MASTER THESIS

-

Optimized solutions for Smart Micro-grids

Author: Mattia Beretta

Academic Supervisor: Prof. Oriol Gomis Bellmunt

Company Supervisors: Hussain Kazmi

Friederik Van Goolen

September 06, 2017

Abstract

The share of distributed energy generation is growing at a rapid pace. The dropping cost of photovoltaic panels and Governments' incentives are making more and more convenient the installation of photovoltaic panels for privates all around the world.

In this thesis, data from 18 houses in the Netherlands is collected and analyzed to verify the effect of a large concentration of photovoltaic energy generation on the distribution grid. The study reveals that during Spring and Summer problems for the grid may arise due to the large amount of current injected into the grid.

Distributed storage, through the installation of batteries, and load shifting are simulated to test their effectiveness in the reduction of the over-injection problem. The results of the physical model are then studied from the economic perspective to verify which option is the most profitable.

Finally, different machine learning algorithms are implemented to predict the load consumption and photovoltaic energy generation one-day ahead.

KEY-WORDS: Solar energy, distributed generation, storage, DSM, Machine Learning.

Acknowledgment

This thesis is the final chapter of a two years long experience that has led me to study and live in many places all along Europe. I have experienced different cultures and met people from all around the world, for this opportunity I want to thank all the people from the RENE program at EIT InnoEnergy that have granted me this opportunity.

I am grateful for all the support that I have received from the people at Enervalis, they gave me the possibility to study and experiment on an extremely interesting topic. To Hussain Kazmi and Friederik Van Goolen goes my gratitude, they have been my mentors during the experience and helped me when I needed.

I want to thank Professor Oriol Gomis Bellmunt, whose courses at UPC and supervision helped me during the development of the thesis.

Finally, I want to mention my family and friends, who have been always there, supporting me during though moments and celebrating the good ones. I will always be thankful for what they have done for me.

Index

List of tables

List of figures

List of abbreviations:

PV: Photovoltaic

W_{DC}: Watt direct current

kW: Kilo-Watt

kWh: Kilo-Watt-hour

DSM: Demand Side Management

IRENA: International Renewable Energy Agency

TSO: Transmission System Operator

DSO: Distribution System Operator

NWP: Numerical Weather Prediction

SVM: Support Vector Machine

SARIMA: Seasonal Autoregressive Integrated Moving Average model

ARIMA: Autoregressive Integrated Moving Average model

AR: Autoregressive model

MA: Moving Average model

ACF: Autocorrelation function

PACF: Partial Autocorrelation function

ANN: Artificial Neural Network

VAT: Value-added tax

NPV: Net Present Value

NPC: Net Present Cost

RMSE: Root mean square error

nRMSE: Normalized root mean square error

MAE: Mean absolute error

MBE: Mean bias error

MAPE: Mean average percentage error

Chapter 1

1.1 Introduction

Climate change and rising awareness toward the effects of greenhouse gasses emissions are promoting the demand for cleaner energy production. Many Governments around the world are investing in renewable energy technologies trying to reduce costs, increasing the share of sustainable sources in the energy mix and create new job opportunities.

Renewable energies, like wind and solar photovoltaic generation, are not easy to predict due to their strong dependence from weather conditions. Storms or overcast days translate into sudden changes in the power output of windfarm or PV installations. The current electric grid is based on the constant balance between electricity generation and consumption, no grid-size storage installation is currently available, except for some pilot projects. The growing amount of unpredictable renewable generation poses a problem in the management of the electricity grid; solutions are needed to guarantee the security of the system without slowing the advance of renewables [1].

Uncertainty in the power generation is not the only problem the grid needs to address. Traditionally electricity has been deployed in a centralized manner. Big power plants produce the necessary electricity and through a voltage grid, organized on different levels, energy reaches the final consumers. This paradigm is being challenged by the rise of "prosumers", individuals that install photovoltaic panels on the rooftops of their houses. The specific cost of PV for residential installations is constantly decreasing, in 2009 was equal to 7.06 $\$/W_{DC}$ in 2016 it was 2.93 $\$/W_{DC}$ and it's forecasted to drop further [2]. In many countries, the injection of energy back into the grid is not only allowed, but encouraged through a system of subsidies and feed-in incentives, the grid is used as a virtual battery where excess energy can be freely dumped. In low voltage sections of the grid, where solar rooftops density is maximum, network congestion problems are becoming more and more frequent. The cable and the transformers installed in this branch of the grid require upgrades to cope with the increasing backflow of electricity. Grid upgrades are not only expensive, but also time consuming, thus system operators are looking for alternative solutions [3].

The above-mentioned challenges are putting in discussion the "status quo" of grid operation, managing the system in a "smarter" way seems mandatory. Recently there has been a significant increase in the

installation of smart meters, devices used to collect data of energy generation and consumption [4]. Access to real-time data is the first step towards a "proactive" and safer grid. Leveraging the high amount of data collected and the continuous advancements in data mining techniques is possible to gain valuable insights on user consumption patterns and even forecast future behaviors. Estimates of future generation and consumption are key elements for better grid planning.

Improvements in the storage technologies are another crucial element for the energy revolution, batteries prices are decreasing [5]. Batteries can be used to provide flexibility covering the discrepancies between load and consumption. Storage coupled with solar rooftops can significantly boosts self-consumption and reduces over-injection during sunny days. More distributed renewables can be integrated in the grid leveraging storage technologies [6].

Flexibility of the grid can also be increased through alternative solutions, like Demand Side Management. Most of the modern appliances can be monitored and controlled remotely, heat pumps, dish washers, laundry machines are all example of interruptible loads. Utilization of controllable loads can be shifted from peak hours to other periods of the day, when generation is abundant and underexploited. These solutions were unlocked by the rollout of smart-meters and by the introduction of information technologies knowledge in the management of the energy system [7].

Current years are a pivotal point in the evolution of the energy grid. The challenges to face are numerous and difficult to solve. Energy field is fertile ground for innovative business models and technologies.

1.2 Scope of the Thesis

This thesis utilizes data from 18 houses, located in Delfzijl a small city in the province of Groenningen, in the north of the Netherlands. Every house is provided with a 4.5 kW peak solar rooftop, heat pump and smart meters.

The scope of the thesis is to use available data to study and observe what happens when a high concentration of renewable generation is present in low voltage sections of the grid. The study of the data highlighted seasonal over injection problems that are addressed making use of storage and DSM. Finally, forecasting algorithms for PV energy production and load consumption are implemented. The mentioned algorithms are needed to develop an online system through which DSOs can interact with customers trying to adjust the status of the grid.

It is important to remark that the thesis is a preliminary study to assess the potential opportunity of using storage and DSM, hence some technical details, like analyzing the power flow of the distribution grid or deciding the optimal position for the installation of the batteries are not considered. Some information, especially grid upgrade costs should also be reviewed in a future study of the problem.

1.3 Organization of the thesis

This thesis is organized in seven chapters, this first chapter serves as a general introduction to the energy sector situation showcasing the most relevant trend and problems to address in the industry. The second chapter is a comprehensive review of the most significant topics in the contest of the study. Technological solutions for the improvement of the grid, forecasting algorithms and energy policies are discussed to prepare the framework within which the thesis is developed. The third chapter is focused on the study of the data, collection of the information and analyses performed are explained together with the main findings. Attention is paid to understanding the consumption patterns along the day and the year. Chapter four presents the simulation prepared to evaluate the installation of batteries and implementation of DSM. The economic profitability of the solutions implemented in Chapter four is investigated in Chapter five. Chapter six focuses on algorithms for the prediction of load consumption and PV generation, these are key tools in the implementation of a real-world energy management application. Different predicting algorithm are explored and their performances are evaluated. Chapter seven wraps up all the findings of the thesis, providing highlights of the results and suggestions for future developments on the topic.

Chapter 2

The transition toward a "smarter" grid is a very relevant problem, many researches with different focuses are undergoing in this study field. This chapter will present some of the previous works in the this research area; the results of the studies are mentioned and used to address the problem that this thesis is focusing on. In addition to the technical review, energy policies are analyzed to better understand the choices that the Dutch Government is taking.

Three important topics have been evaluated:

- *Usage of distributed storage*
- *Demand Side Management*
- *Forecasting algorithms for photovoltaic energy generation and load consumption*

2.1 Distributed storage using batteries

Batteries are seen by many experts as an effective tool for balancing the grid. Having some storage capacity allows to accumulate energy, when generation is greater than consumption and use it later in the day. Additionally, stored energy can be used to increase the security of the grid guaranteeing supplies in case of black-outs or adjusting unbalances of the network. While storage technology has been known from many years it has become relevant only recently, thanks to the technological improvements and the increasing amount of variable generation added to the grid.

In [8] the authors show the necessity of batteries and power electronics in the management of the grid, they also highlight the importance to create different revenue streams using batteries for more than one task. Household storage can be used to increase the reliability of the service in case of black-outs as well as a tool to boost self-consumption. The main result of the study is the simulation of two systems, a 5 kW PV system for home application and a 100 MW wind farm, both systems showed their potential to make the grid more stable. Another important finding is the influence of local condition of the grid and its topology on the design of the optimal solution.

[9] is a comprehensive report from IRENA summarizing the state of the art of battery technologies and addressing some potential usage opportunities. The report shows that storage market is becoming more and more important, prices are steadily decreasing due to recent technological breakthrough and large scale production of the batteries. Lithium-ion batteries is the rising technology, due to a good energy and power density, long lifetime and decreasing costs.

[10] is another study that exposes the different solution that are available and investigates the possible opportunities for the different shareholders, ranging from TSO to single customers. It has been found that batteries can be particularly beneficial for DSO as an alternative solution to grid upgrades and as a tool to decrease the thermal stress of the substations components. Batteries can also be utilized to improve the profile of the local voltage and finally they can provide back-up in case of emergency situations.

In the contest of this thesis Lithium-Ion batteries are used to relieve the stress on the grid caused by the high share of photovoltaic energy produced by the houses analyzed.

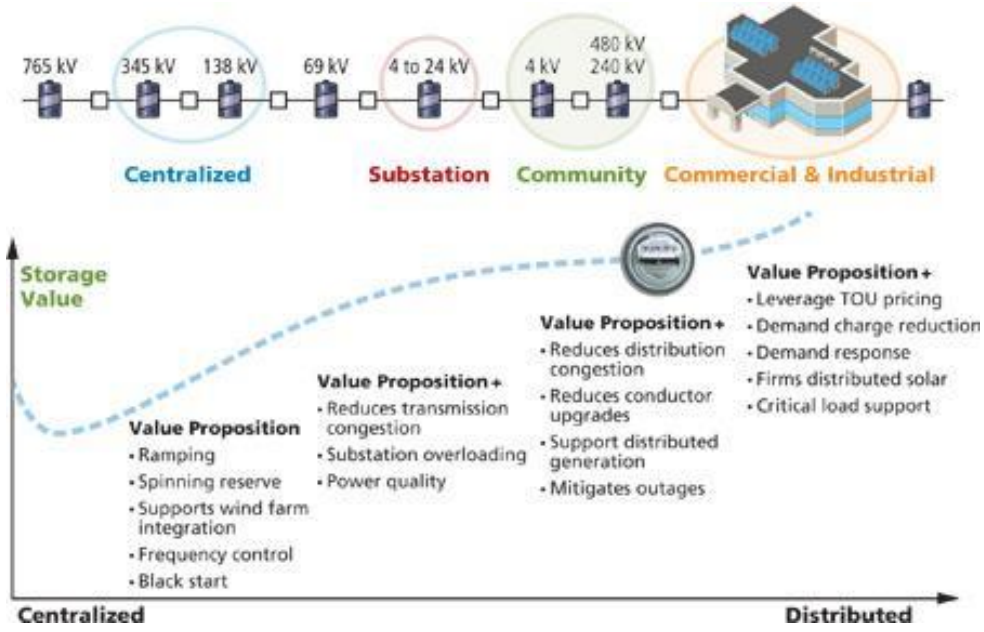

FIGURE 1: VALUE PROPOSITION OF STORAGE [11]

2.2 Demand Side Management opportunities

Demand Side Management is defined as a portfolio of measures to improve the energy system at the consumption side [7]. DSM is considered as one of the main enabler in the transition toward a "smarter" energy system.

Reference [12] analyzes the possibility to use heat pumps as a source of flexibility, different business models are presented. The results show that heat pump flexibility has high potential in driving down grid costs. An important remark from the study is the necessity of a clear regulation for these kind of activities, the creation of a new actor in the energy field, the aggregator and the deployment of an infrastructure that guarantees the interconnection of all the interested parties.

In this thesis is studied the possibility to shift the production of hot water and consuming part of the energy produced by the PV panels, instead of injecting it in an already congested grid.

