSO and aggregator. Both parties have to understand the method if it is to be used for validation of activated flexibility in the settlement process. It will also be important that the training of the method has low computational time. If connection points to the substation are changed, removed or added, the method would have to be trained again. In such a case, it may be beneficial for both participants that the training of the method with the new data is computed relatively quick.

As MLR is a more straightforward method, it may be more understandable for the DSO and aggregator to implement and evaluate. However, as the baseline estimation results showed, this method does not always provide the most accurate baseline estimations. ANN might be more accurate, though also more advanced in terms of implementation. Ideally, there would be a balance between simplicity and accuracy for both participants, as one factor may not be more important than the other.

Chapter 3 mentioned several load forecasting methods, of which some could have been implemented instead of ANN and MLR to estimate baselines. An averaging method could have been an alternative. These methods are simple to implement and could hence have been accepted by both the DSO and aggregator. Nevertheless, these methods can lead to large errors considering that aggregated residential load patterns tend to have stronger heterogeneity compared to industrial and commercial loads. It is also uncertain if any of the methods mentioned in the chapter would have been able to capture the high fluctuating frequency of the actual baseline when implementing the strategies. Since both ANN and MLR are commonly used methods for baseline estimation, and neither of these could estimate the high frequency of the actual baseline, it is debatable if any of the other methods would have achieved more accurate estimations.

The aggregator role can be reflected on, as there is a possibility to include DR activation directly through the DSO, rather than including an aggregator as an intermediary. This would have included a contractual agreement with the end-user, allowing the DSO to control their loads. This could have reduced the transaction costs and complexity for the DSO. However, such an approach would have made it more difficult to provide efficient aggregation of flexibility. Introducing an aggregator would ensure fair pricing and efficient aggregation in a market platform. Such a market-based approach could, on the other hand, possibly have led to a more complex system with numerous interactions and remodelling of the current electricity network. However, by including the aggregator, the DSO could focus on the flexibility request rather than the delivery responsibility.

Consumer privacy and GDPR rules are fundamental concerns for the implementation of DR solutions in the flexibility market. Project stakeholders agree on respecting the data privacy of individual customers. However, access to data, such as AMS measurements, is of the essence to adequate forecasting and flexibility procurement. Contracts signed by end-users participating in DR programs usually contain several long documents. These would include a detailed description of all service operations to be performed with the customer data. The complexity of such paperwork is very new and often too high for the average resident. A possible option could be to simplify the forms without lowering the level of privacy.

In order for the DSO to estimate the baseline as accurately as possible, there should be implemented some regulation in regard to the LFM, in particular which parties the aggregator or consumer can sell flexibility to and which third parties have insight to the flexibility trade. For a DSO, it is essential to calculate all admissible days correctly for a baseline estimation. Hence, they would need insight into the amount of flexibility the consumer sell to other participants in the LFM. However, it will be complicated for all players in a market to agree on such an agreement that is also in concurrence with the law regarding GDPR.

The rebound effect explained in section 3.4 is another challenge regarding the baseline estimation, which should be addressed in a future aggregator framework and contracts to prevent any risk of a negative financial impact on the end-user. The DSO could either incentives or punish the flexibility supplier, based on their change in consumption after an up or down-regulation in a DR-vent. In a worst-case scenario, such as a complete rebound after load curtailment in every DR event, the rebound would be higher than the estimated baseline, punishing the end-user. For the aggregator however, the impact would depend on the portion of such customers they have in their portfolio. This impact could hence be more distinct and noticeable for smaller and emerging flexibility suppliers.

On the contrary, if a compensation mechanism for the rebound effect has to be implemented in a contract agreement, it could encourage aggregators to only interact with non-rebounding assets. This would, however, go against the current European goals to actively involve consumers in the provision of ancillary services such as flexibility in the LFM. The neglect of specific ancillaries or customers could further decrease the use of the potential flexibility available on the demand side.

The historical data used in the baseline estimation today is usually based on data from the prior 1 to 3 years. However, the electricity consumption profile is rapidly changing with the high integration of electrical loads in the market. In the upcoming years, it will hence not be advantageous to use the historical data from the late 2010s, as these consumption patters would not resemble the modern electricity market. In the upcoming years, the flexibility market might be established and the consumption data from the 1 to 3 prior years would hence include flexibility activation as well. In this scenario, baseline estimations such as the ones attained in this thesis would not be adequate. The baseline estimation simulations used in this thesis include historical data of consumption without flexibility. However, the historical data without flexibility from the late 2010s would not resemblance the consumption pattern of an established flexibility market, as the consumption would assumable be higher in the future.

8 Conclusion and further work

The aim of this thesis was to research how the DSO could validate demand-side flexibility at substation level activated by the aggregator in the settlement process, in addition to how validation could be implemented in a realistic scenario using consumption data available to the DSO. Load forecasting methods for baseline estimation can be used for validating activated flexibility. Two regression methods, ANN and MLR, were implemented to estimate baselines with artificially created substation data in order to validate the demand-side activated flexibility in this thesis.

To validate activated flexibility with consumption data available to the DSO, a recursive strategy was applied to both regression methods. ANN provided a more accurate baseline estimation with this strategy compared to MLR. Both methods followed the trend of the actual baseline, but they were not able to capture the high fluctuating frequency of the actual baseline. As neither method showed promising results after implementing the strategy, the methods are not considered accurate enough for a DSO to validate flexibility activation in the settlement process.

A rectifying strategy was implemented to improve the estimations from the recursive strategy. The new strategy did improve the baseline estimation results to some degree. The accuracy of the methods after implementing this strategy is moderate. There is scope for future improvement through the use of appropriate explanatory variables and advanced machine learning algorithms, among other factors.

The artificially created substation data was tested with both 1-minute and 5-minute frequency to research the effect of data frequency on the baseline estimation performance. The frequency adjustment had little effect on the accuracy of the baseline estimations with ANN. 5-minute frequency provided some improvement to the results compared to 1-minute for MLR after implementing the rectifying strategy. Given that the frequency adjustment had little or some effect on the accuracy, neither of the two frequencies will be more favorable for the DSO when implementing the regression methods. Of the two methods, ANN had more accurate results compared to MLR.

Baseline estimations on individual households were included in the thesis to further benefit the DSO in validating flexibility on the demand side by using available data. The estimated baselines of individual households did not capture the high fluctuating frequency of the actual baseline as the consumption patterns are more volatile at residential level, and residential load patterns can not maintain the same stable level as an artificial substation. The contribution from the individual house estimations will hence be negligible to the DSO in the settlement process.

The complexity of validation using baseline estimation has been established in this thesis, where even the most common methods fail due to the nature of the problem. Further improvements should be implemented to the strategies in order to obtain accurate estimations with data available to the DSO in a real-world scenario.

The suggestions for possible improvements of the simulation experiments presented in this thesis and further work are listed as follows:

- Use real substation data to test the regression methods, instead of artificially created substation data. In addition, test the methods on larger substation data, for example a substation of 50 households.
- Include a whole year of data for training of the methods and test the simulations the following year to ensure training of all the months and their different patterns.
- Implement the recursive strategy for the duration of a DR event to examine how the strategy performs in a smaller time frame compared to over a whole year, as conducted in this thesis.
- Implement other methods as the second model in the rectifying strategy. Also, test different explanatory variables in the second model to optimize the rectifying improvement. To improve the results further, the rectifying strategy can be implemented multiple times.
- Test several method combinations in the rectifying strategy to optimize the performance improvement further and create an ideal compromise of the bias-variance trade-off. Model 1 should have low variance and high bias, while model 2 should have high variance and low bias.
- Analyze what factors can cause the methods to capture the high fluctuating frequency of the actual baseline better, instead of only estimating the overall trend.
- Different strategies than the two presented in this thesis can be explored to analyze their effect on baseline estimation accuracy and their ability to capture high frequency.
- More advanced machine learning methods can be explored to research their ability to capture the high fluctuating frequency of the actual baseline.
- Other topics that can be further researched include rebound effect and baseline manipulation.