2.3 Forecasting algorithms for photovoltaic energy generation and load consumption

The balance of the grid is a difficult task to accomplish, knowing the future values of generation and consumption is crucial for a better grid management. Many efforts have been done in the creation of forecasting tools of renewable energy generation and load consumption. Three approaches can be followed to generate weather predictions:

- Physical approach
- Statistical modelling
- Hybrid approach

Different input data is needed depending on the model used, outputs are also very different in terms of time and space resolution.

Numerical Weather Prediction (NWP) belongs to the physical methods class. This kind of approach is based on the solution of physical equations that describe the evolution of the weather given some initial conditions. NWP models are usually used to generate predictions with a large resolution both in time and space. Solving the physical equations requires high computing power, therefore this service is typically provided by national weather services. The output of these models often requires additional manipulations to remove the biases or extrapolate information about specific areas [13].

Statistical models are more general approaches, no prior knowledge of the system is required, the algorithms used are designed to "learn" from the data actualizing parameters and generating a function that maps the relation between input and output data. These methods are particularly interesting since they are very general, the same forecasting algorithm can be applied for different locations with some minor adjustments. It is also important to remember that the results of the model are heavily influenced by the quality and quantity of input data [13].

Hybrid models are an attempt to combine the two previous approaches. Using statistical models on top of NWP often results in very accurate predictions.

Previous works used different types of machine learning techniques. The decision of the algorithm that better fits the data is one of the key decision in a machine learning problem. Performance of statistical methods are heavily influenced by the amount and quality of input data, for example some locations may have a more stable and predictable weather compared to others. Comparing the results of different researches is difficult, since there is no consensus on the error metrics to use when presenting the results. Some attempts have been made to address this problem. Technical reviews, where different models are compared, have been written by several authors [15] [14].

In [16] SARIMA models in addition to Support Vector Machine (SVM) are used, interesting to notice how the combination of the two approaches results in an improvement of the performance. SARIMA models are good at capturing linear trends in the data, whereas SVM is used to capture the non-linear components.

In [17] quantile random forest is implemented to forecast the production of five PV plants using weather data as input for the model, the paper thoroughly explains the methodology and documents the advantages of using random forest – based algorithm in PV forecasting.

Reference [18] is a detailed study in which ANNs are used to forecast load profiles, different aggregation sizes of the dataset are also tested to observe their influence of the results. Interestingly the higher the aggregation level of the data the smoother the load profiles become and consequently forecasts' quality improves.

2.4 Dutch energy policies review

The Dutch Government, being one of the countries who signed the Paris Agreement, has pledged to drastically cut its carbon emissions in the following years. The current objectives are to produce at least 16% of the total energy from renewables by 2023. By 2050 the emissions should be reduced by 95% [19]. The Government is committed to reach its goals, an inclusive discussion where all the parties involved are listened has been started, to draft a plan that will allow the Netherlands to achieve these ambitious goals. Three main pillars to future policies have been identified:

- Reduction of CO_2 emissions
- Take advantage of all the economic opportunities that energy transition offers
- Integrate energy in spatial planning policy

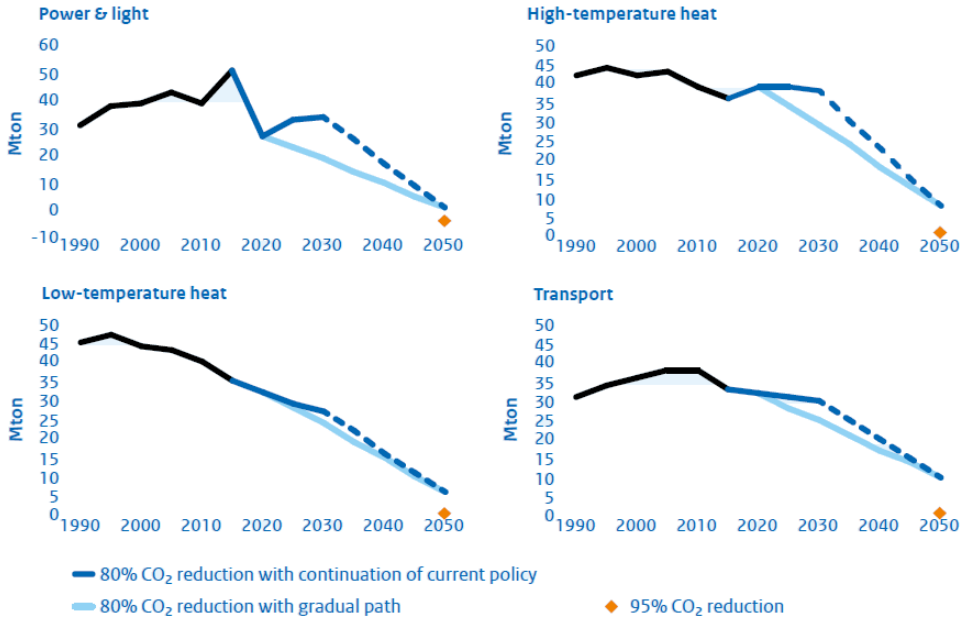

FIGURE 2: DUTCH GOVERNMENT 2050 GOALS [20]

Four working area are highlighted:

- Power & light
- High-temperature heat
- Low-temperature heat
- Transport

Regarding power and light sector, which is the most relevant for this thesis, the objectives are to reduce the carbon footprint of the energy industry, improve the Northwestern European electricity market and adapting the electricity system to accommodate the increasing decentralized supply of electricity and boosting the flexibility of the whole system. The measures that are being considered to achieve these goals are:

- Extending the energy production incentive scheme (SDE+)
- Collaborate with neighboring countries to avoid competition for subsidy tools between nations
- Proceed with the large-scale rollout of offshore wind energy
- Applying the successful approach used for offshore wind and utilize it for other technologies
- Encourage local renewable energy production

The Dutch Government strongly believes that energy transition will largely take place at regional and local level, central institutions should cover a support role, providing guidance and incentives to boost progress [20].

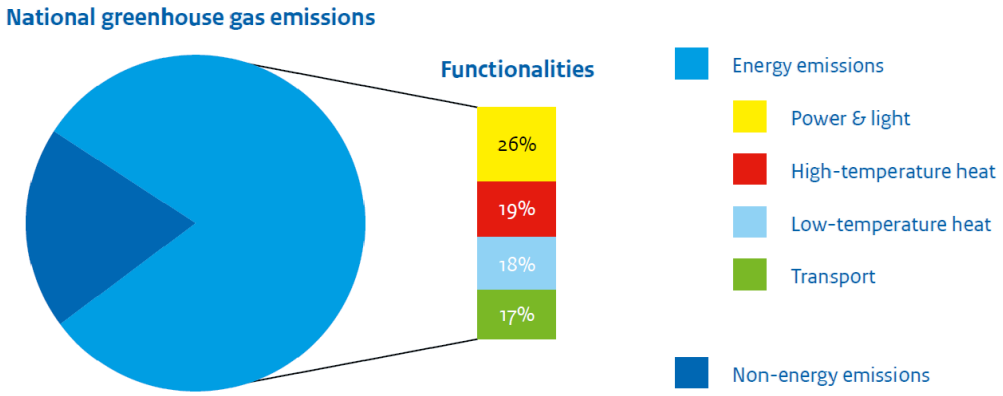

FIGURE 3: DUTCH EMISSION BREAKDOWN BY SECTOR [20]

High-temperature heat is important to address since in the Netherlands there are numerous energy-intensive companies that accounts for almost 25% of the total carbon emissions of the country. The challenge to face in the sector is particularly difficult, severe cuts in the carbon emission are needed, but at the same time the cost-competitiveness of the industry has to be preserved. Carbon tax schemes have already been applied by the European Union, but the results of such measures revealed to be

insufficient. The strategy of the Dutch Government is to utilize a mix of incentives and regulations. Energy conservation measures, more effective tax schemes, alternative heating methods and finally CO2 capture and storage are the tools chosen to make Dutch industrial sector more sustainable. A measure to reduce emissions that is deemed particularly effective is exploitation of deep geothermal heat, that according to a study made by ECN could cover up to 30% of the total heat demand of the industry [20].

Residential sector emissions are addressed in the low-temperature heat policies. This sector accounts for more than 30% of the total energy consumption of the country. Energy conservation and a reduction in the usage of natural gas appear as the main challenges to accomplish. No new gas infrastructure will be created in newly built districts, moreover the requirement to provide a gas connection will be replaced by a more general right to a heating infrastructure connection. The overall objective is to meet the heating requirement through local solutions like heat pumps, solar boilers, district heating and biogas installations, for these reasons a lot of decisional power is left in the hand of local administrations [20].

The fourth focus area is the transportation sector, which is still dominated by fossil fuels. The key words here are: fuel saving, sustainable biofuels and zero emission vehicles. The working areas are not only limited to technological development of solutions to address this problem, great importance is reserved to behavioral change. Transportation sector has some outstanding targets, by 2035 all the newly sold passenger cars should have zero emissions and by 2050 all the circulating cars should be emission free. Upgrades to the rail and road network have been planned for the future, charging points for electric cars will be deployed on the territory and the usage of locally produced fuel and renewable energies are promoted [20].

The current situation in terms of local energy production revolves around net metering. The Dutch Government realized the importance of distributed generation as a mean to foster social awareness toward energy transition. Even though distributed generation is not the most effective way to produce energy it is believed that its social function is worth to be supported. The form of net-metering applied in the Netherlands is particularly convenient for the users, since no energy taxes, no renewable energy surcharge and no VAT are applied on the netted electricity. This scheme is based on the usage of the grid as a virtual battery; electricity can be injected and extracted from the grid without any limitation.

In order to guarantee the security of the network some upgrades are needed. Interconnection with neighboring countries are planned, an extended grid can accommodate increased incoming and outcoming energy flows. Flexibility is also another key aspect in the reinforcement of the grid. Small-scale users' flexibility is fostered through the installation of smart meters that collecting data enables new possibilities in the DSM. Moreover, new actors are emerging in the energy field: aggregators. Their role is to gather group of small-consumer and provide market services such as flexibility through demand response. Another crucial point that will be addressed by the Government is the regulation of the electricity storage market regarding the fiscal policies that are applied to stored energy.

Chapter 3

The chapter's focus is on the data that has been used in the thesis. The data pipeline is presented in all its passages, from collection to visualization, passing through the "cleaning" phase where anomalies are handled. Main findings are also presented.

3.1 Data pipeline

3.1.1 Data collection

As mentioned in the first chapter, the data used come from one of the active projects of Enervalis, the company that proposed this thesis. Eighteen houses located in the city of Delfzjil, in the north of the Netherlands, are monitored. The houses are equipped with 4.5 kW peak PV panels, smart meters, a 200 liters hot water tank and heat pumps to warm up water and provide space heating.

FIGURE 4: DELFZJIL LOCATION

The readings from the smart meters are sorted in different categories: household load, PV production and heat pump consumption. Sensors do not track only electricity usage but also additional information such as hot water consumption and the level of water in the tank. The sampling rate can be adjusted to different values, starting from a 5-minute resolution.

Data utilized for simulation purposes has a sampling rate of 15 minutes, since this is the relevant time horizon for grid planning. Hourly data is used for analyses, the quarterly hour precision is not needed to discover the information hidden in the data and the resulting plot would be less readable. Since the variation of the measured quantities are expected to be dependent on the season, a year of data is collected from 1/7/2016 to 28/6/2017.

Python is the programming language used for data collection and elaboration. Data is queried using the Application Programming Interface (API) of Enervalis; information is requested remotely and provided to the user in the form of a JSON (JavaScript Object Notation) object, in which timestamp and desired values are stored.

3.1.2 Data cleaning and aggregation

Data is collected from each of the eighteen houses individually, but for visualization and analysis purposes it is aggregated. Consumption profile of a single house may vary a lot during time, whereas aggregated ones are easier to understand and predict. Moreover, for grid planning purposes is not important to know the details of each single house; regulation is done on groups of them.

Sometimes sensors go offline for several reasons, when it happens data shows anomalies that need to be corrected. In the case of consecutive wrong readings, the values are substituted with the ones measured the previous day at the same hour. Isolated anomalies are easier to treat, an average of the adjacent values can be used. Luckily very few times data needed to be corrected, hence the overall results of the analysis were not modified excessively.

3.1.3 Data Visualization

Visualizing the data helps to capture the most relevant information. The different components of the system are presented individually and analyzed. First, electricity generation and consumption is discussed, a separate analysis is devoted to hot water consumption due to the usage of heat pumps' flexibility.

3.2 Photovoltaic panels' data

The PV panels' energy profile is both daily and yearly time dependent. Several factors influence the final production, the most important ones are: solar irradiation, temperature and humidity all of which vary with the seasons. The following figure represents the daily PV production by month. The graph used is a boxplot, the average value is shown with a red triangle, the median with a green line and the black whiskers show how much the data is sparse.