References

[1] United Nations Climate Change. *The Paris Agreement*. 2021. URL: https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement (visited on Dec. 5, 2021).

[2] The Norwegian Government. *Klimaendringer og norsk klimapolitikk*. 2021. URL: https://www.regjeringen.no/no/tema/klima-og-miljo/innsiktsartikler-klima-miljo/klimaendringer-og-norsk-klimapolitikk/id2636812/ (visited on Dec. 5, 2021).

[3] Viktorija Dudjak et al. *Impact of Local Energy Markets on the Distribution Systems: A Comprehensive Review*. 2018.

[4] Shuangyuan Wang et al. "Regional nonintrusive load monitoring for low voltage substations and distributed energy resources". In: *Applied Energy* 260 (2020).

[5] Jayaprakash Rajasekharan. *Project: Flexibility aggregation from distributed energy sources*. 2021. URL: https://arkiv.iel.ntnu.no/fagvalg/prosjekt-detaljer.php?Sortering=1045&fbclid=IwAR3wPQh2n6YTtpWRJ5WzF_rGEdcUFzc0yy1wH3XpXGQyiNer71fdS8SXA7I (visited on Dec. 7, 2021).

[6] Ioannis Lampropoulos et al. "A system perspective to the deployment of flexibility through aggregator companies in the Netherlands". In: *Energy Policy* 118 (2018), pp. 534–551.

[7] Xinkang Wang et al. "Customer Baseline Load Bias Estimation Method of Incentive-Based Demand Response Based on CONTROL Group Matching". In: *2nd IEEE Conference on Energy Internet and Energy System Integration (EI2)*. 2018, pp. 1–6.

[8] Stephen Haben et al. "Review of low voltage load forecasting: Methods, applications, and recommendations". In: *Applied Energy* 304 (2021).

[9] Rik Fonteijn et al. "Baselining Flexibility from PV on the DSO-Aggregator Interface". In: *Applied Sciences Switzerland* 11 (2021), pp. 1–25.

[10] Xiaolong Jin, Qiuwei Wu, and Hongjie Jia. "Local flexibility markets: Literature review on concepts, models and clearing methods". In: *Applied Energy* 261 (2020).

[11] Luca Mendicino et al. "DSO Flexibility Market Framework for Renewable Energy Community of Nanogrids". In: *Energies* 14.12 (2021).

[12] Simran Jit Kaur Sandhu and Marthe Vågen. *Load Forecasting and Monitoring for Flexibility Aggregation from Distributed Energy Sources*. Project report in TET4510. Department of Electrical Power Engineering, NTNU – Norwegian University of Science and Technology, 2021.

[13] Stian Hackett et al. "Market design options for procurement of flexibility". In: *Nordic Energy Research* (2021).

[14] Kari Dalen et al. *Statnett: Distributed balancing of the power grid - Results from the eFleks pilot in the mFRR-market 2019/2020*. 2021.

[15] Nordic Council of Ministers. *Demand side flexibility in the Nordic electricity market - From a Distribution System Operator Perspective*. 2017.

[16] Pol Olivella-Rosell et al. "Local Flexibility Market Design for Aggregators Providing Multiple Flexibility Services at Distribution Network Level". In: *Energies* 11.4 (2018).

[17] Pau Lloret et al. *EU Horizon 2020 Research and Innovation programme: Overall INVADE architecture*. 2017.

[18] CEN-CENELEC-ETSI. *Smart Grid Coordination Group - Smart Grid Reference Architecture*. 2012.

[19] Chenghua Zhang et al. "Peer-to-Peer energy trading in a Microgrid". In: *Applied Energy* 220 (2018), pp. 1–12.

[20] Abouzar Estebsari et al. "A SGAM-Based Test Platform to Develop a Scheme for Wide Area Measurement-Free Monitoring of Smart Grids under High PV Penetration". In: *Energies* 12 (2019).

[21] Alexandre Lucas et al. "Load Flexibility Forecast for DR Using Non-Intrusive Load Monitoring in the Residential Sector". In: *Energies* 12.14 (2019).

[22] ENEDIS. *InterFlex: Interactions between automated energy systems and Flexibilities brought by energy market players*. 2021. URL: https://cordis.europa.eu/project/id/731289 (visited on Mar. 11, 2022).

[23] FLEXCoop. *Demand response for energy cooperatives*. 2019. URL: http://www.flexcoop.eu/ (visited on Apr. 15, 2022).

[24] Computer Engineering Spa Italy. *eDREAM - enabling new Demand REsponse Advanced, Market oriented and Secure technologies, solutions and business models*. 2022. URL: https://cordis.europa.eu/project/id/774478 (visited on May 9, 2022).

[25] National Center for Research and Technological Development Greece. *DRIMPAC: Unified DR interoperability framework enabling market participation of active energy consumers*. 2022. URL: https://cordis.europa.eu/project/id/768559 (visited on May 9, 2022).

[26] Jelena Ponocko and Jovica V. Milanovic. "Forecasting Demand Flexibility of Aggregated Residential Load Using Smart Meter Data". In: *IEEE Transactions on Power Systems* 33.5 (2018), pp. 5446–5455.

[27] DNVGL. *Baseline Methodology Assessment - Energy Networks Association*. 2020.

[28] Leslie Hatton, Philippe Charpentier, and Eric Matzner-Løber. "Statistical Estimation of the Residential Baseline". In: *IEEE Transactions on Power Systems* 31.3 (2016), pp. 1752–1759.

[29] A. Gabaldón et al. "Improvement of customer baselines for the evaluation of demand response through the use of physically-based load models". In: *Utilities Policy* 70 (2021).

[30] Elhub AS c/o Statnett SF. *About Elhub*. 2022. URL: https://elhub.no/en/about-elhub/ (visited on Mar. 10, 2022).

[31] Maria Georges. *Elhub AS*. Personal communication. E-mail. Mar. 8, 2022.

[32] Stephen Haben et al. "Short term load forecasting and the effect of temperature at the low voltage level". In: *International Journal of Forecasting* 35.4 (2019), pp. 1469–1484.

[33] Tao Hong and Shu Fan. "Probabilistic electric load forecasting: A tutorial review". In: *International Journal of Forecasting* 32.3 (2016), pp. 914–938.

[34] Tao Hong et al. “Energy Forecasting: A Review and Outlook”. In: *IEEE Open Access Journal of Power and Energy* 7 (2020), pp. 376–388.

[35] Meritxell Gómez-Omella et al. “k-Nearest patterns for electrical demand forecasting in residential and small commercial buildings”. In: *Energy and Buildings* 253 (2021).

[36] Carlos Santos Silva. *Técnico Lisboa: Presentation, Energy Services - Feature Selection.*

[37] Christopher J. Bennett, Rodney A. Stewart, and Jun Wei Lu. “Forecasting low voltage distribution network demand profiles using a pattern recognition based expert system”. In: *Energy* 67 (2014), pp. 200–212.

[38] Isaac Kofi Nti et al. “Electricity load forecasting: a systematic review”. In: *Electrical Systems and Inf Technol* 7.13 (2020).

[39] Rachit Srivastava, A.N. Tiwari, and V.K. Giri. “Solar radiation forecasting using MARS, CART, M5, and random forest model: A case study for India”. In: *Heliyon* 5.10 (2019).

[40] Zach. *RMSE vs. R-Squared: Which Metric Should You Use?* 2021. URL: `https://www.statology.org/rmse-vs-r-squared/` (visited on Dec. 8, 2021).

[41] AskPython. *RMSE – Root Mean Square Error in Python.* URL: `https://www.askpython.com/python/examples/rmse-root-mean-square-error` (visited on Dec. 8, 2021).

[42] Yang Weng, Jiafan Yu, and Ram Rajagopal. “Probabilistic baseline estimation based on load patterns for better residential customer rewards”. In: *International Journal of Electrical Power & Energy Systems* 100 (2018), pp. 508–516.

[43] Shubhasmita Pati, Satish J. Ranade, and Olga Lavrova. “Methodologies for Customer Baseline Load Estimation and their Implications”. In: *IEEE Texas Power and Energy Conference (TPEC)* (2020), pp. 1–5.