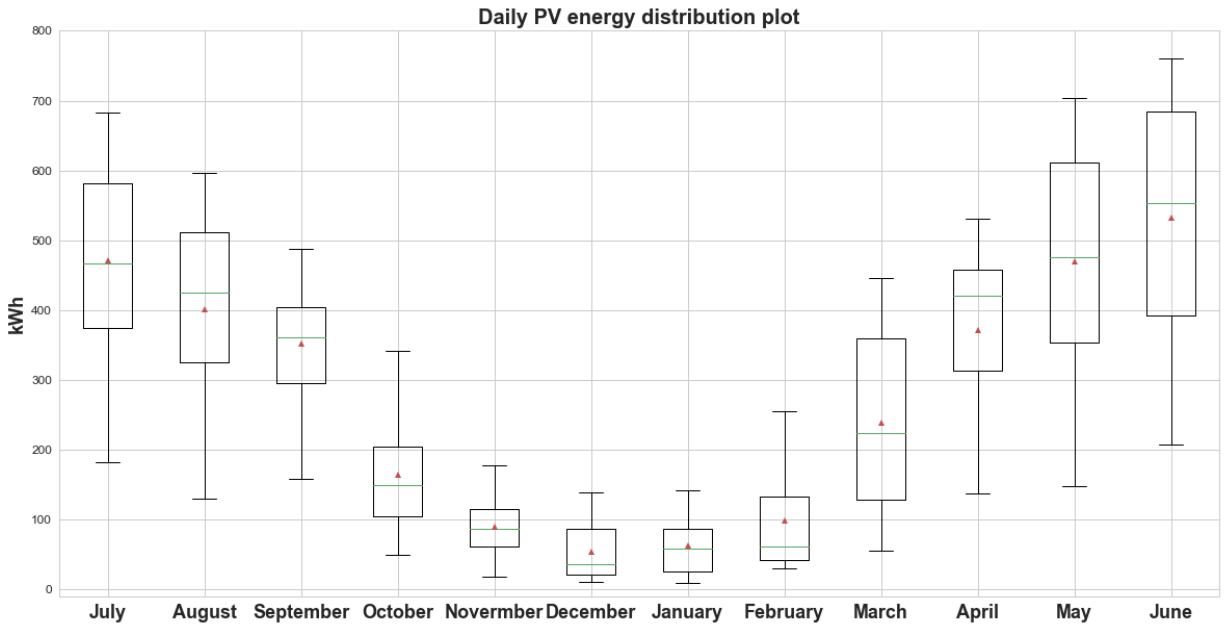

FIGURE 5: DAILY PV ENERGY BOXPLOT

Energy generation is minimum during Winter months and maximum during Summer, also notable is the high amount of electricity produced in May. Late Spring is a very convenient period for PV production due to the high irradiance and mild temperature. Another important observation is the large difference between the maximum and minimum values in Summer and Spring. In the Netherlands, these periods are characterized by great variability of weather conditions, while some days are hot and sunny others are windy and rainy.

3.3 Load data

Electrical load consumption need to be divided in its controllable and fixed components. The installed smart meters differentiate only between household load (meaning all the appliances, lightning, etc.) and the consumption of the heat pumps. Household load is considered as fixed; while some smart appliances may be installed, no information about their individual consumption is available, hence nothing can be done with this information. Heat pumps usage has a dedicated reading from the meters. Space heating and hot water production are provided through heat pumps, the latter function is particularly interesting due to the presence of a hot water tank that allows to decouple the production of hot water from its consumption.

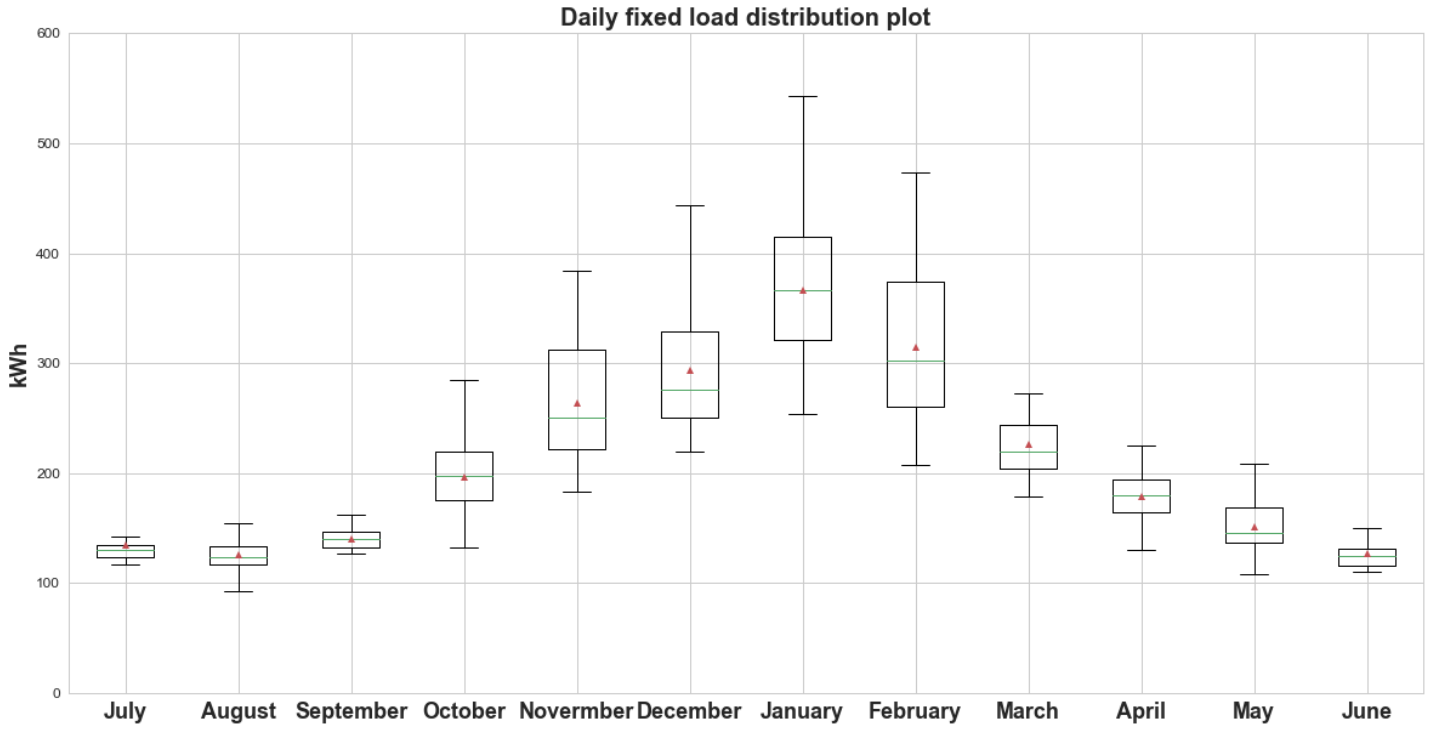

FIGURE 6: DAILY FIXED LOAD BOXPLOT

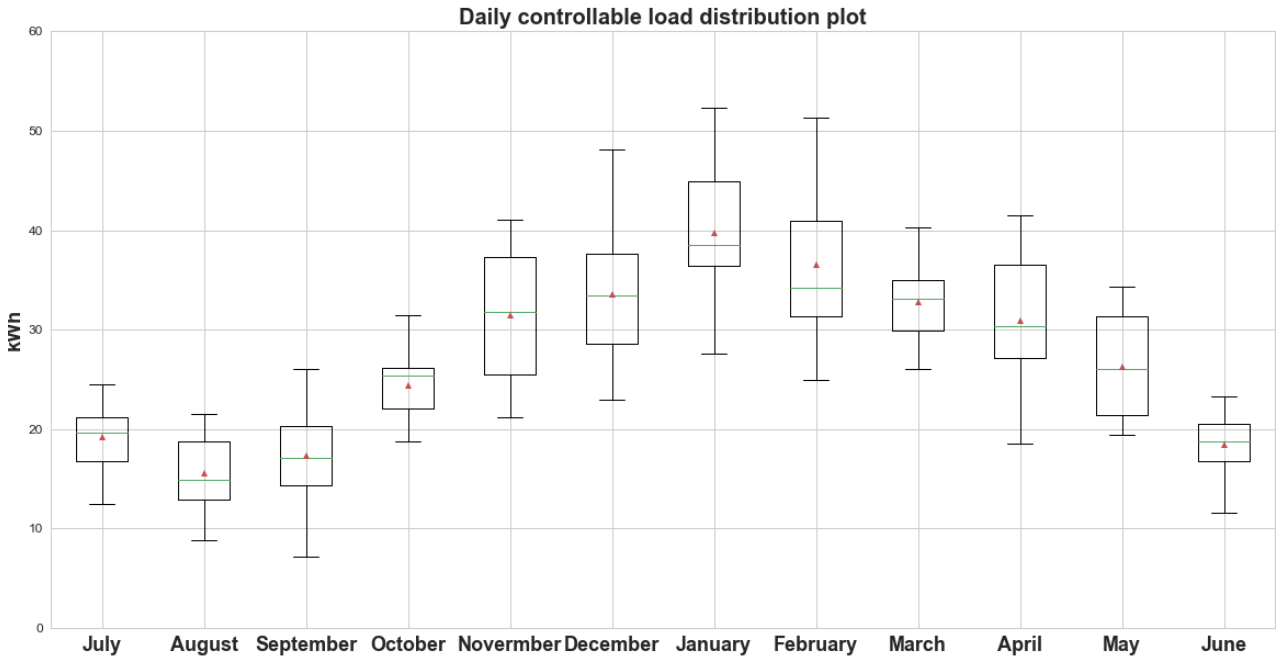

FIGURE 7: DAILY CONTROLLABLE LOAD BOXPLOT

The above figures compare the fixed and controllable load energy consumption along the year. The magnitudes of the two load categories are very different. Controllable load is just a small fraction of the total consumption, but controlling it helps to change the shape of the daily load curve.

3.4 Excess energy

So far production and consumption have been analyzed, but what is truly interesting to observe is the net difference of these values: the excess energy that, without storage, would be injected into the grid.

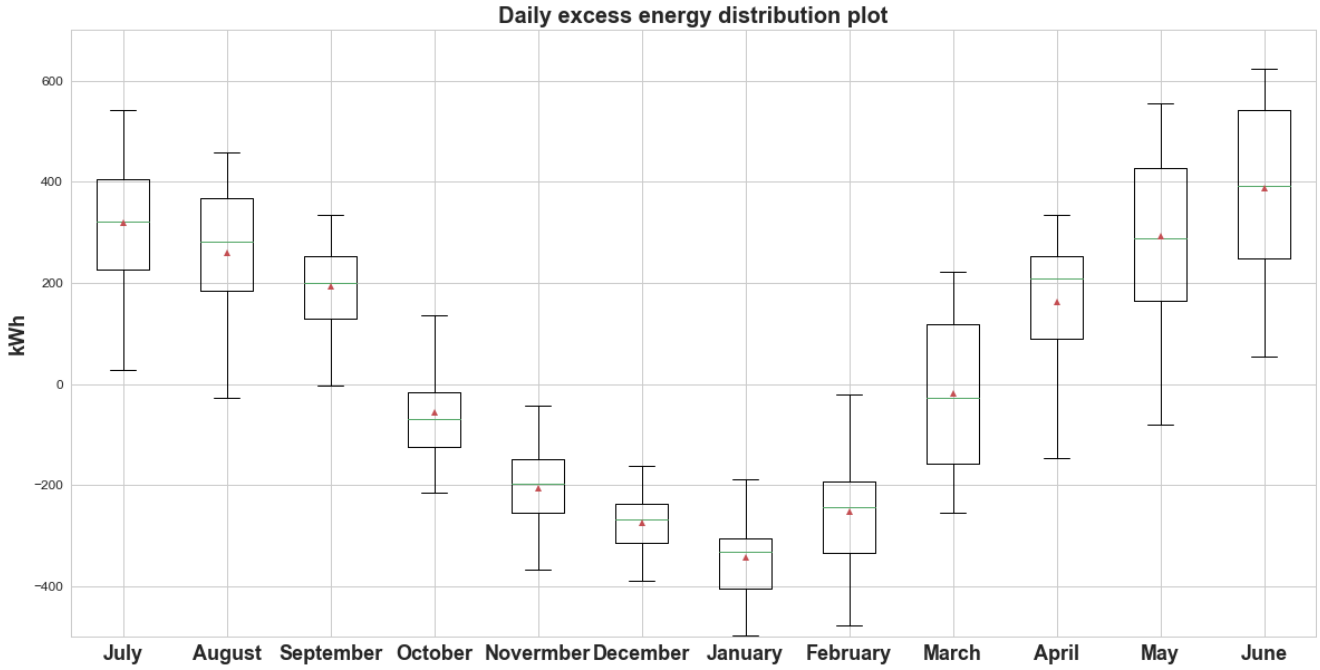

FIGURE 8: DAILY EXCESS ENERGY BOXPLOT

Late Spring and Summer are the period in which there is a constant daily net injection of energy into the grid, while Fall and Winter are characterized by net import from the network. The amount of energy injected during the hot season is considerable and could lead to problems in the system. Spring and Summer situation is analyzed more carefully, time-series plots of load and consumption are created.