[44] AskPython. *Coefficient of Determination – R squared value in Python.* URL: `https://www.askpython.com/python/coefficient-of-determination` (visited on Dec. 8, 2021).

[45] Javad Jazaeri et al. “Baseline methodologies for small scale residential demand response”. In: *IEEE Innovative Smart Grid Technologies - Asia (ISGT-Asia)* (2016), pp. 747–752.

[46] ITBodhi. *Bias and Variance Trade off.* 2020. URL: `https://medium.com/@itbodhi/bias-and-variance-trade-off-542b57ac7ff4` (visited on May 25, 2022).

[47] G. B. M. A. Litjens, E. Worrell, and W. G. J. H. M. van Sark. “Assessment of forecasting methods on performance of photovoltaic-battery systems”. In: *Applied Energy* 221 (2018), pp. 358–373.

[48] Nick MacMackin, Lindsay Miller, and Rupp Carriveau. “Modeling and disaggregating hourly effects of weather on sectoral electricity demand”. In: *Energy* 188 (2019).

[49] A. Kipping and E. Trømborg. “Modeling and disaggregating hourly electricity consumption in Norwegian dwellings based on smart meter data”. In: *Energy and Buildings* 118 (2016), pp. 350–369.

[50] Young M. Lee et al. “Applying science and mathematics to big data for smarter buildings”. In: *Annals of the New York Academy of Sciences* 1295 (2013).

[51] Andrei Marinescu et al. "Residential electrical demand forecasting in very small scale: An evaluation of forecasting methods". In: *2nd International Workshop on Software Engineering Challenges for the Smart Grid (SE4SG)* (2013), pp. 25–32.

[52] Gde Dharma Nugraha et al. "Lambda-Based Data Processing Architecture for Two-Level Load Forecasting in Residential Buildings". In: *Energies* 11 (2018).

[53] Barry P. Hayes, Jorn K. Gruber, and Milan Prodanovic. "A Closed-Loop State Estimation Tool for MV Network Monitoring and Operation". In: *IEEE Transactions on Smart Grid* 6.4 (2015), pp. 2116–2125.

[54] Kaushik Katari. *Multiple Linear Regression model using Python: Machine Learning*. 2020. URL: `https://towardsdatascience.com/multiple-linear-regression-model-using-python-machine-learning-d00c78f1172a` (visited on Dec. 1, 2021).

[55] Thierry Zufferey, Sandro Renggli, and Gabriela Hug. "Probabilistic State Forecasting and Optimal Voltage Control in Distribution Grids under Uncertainty". In: *Electric Power Systems Research* 188 (2020).

[56] Jicheng Liu and Yinghuan Li. "Study on environment-concerned short-term load forecasting model for wind power based on feature extraction and tree regression". In: *Journal of Cleaner Production* 264 (2020).

[57] Tairen Chen et al. "Distribution Feeder Level Day-ahead Peak Load Forecasting Methods And Comparative Study". In: *IET Generation, Transmission and Distribution* 12 (2018).

[58] Youlong Yang et al. "Sequential grid approach based support vector regression for short-term electric load forecasting". In: *Applied Energy* 238 (2019), pp. 1010–1021.

[59] Jinran Wu et al. "Support vector regression with asymmetric loss for optimal electric load forecasting". In: *Energy* 223 (2020).

[60] Hui He et al. "Short-term load probabilistic forecasting based on quantile regression convolutional neural network and Epanechnikov kernel density estimation". In: *Energy Reports* 6 (2020), pp. 1550–1556.

[61] Yi Wang et al. "Probabilistic individual load forecasting using pinball loss guided LSTM". In: *Applied Energy* 235 (2019), pp. 10–20.

[62] Yandong Yang et al. "Bayesian Deep Learning-Based Probabilistic Load Forecasting in Smart Grids". In: *IEEE Transactions on Industrial Informatics* 16.7 (2020), pp. 4703–4713.

[63] Wanying Zhang, Yaoyao He, and Shanlin Yang. "Day-ahead load probability density forecasting using monotone composite quantile regression neural network and kernel density estimation". In: *Electric Power Systems Research* 201 (2021).