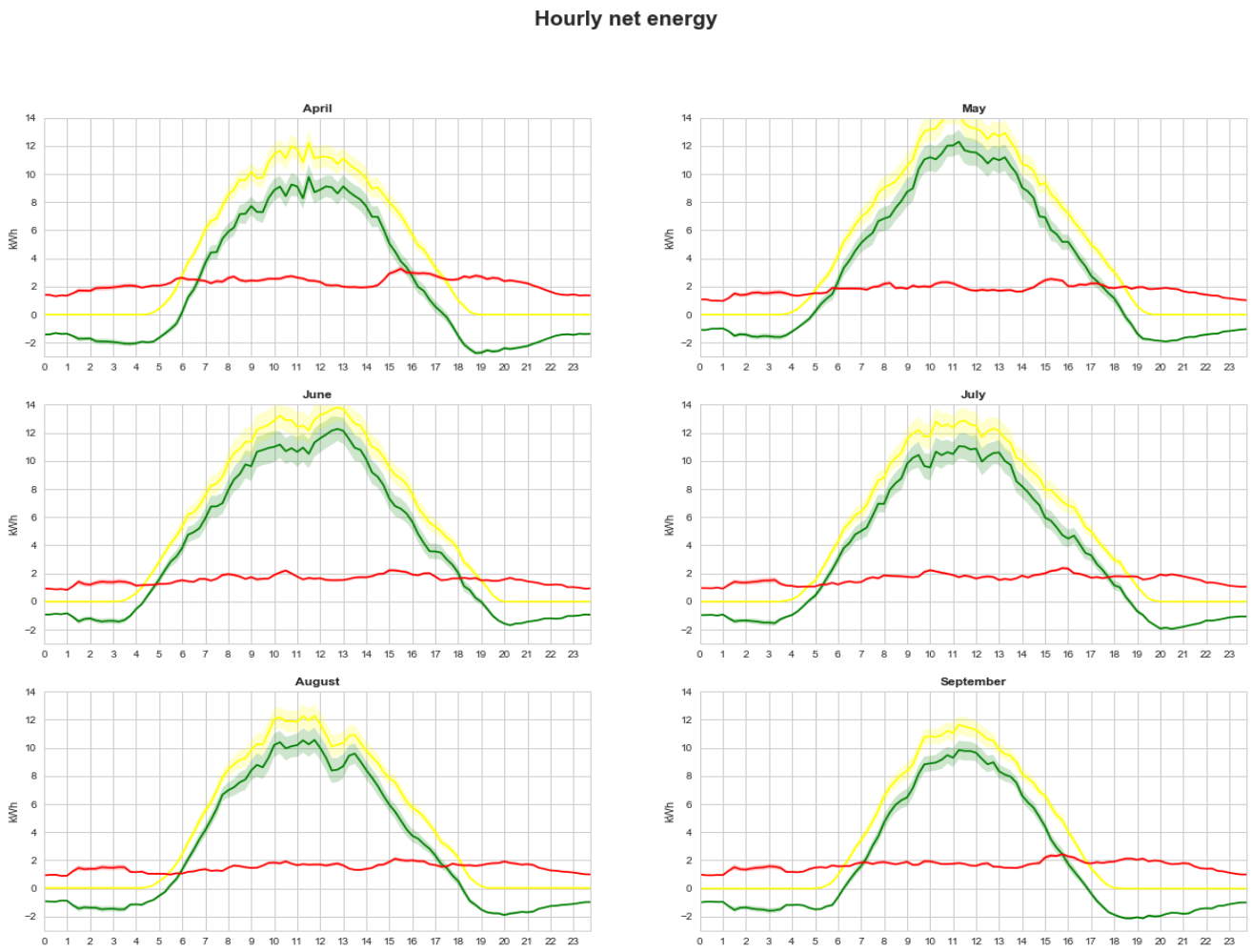

FIGURE 9: NET ENERGY HOURLY TIME-SERIES

The plots show several interesting information, first the difference between load and consumption is significant, thus the large amount of excess net energy. Secondly, the production is highest when load is fairly low, charging batteries during this part of the day or shifting the hot water production could be good strategies for reducing current injection during the day.

3.5 Heat pumps data

Heat pumps data is particularly interesting since the information available does not only include the consumption, but also the operating mode of the devices. Moreover, the equipment can be controlled remotely, their usage can be scheduled. The mentioned operating mode are three: space heating, hot water production and a sterilization cycle.

3.5.1 Space heating

Heat pumps can be used to warm the rooms of the houses, the devices are not able to provide cooling. Space heating is not a flexible service, it is heavily influenced by the presence of people in the house. Thermal inertia of the building could be utilized to activate the heat pump in advance, but the implementation of such as a system would not be trivial and could lead to unexpected increases in the electricity consumption.

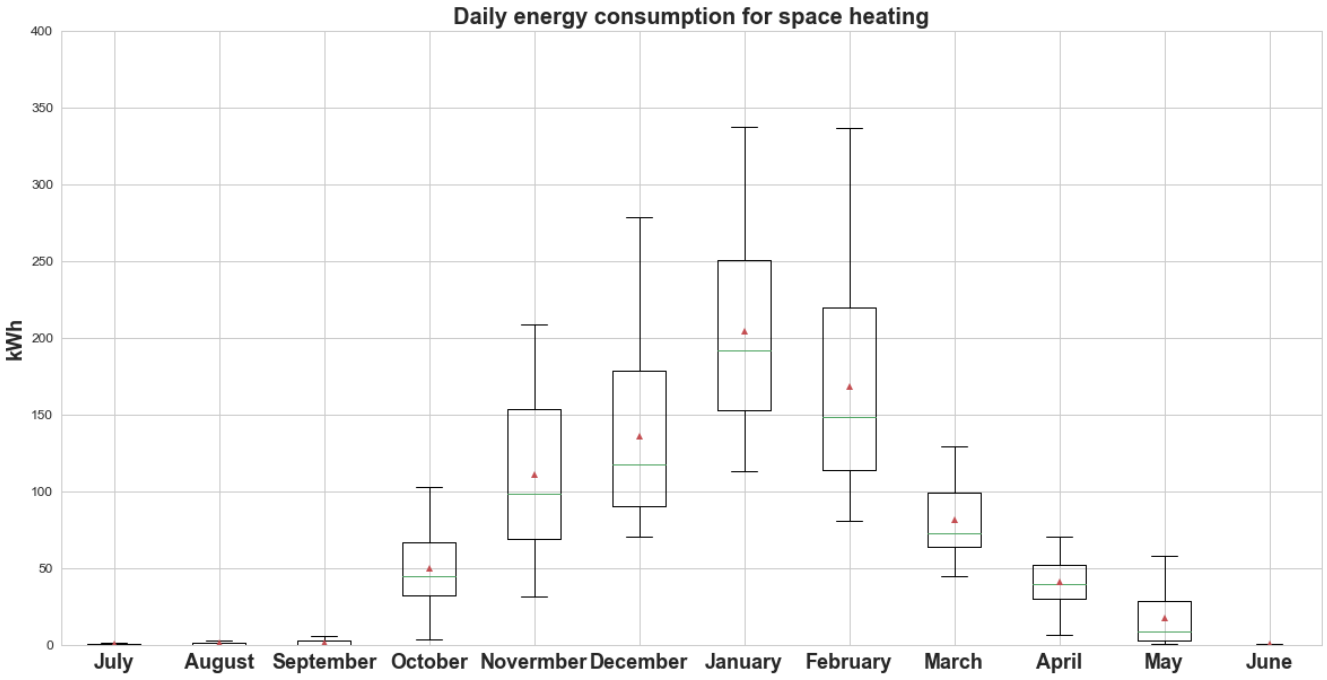

FIGURE 10: DAILY SPACE HEATING ENERGY CONSUMPTION BOXPLOT

Since no cooling can be provided by the installed heat pumps, consumption during Summer months is very low, almost nihil. Overall space heating is not interesting in the analysis, since it cannot be controlled and its usage is low during the critical period for over-injection.

3.5.2 Sterilization cycle

Warm stagnant water is the perfect habitat for bacteria proliferation, that is why the tanks need to be sterilized regularly. To do so, the tanks are heated up to 65°C; while the activation procedure can be easily scheduled to the most convenient time of the day its activation is required only once a week, hence its overall energy consumption is not relevant.

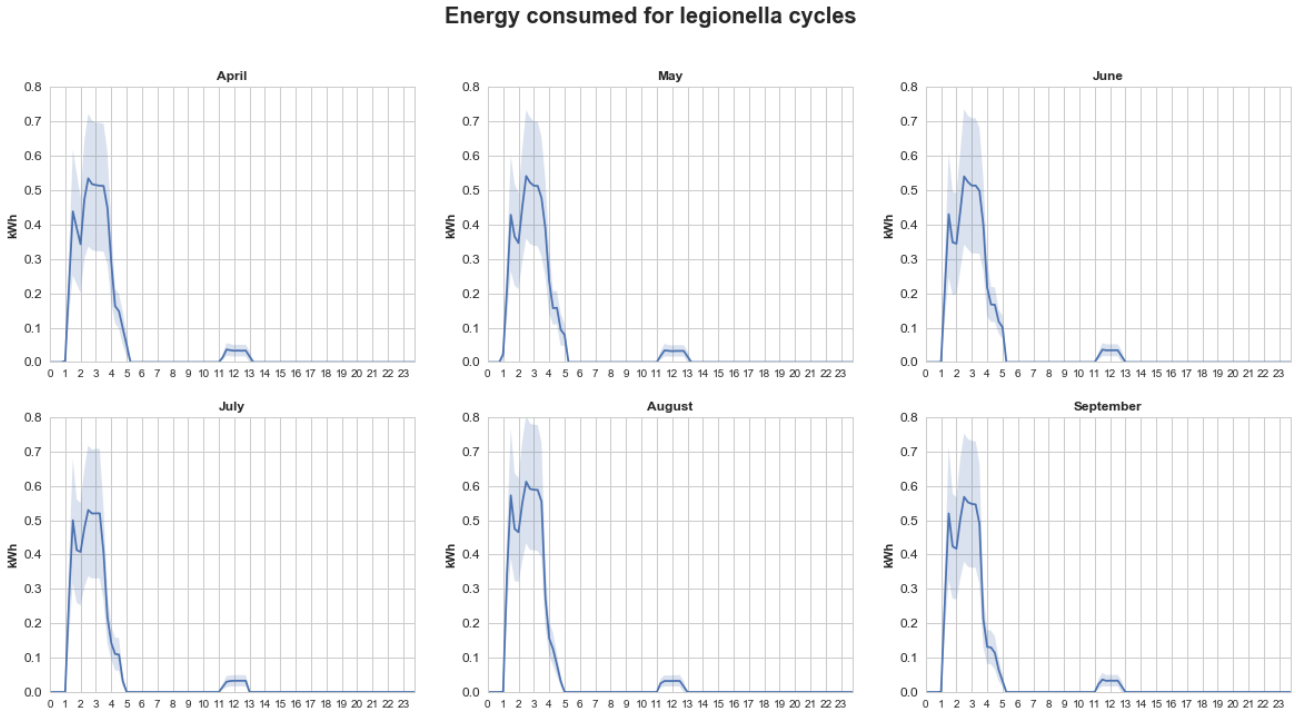

FIGURE 11: HOURLY ENERGY CONSUMPTION STERILIZATION CYCLE

3.5.3 Hot water production

This function is the most interesting one, due to its flexibility and limited disturbance of users' habits. The average daily consumption is around 20 to 25 kWh/day during the Spring-Summer period, on top of that the presence of the hot water tank makes it controllable. Energy consumption is significantly higher during Winter and Fall while water consumption is not much higher, this is related to the different temperature increase required and the consequent reduced thermodynamic efficiency of the system.

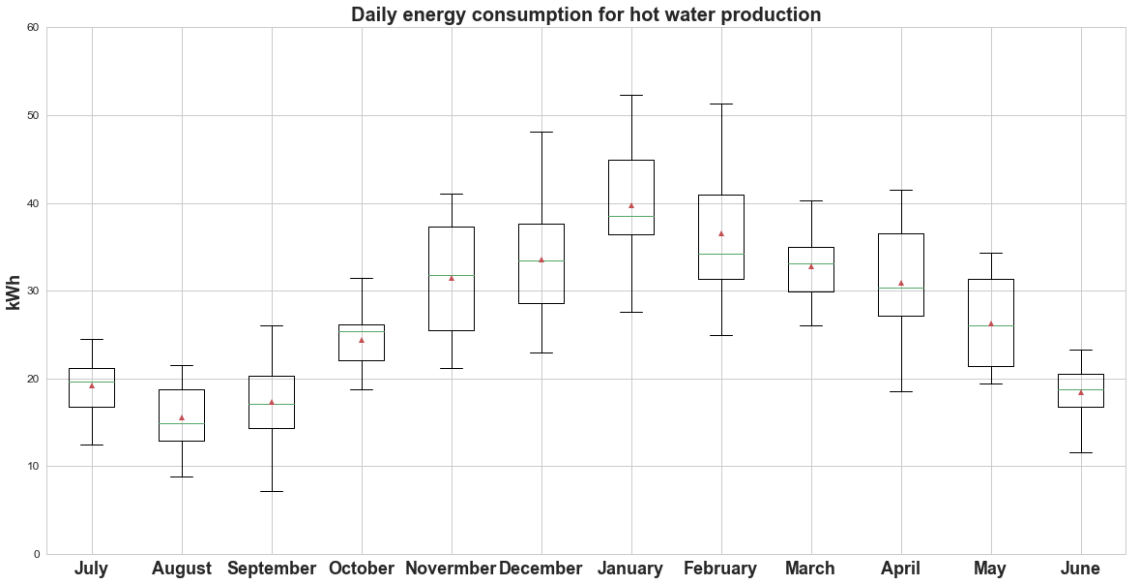

FIGURE 12: DAILY ENERGY CONSUMPTION FOR HOT WATER PRODUCTION BOXPLOT

It is important to verify the time at which energy is consumed during the day to warm up water, boxplot do not provide this kind of information, time-series plots are better suited for the task. These graphs show the average daily profile, which is drawn using a solid line while the shadowed area is used to show how much the data varies during the observation period.

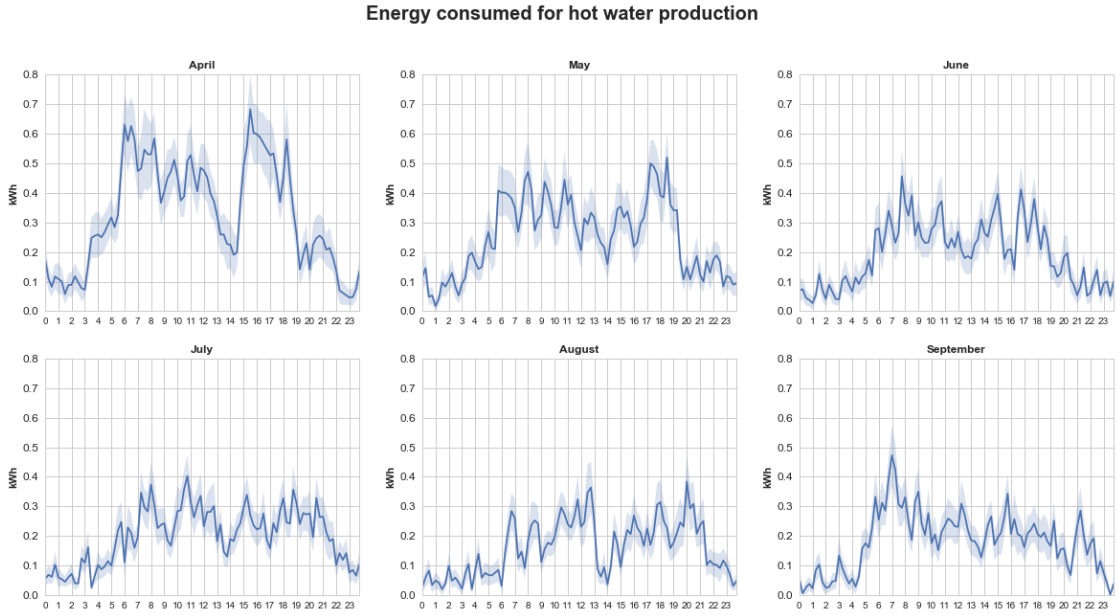

FIGURE 13: ENERGY CONSUMPTION TIME-SERIES FOR HOT WATER PRODUCTION

Hot water is used mainly during the day and depending on the month two or three peaks are visible. Data show high variation along the day, but it looks like that from noon to 2PM consumption is generally low. Variation from one month to the other can be partially related to the vacation periods in the Netherlands, in particular it can be observed that, during Summer, peaks appear slightly later during the day.

3.6 Water consumption data

Consumption patterns need to be analyzed to assess the feasibility of shifting hot water production. The tank capacity is 200 liters, the daily consumption should be lower than the storage volume and the peak consumption hours should be sufficiently separated one from the other, so that water can be warmed in the meantime.

The first plot presented is the aggregated daily average consumption of the 18 houses month by month.

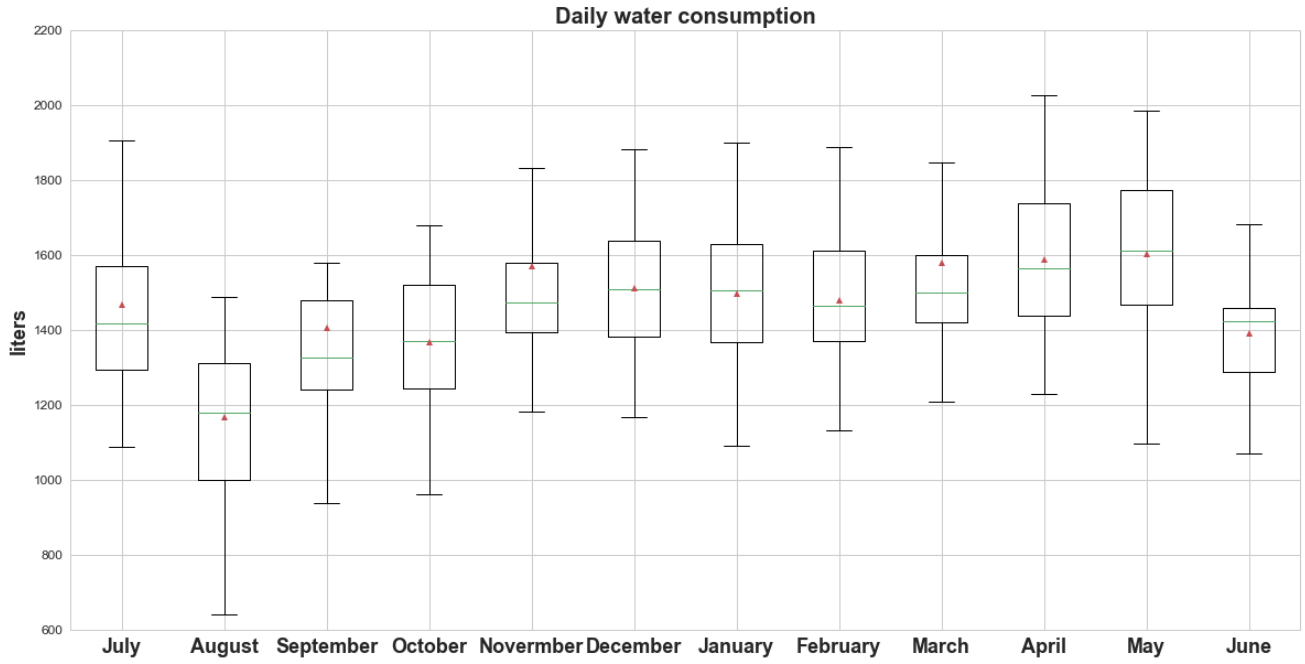

FIGURE 14: DAILY WATER CONSUMPTION BOXPLOT

The average daily consumption for the entire year is around 1470 liters per day, that divided by the number of houses, gives 86.5 liters per day per house of hot water. Also, it is important to observe that the consumption is relatively constant throughout the year, a slight increase is visible during Spring. The lowest consumption is registered in August, most likely because of vacations. The installed tanks are much larger than the daily average consumption, thus no problem should arise by shifting the hot water production.

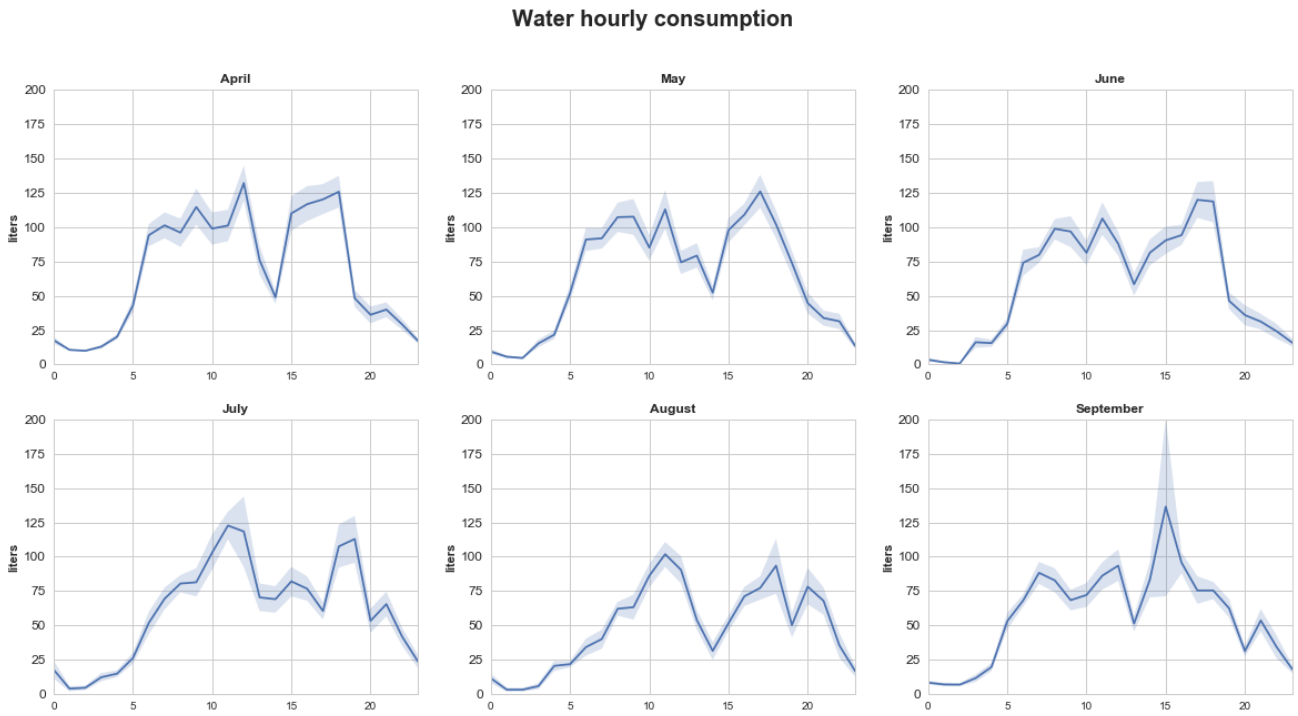

FIGURE 15: HOURLY WATER CONSUMPTION TIME-SERIES

Besides general statistics values, it is useful to observe the time-series plot of hourly water consumption to identify the peak hours. Only Spring and Summer months are analyzed, since they are the periods during which over-injection is more likely to be a problem. Depending on the month two or three peaks are visible, moreover it can be seen how these are concentrated during morning and late afternoon, while around noon consumption is low. Excess energy is often high around midday, since the solar irradiance is particularly strong and the consumption is low, hot water production could be shifted around this time harvesting the heat necessary to warm up the water for the entire day.

Additional elaboration on the data is presented to pinpoint the highest consumption hours, since interpreting time-series plot is somehow subjective. The first approach utilized consists in summing the maximum hourly values one by one until the 75% of the daily consumption is covered, the selected hours are stored and their frequency showed through histograms.