[64] Can Bikcora, Lennart Verheijen, and Siep Weiland. "Density forecasting of daily electricity demand with ARMA-GARCH, CAViaR, and CARE econometric models". In: *Sustainable Energy, Grids and Networks* 13 (2018), pp. 148–156.

[65] Charalampos Ziras, Carsten Heinrich, and Henrik W. Bindner. "Why baselines are not suited for local flexibility markets". In: *Renewable and Sustainable Energy Reviews* 135 (2021).

[66] Xinan Wang et al. "Residential Customer Baseline Load Estimation Using Stacked Autoencoder With Pseudo-Load Selection". In: *IEEE Journal on Selected Areas in Communications* 38.1 (2020), pp. 61–70.

[67] Kangping Li et al. "A Baseline Load Estimation Approach for Residential Customer based on Load Pattern Clustering". In: *Energy Procedia* 142 (2017). Proceedings of the 9th International Conference on Applied Energy, pp. 2042–2049.

[68] Xiaoyang Zhou et al. "Forecast load impact from demand response resources". In: *IEEE Power and Energy Society General Meeting (PESGM)* (2016), pp. 1–5.

[69] Saeed Mohajeryami, Milad Doostan, and Peter Schwarz. "The impact of Customer Baseline Load (CBL) calculation methods on Peak Time Rebate program offered to residential customers". In: *Electric Power Systems Research* 137 (2016), pp. 59–65.

[70] Yongbao Chen et al. "Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings". In: *Applied Energy* 195 (2017), pp. 659–670.

[71] Jayesh Priolkar, E. S. Sreeraj, and Anita Thakur. "Analysis of Consumer Baseline for Demand Response Implementation: A Case Study". In: *7th International Conference on Signal Processing and Integrated Networks (SPIN)* (2020).

[72] Mingyang Sun et al. "Clustering-Based Residential Baseline Estimation: A Probabilistic Perspective". In: *IEEE Transactions on Smart Grid* 10.6 (2019), pp. 6014–6028.

[73] Yi Zhang et al. "A Cluster-Based Method for Calculating Baselines for Residential Loads". In: *IEEE Transactions on Smart Grid* 7.5 (2016), pp. 2368–2377.

[74] Eunjung Lee et al. "Defining virtual control group to improve customer baseline load calculation of residential demand response". In: *Applied Energy* 250 (2019), pp. 946–958.

[75] Princeton University Library. *Interpreting Regression Output*. URL: `https://dss.princeton.edu/online_help/analysis/interpreting_regression.htm#:~:text=a%20miniscule%20effect.-,Coefficients,the%20direction%20of%20the%20effect.` (visited on May 29, 2022).

[76] Ernesto Aguilar Madrid and Nuno Antonio. "Short-Term Electricity Load Forecasting with Machine Learning". In: *Information* 12.2 (2021).

[77] Jasleen Kaur et al. "Solar power forecasting using ordinary least square based regression algorithms". In: *IEEE Delhi Section Conference (DELCON)*. 2022, pp. 1–6.

[78] Y.-S. Park and S. Lek. "Chapter 7 - Artificial Neural Networks: Multilayer Perceptron for Ecological Modeling". In: *Ecological Model Types*. Vol. 28. Developments in Environmental Modelling. Elsevier, 2016, pp. 123–140.

[79] Rendyk. *Tuning the Hyperparameters and Layers of Neural Network Deep Learning*. 2021. URL: `https://www.analyticsvidhya.com/blog/2021/05/tuning-the-hyperparameters-and-layers-of-neural-network-deep-learning/` (visited on May 21, 2022).

[80] TIBCO. *What is a Neural Network?* URL: `https://www.tibco.com/reference-center/what-is-a-neural-network` (visited on May 28, 2022).

[81] DeepAI. *What is a Hidden Layer?* URL: https://deepai.org/machine-learning-glossary-and-terms/hidden-layer-machine-learning (visited on May 21, 2022).