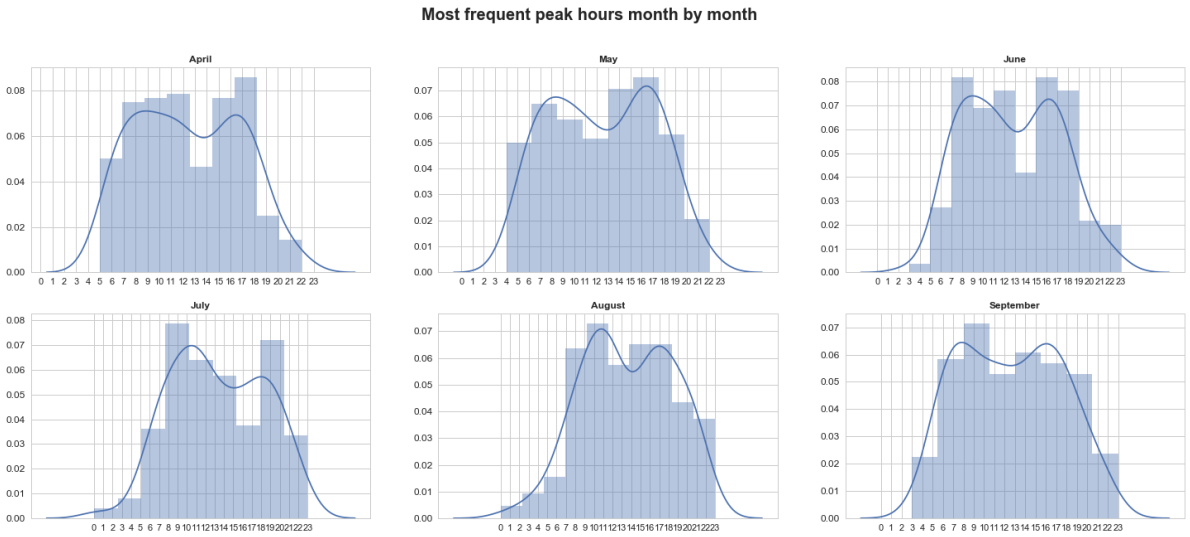

FIGURE 16: WATER CONSUMPTION PEAK HOURS HISTOGRAMS

A second approach is to first decide the maximum number of hours to sum for each day and then check how much of the total consumption is covered by the sum of the selected values.

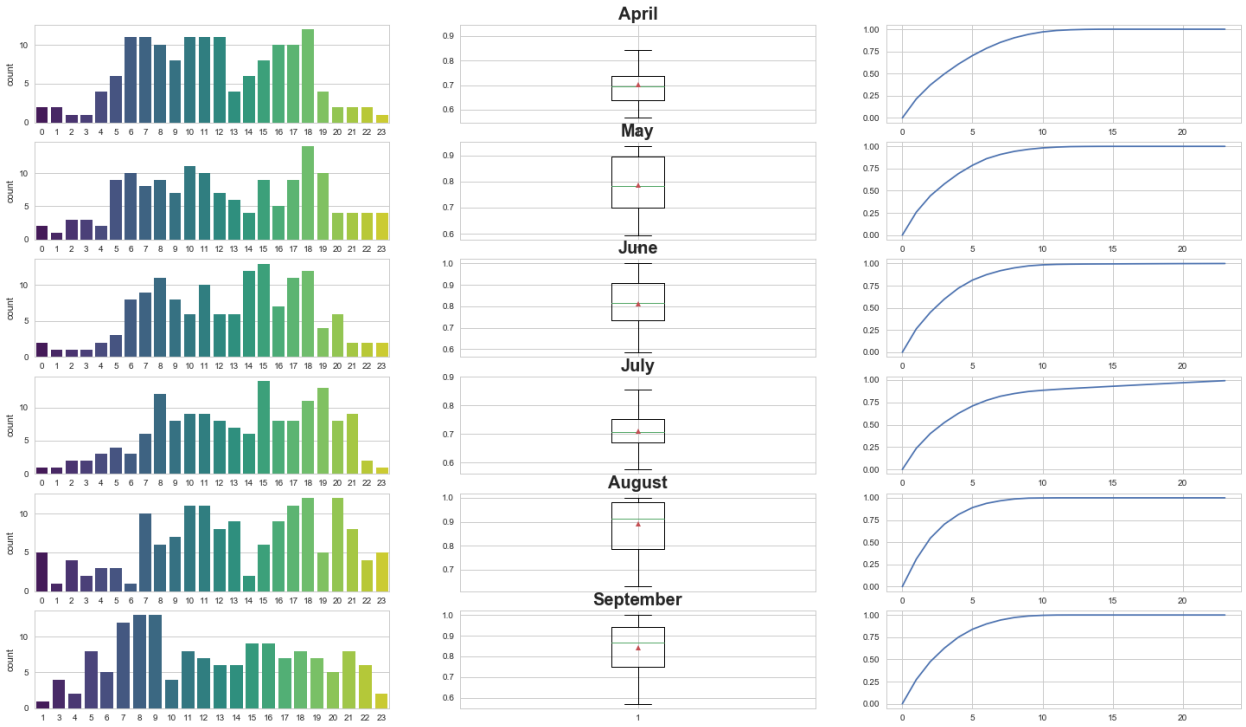

FIGURE 17: WATER CONSUMPTION PEAK HOURS HISTOGRAMS AND BOXPLOTS

The resulting plot is composed of three subplots, on the left is the count of how many time a certain hour has been selected, the central boxplot shows how much of the daily consumption was covered selecting a certain number of relevant hours, finally the plot on the right is a graph that shows how much of the total consumption is covered by selecting a certain number of hours. Using the first 5 maximum consumption hours of each day the 75% of the daily consumption is represented.

Heatmaps are another useful tool to quickly visualize the top consumption hours over a long period of time.

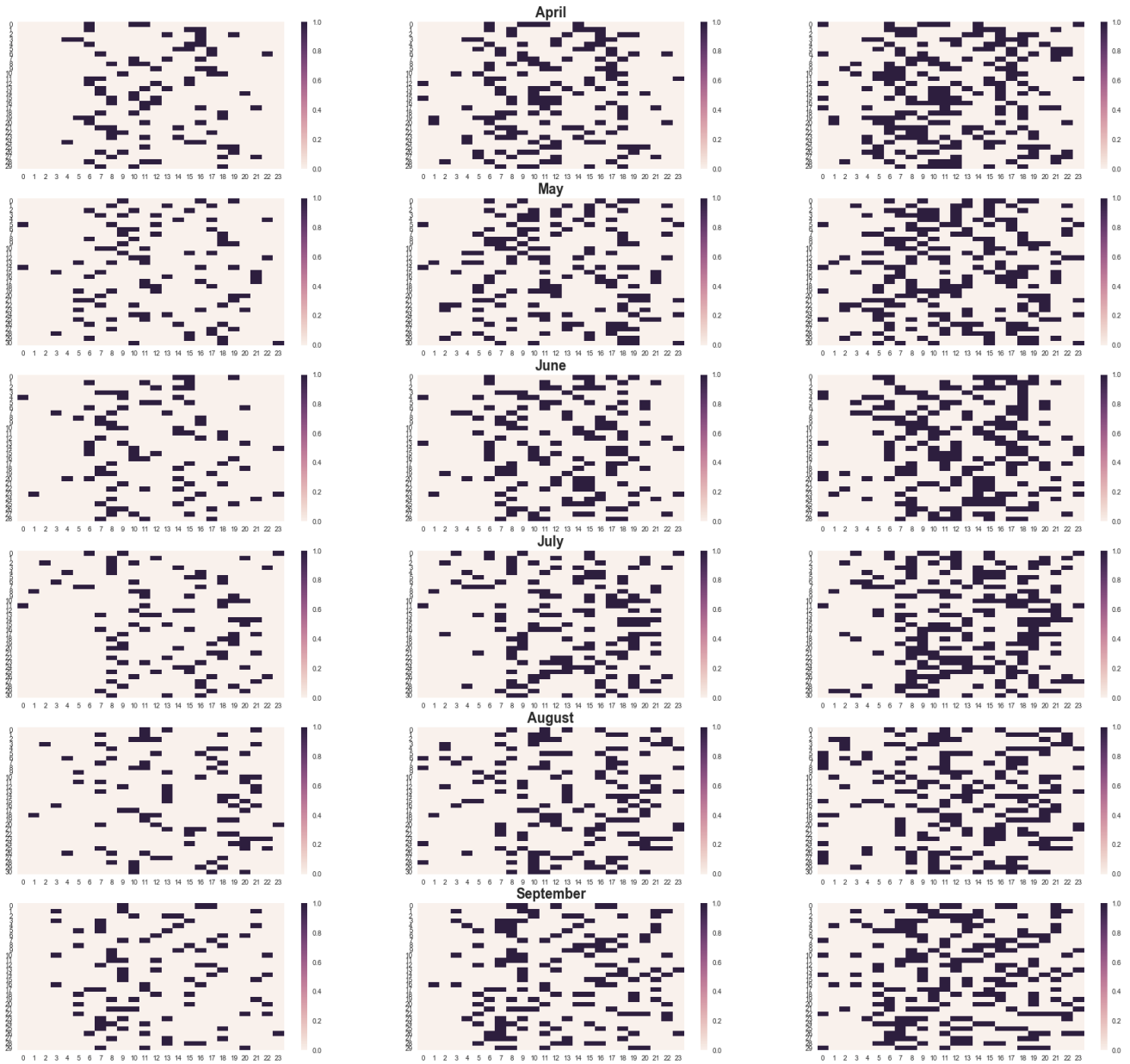

FIGURE 18: WATER CONSUMPTION PEAK HOURS HEATMAPS

Each row represents a specific month, the first column is created using the first three peak hours, the second using five, the third one seven. A black square means that the corresponding hour is a peak consumption one for the specific day. This kind of representation is a quick way to summarize high dimensional data in one picture and understand the consumption patterns over several months in one plot. Using the two approaches and the resulting plots, the initial intuition formulated from the time-series plot can be confirmed, morning and late afternoon hours are the periods during which consumption is higher, midday is a convenient time to refill the tank and cover the consumption of the entire day.

Chapter 4

To verify the effectiveness of batteries and DSM in solving the over-injection problem a simulation is prepared. In this chapter the key assumptions for the simulation, the flowchart and the operative strategy of the system, the presentation of the results and their discussion from a technical standpoint can be found.

4.1 Key assumptions

Due to lack of some information and to simplify the analysis some assumptions are made. The most critical one is the current injection threshold that has been set to 48.6 kWh, corresponding to the 60% of the total peak production of the eighteen houses. The value used is not official, no project is currently undergoing with the DSO of the studied area and it is also extremely dependent on the local condition and topology of the grid. The used value has been suggested by co-workers at Enervalis as a reasonable starting point considering their experience on similar projects. Another assumption related to the over-injection is how to deal with surplus energy, in this analysis it is curtailed. Batteries simulation required some simplifications, round-trip efficiency is set to 90%. No studies on the effect of full discharges, or on the optimal positioning in the grid are made. It is important to remark that the scope of this study is to evaluate the profitability of different alternatives to cope with the back-flow problem. No detailed data about the topology or state of the grid is available, hence very refined technical analyses cannot be prepared.

4.2 Starting conditions - curtailment

A preliminary study on the curtailment is beneficial, the most problematic periods of the years can be identified. Quarterly hour data is studied and the following plots made.

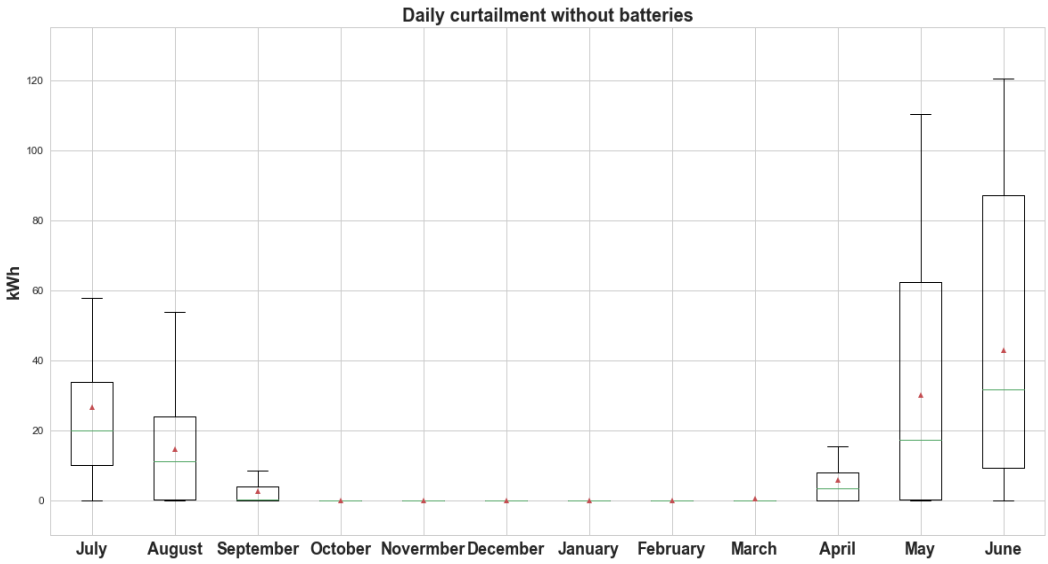

FIGURE 19: DAILY CURTAILMENT BOXPLOT

It can be observed that curtailment is needed only in Summer and Spring, thus different operative strategies for the batteries will be necessary. Winter and Fall should take advantage of the batteries to increase self-consumption, whereas in Spring and Summer the focus should be on reducing the injection into the grid.