[82] Christian Versloot. *Relu, sigmoid and tanh- todays most used activation functions*. 2022. URL: https://github.com/christianversloot/machine-learning-articles/blob/main/relu-sigmoid-and-tanh-todays-most-used-activation-functions.md (visited on May 28, 2022).

[83] Emerson Rodolfo Abraham et al. *Time Series Prediction with Artificial Neural Networks: An Analysis Using Brazilian Soybean Production*. 2020.

[84] Jason Brownlee. *Difference Between a Batch and an Epoch in a Neural Network*. 2019. URL: https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/ (visited on May 21, 2022).

[85] Neha Seth. *How does Backward Propagation Work in Neural Networks?* 2021. URL: https://www.analyticsvidhya.com/blog/2021/06/how-does-backward-propagation-work-in-neural-networks/ (visited on May 28, 2022).

[86] UC Business Analytics R Programming Guide. *Regression Artificial Neural Network*. URL: http://uc-r.github.io/ann_regression (visited on May 28, 2022).

[87] Ivar Thokle Hovden. *Optimizing Artificial Neural Network Hyperparameters and Architecture*. 2019.

[88] Jason Brownlee. *How to Control the Stability of Training Neural Networks With the Batch Size*. 2019. URL: https://machinelearningmastery.com/how-to-control-the-speed-and-stability-of-training-neural-networks-with-gradient-descent-batch-size/ (visited on May 28, 2022).

[89] Michael Middleton. *Deep Learning vs. Machine Learning — What's the Difference?* 2021. URL: https://flatironschool.com/blog/deep-learning-vs-machine-learning/ (visited on May 21, 2022).

[90] Xiaochu Wang and Wenyuan Tang. "Analysis and Evaluation of Baseline Manipulation in Demand Response Programs". In: *IEEE* (2020).

[91] Emil Larsen, Kenneth Rosenørn, and Anna Jónasdóttir. "Baselines for evaluating demand response in the EcoGrid 2.0 project". In: *CIRED 2019 Conference* (2019), p. 773.

[92] Kangping Li et al. "A Baseline Load Estimation Approach for Residential Customer based on Load Pattern Clustering". In: *Energy Procedia* 142 (Dec. 2017), pp. 2042–2049.

[93] Ioannis Lampropoulos et al. "A system perspective to the deployment of flexibility through aggregator companies in the Netherlands". In: *Energy Policy* 118 (2018), pp. 534–551.

[94] Souhaib Ben Taieb and Rob J Hyndman. "Recursive and direct multi-step forecasting: the best of both worlds". In: *Department of Econometrics and Business Statistics Working paper series* 19.12 (2012).

[95] Pecan street. *Pecan Street Dataport*. URL: https://www.pecanstreet.org/dataport/ (visited on Jan. 14, 2022).

[96] World Weather Online. *World Weather Online*. URL: https://www.worldweatheronline.com/ (visited on Apr. 17, 2022).

[97] Tyler Hodge. *U.S. Energy Information Administration: Hourly electricity consumption varies throughout the day and across seasons*. 2020. URL: https://www.eia.gov/todayinenergy/detail.php?id=42915 (visited on May 1, 2022).

[98] Optimizely. *Heatmap*. URL: https://www.optimizely.com/no/optimization-glossary/heatmap/ (visited on Dec. 9, 2021).

[99] Jason Brownlee. *How to Calculate Correlation Between Variables in Python*. 2020. URL: https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/#:~:text=A%20correlation%20could%20be%20positive,that%20the%20variables%20are%20unrelated. (visited on May 1, 2022).

[100] Raghav Vashisht. *When to perform a Feature Scaling?* 2021. URL: https://www.atoti.io/articles/when-to-perform-a-feature-scaling/ (visited on May 28, 2022).

[101] Jason Brownlee. *4 Strategies for Multi-Step Time Series Forecasting*. 2019. URL: https://machinelearningmastery.com/multi-step-time-series-forecasting/ (visited on May 6, 2021).

Simran Jit Kaur Sandhu and Marthe Vågen

Baseline Estimation for Flexibility Validation

NTNU

Norwegian University of
Science and Technology