Another important finding is that energy is curtailed in the central part of the day, thus hot water production is concentrated between 9AM and 3PM.

4.3 Battery strategies

Battery utilization strategies should be set accordingly to the goal to achieve. Two strategies for charging and discharging are considered. The naïve one is to charge the battery every time the energy produced by the PV panels is higher than the load. An alternative operative mode is to charge the battery only when the injection threshold is not respected, the main goal of this second approach is to minimize the curtailed energy using the storage of the battery to relieve some stress from the grid. Considering the findings about the curtailed energy the first battery strategy should be used during Winter and Fall, whereas the second one is better suited for Spring and Summer.

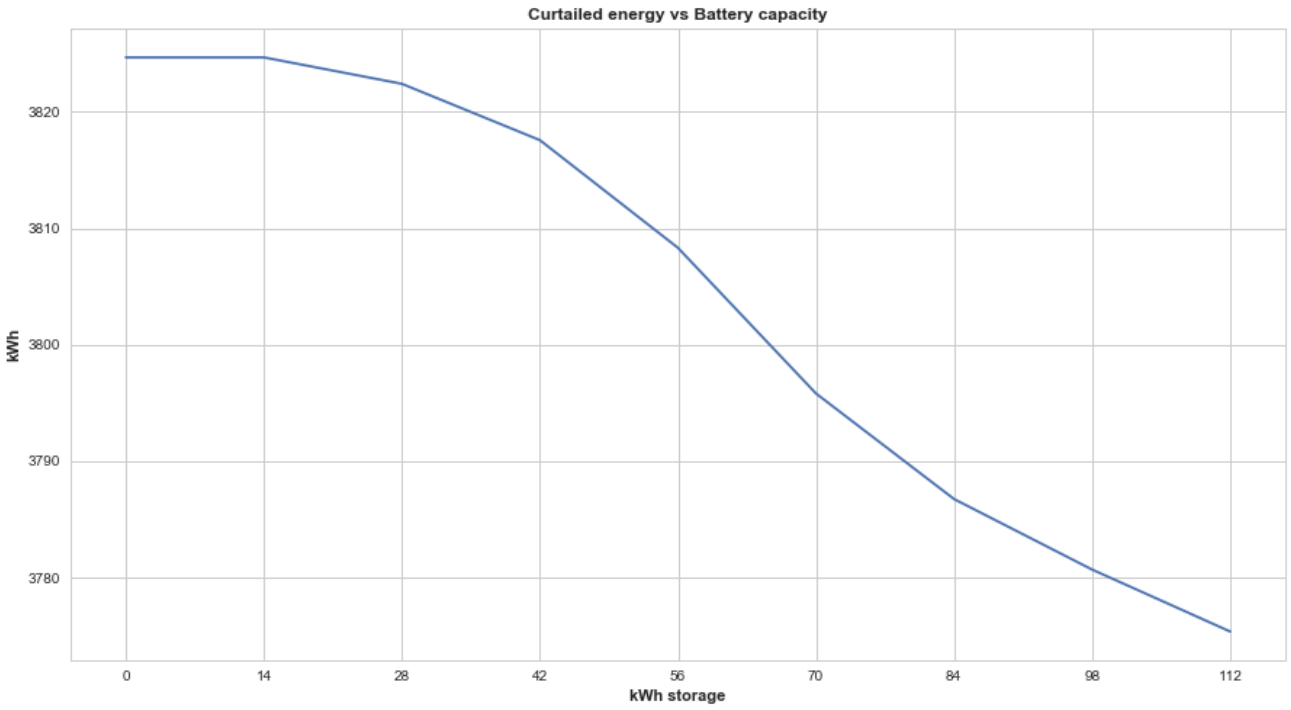

FIGURE 20: CURTAILED ENERGY AS A FUNCTION OF BATTERY SIZE

Charging the battery every time production is higher than consumption leads to very poor results, as figure 20 shows. The reduction in curtailed energy is extremely low, thus a different strategy for the management of the battery is adopted.

4.4 Simulation diagram

Different battery sizes are tested to find the optimal size of storage to install. Historical data is utilized for the simulation, a quarter hour time span has been used, since this is the relevant time horizon in grid planning. Two different flowcharts are designed to account for the different operative conditions during Fall-Winter and Spring-Summer.

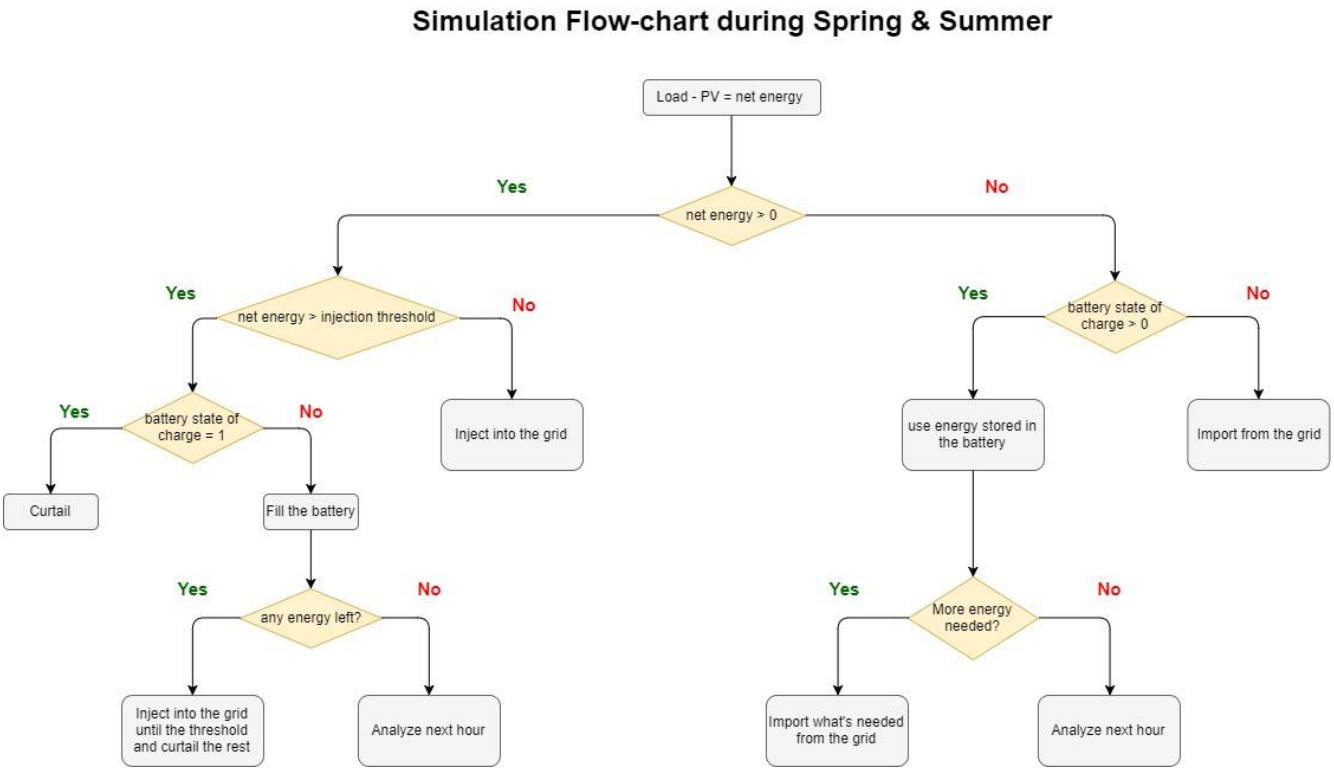

FIGURE 21: SIMULATION DIAGRAM FOR SPRING AND SUMMER

The first step in the simulation is to compute the difference between the energy produced by the PV panels and the load consumption. When the resulting quantity is larger than the injection threshold the capacity of the battery is checked, if the battery is not full, energy is injected into it. The energy left is exported to the grid, until the injection limit is reached, the remainders are curtailed. Whenever consumption is larger than production energy from the battery is used to cover the deficit, if it is not enough the rest of the needed energy is bought from the grid.

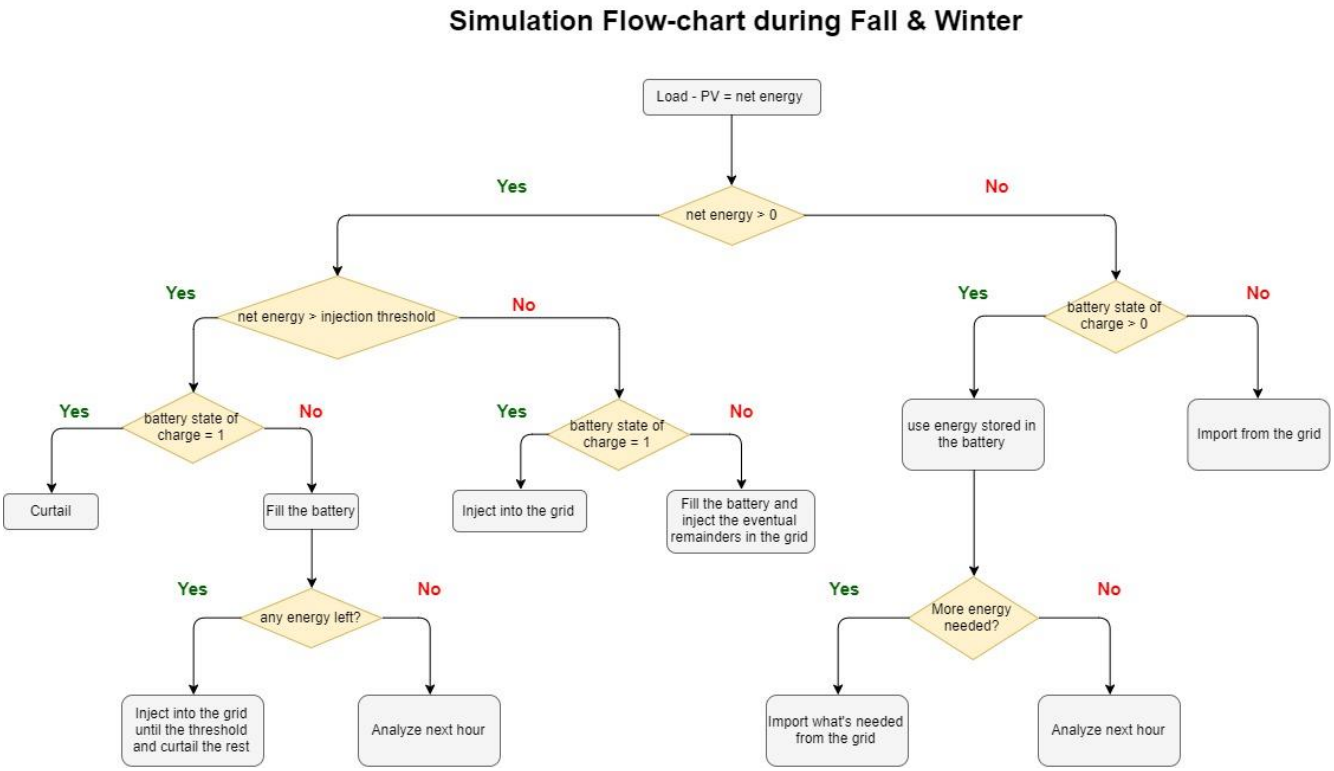

FIGURE 22: SIMULATION DIAGRAM FOR WINTER AND FALL

The Winter and Fall simulation is almost identical to the Summer and Spring one, in fact the right branch of the flowchart is left unchanged. The differences appear when production is larger than consumption, here the battery is charged whenever storage capacity is available and not only during peak hours.

4.5 Simulation scenarios

4.5.1 Battery and Business as Usual energy management

In this scenario, several batteries of different size are tested while the energy utilization is not modified. The results of the simulation serve as baseline in the comparisons with the other alternatives, moreover the effect of installing storage only can be observed.

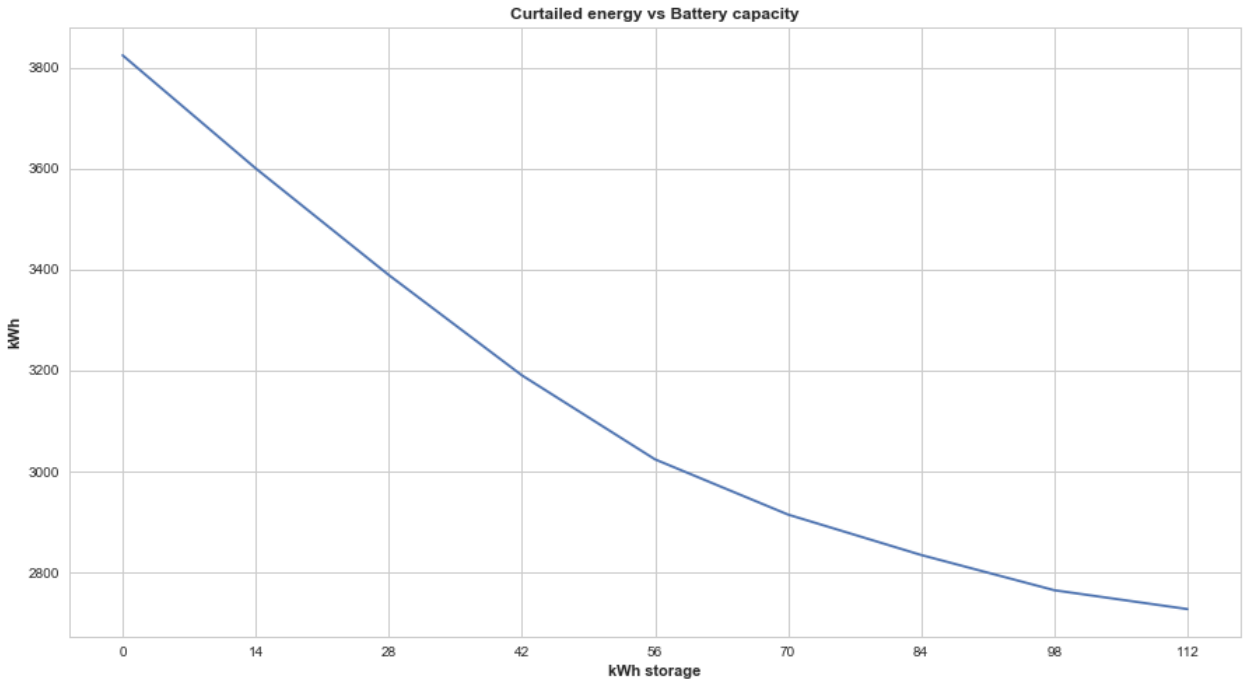

FIGURE 23: CURTAILED ENERGY AS A FUNCTION OF BATTERY CAPACITY BAU SCENARIO

Storage clearly helps to decrease the amount of curtailed energy. In one year, hundreds of kWh of electricity can be saved with the installation of a relatively small amount of batteries.

The amount of curtailed energy does not decrease linearly with the storage size thus, installation of additional capacity should be carefully evaluated to ensure its economic profitability. The physical reason behind the shape of the curve can be understood analyzing the state of charge of the batteries hour by hour, day by day for different battery capacity.

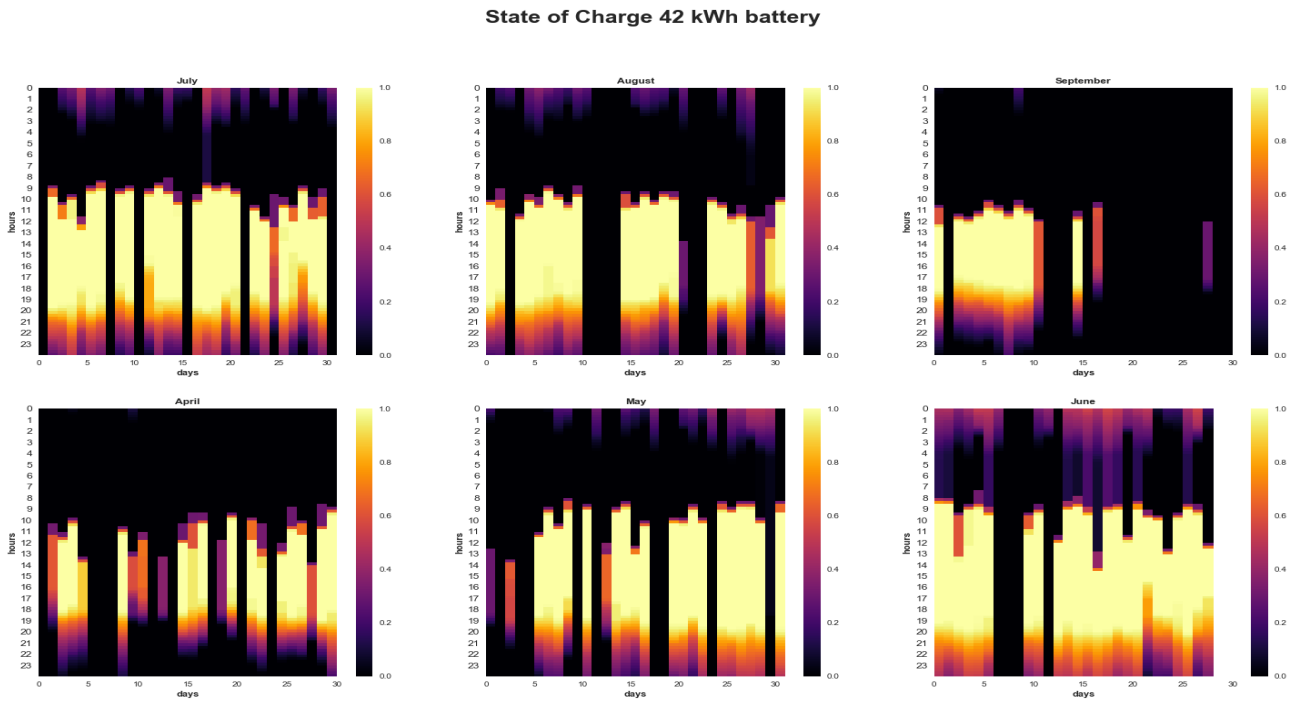

FIGURE 24: STATE OF CHARGE 42 KWH BATTERY SPRING AND SUMMER BAU SCENARIO

The presented heatmaps show the evolution of the battery charge in the relevant months for the analysis, on the x-axis are present the day of the months, on the y-axis the hour of the day, the color scale is the state of charge of the storage device, the brighter the higher the charge.

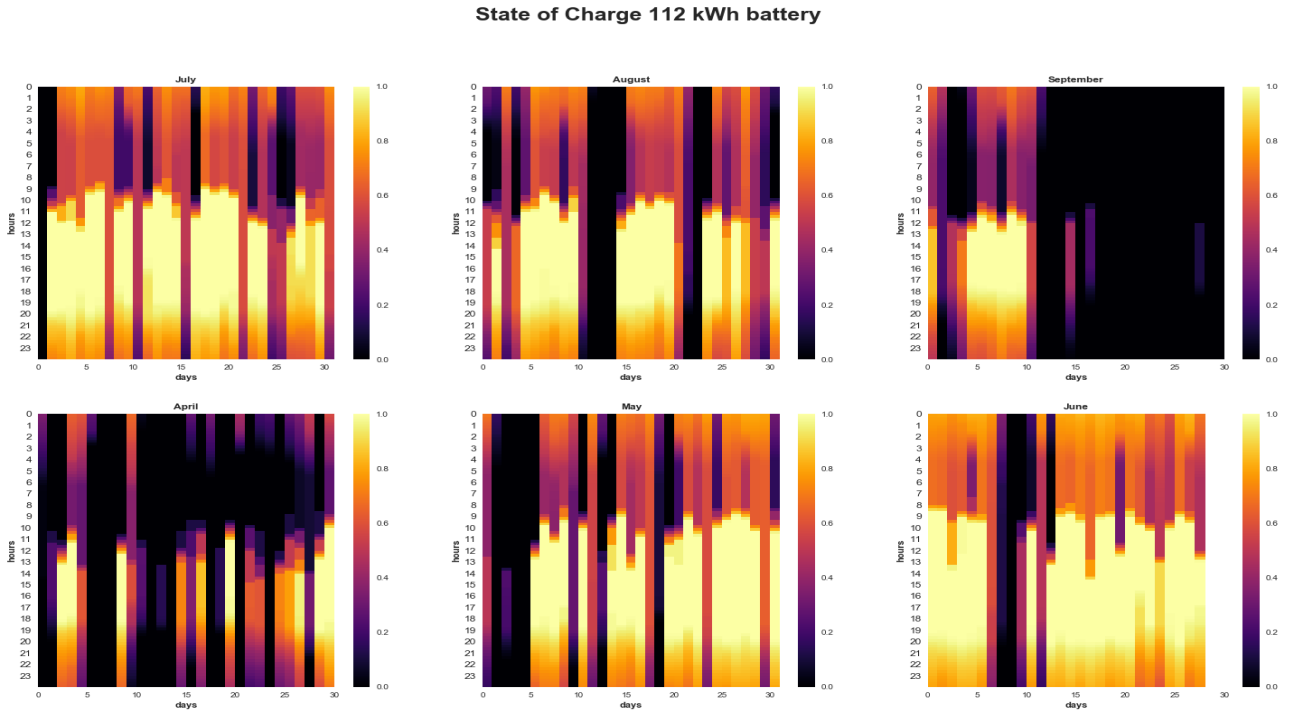

FIGURE 25: STATE OF CHARGE 112 KWH BATTERY SPRING AND SUMMER BAU SCENARIO

The large difference between production and consumption is the reason of the reduced effectiveness of bigger batteries. Two exemplary battery sizes are presented here, to see how much the state of charge of the battery is influenced by the total storage size. It is clear that, large batteries cannot fully discharge from one day to the other, hence their storage capacity cannot be fully exploited during curtailment hours. Considering the information about the state of charge of different size batteries and the curtailed energy plot it can be seen that in between 42 and 56 kWh lies the "sweet-spot" for storage capacity. Choosing to install batteries of this dimension allows to save 600 – 800 kWh/year.

It is also relevant to highlight how the sole installation of batteries cannot solve entirely the over-injection problem. Very large size batteries should be used to cancel curtailment completely, but their convenience would be very limited and the type of storage would be seasonal. Alternative storage solutions could be considered, but at the moment they are not completely reliable nor cost convenient.

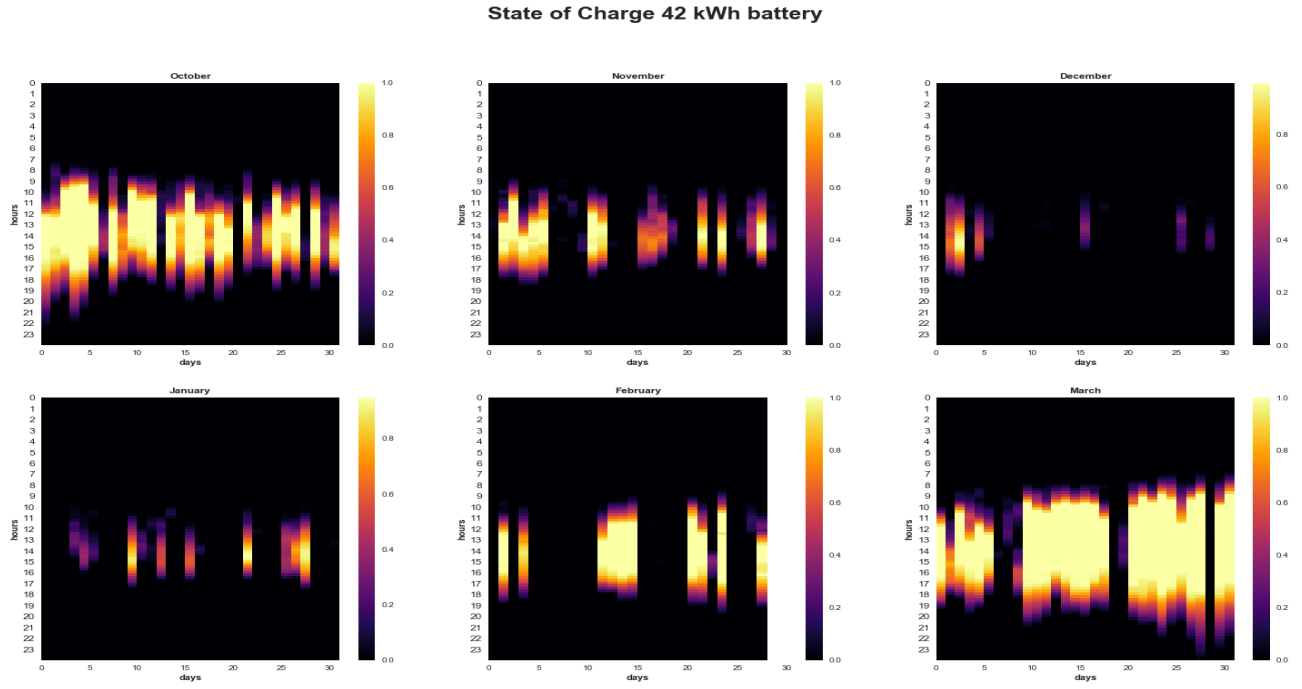

FIGURE 26: STATE OF CHARGE 42 KWH BATTERY FALL AND WINTER BAU SCENARIO

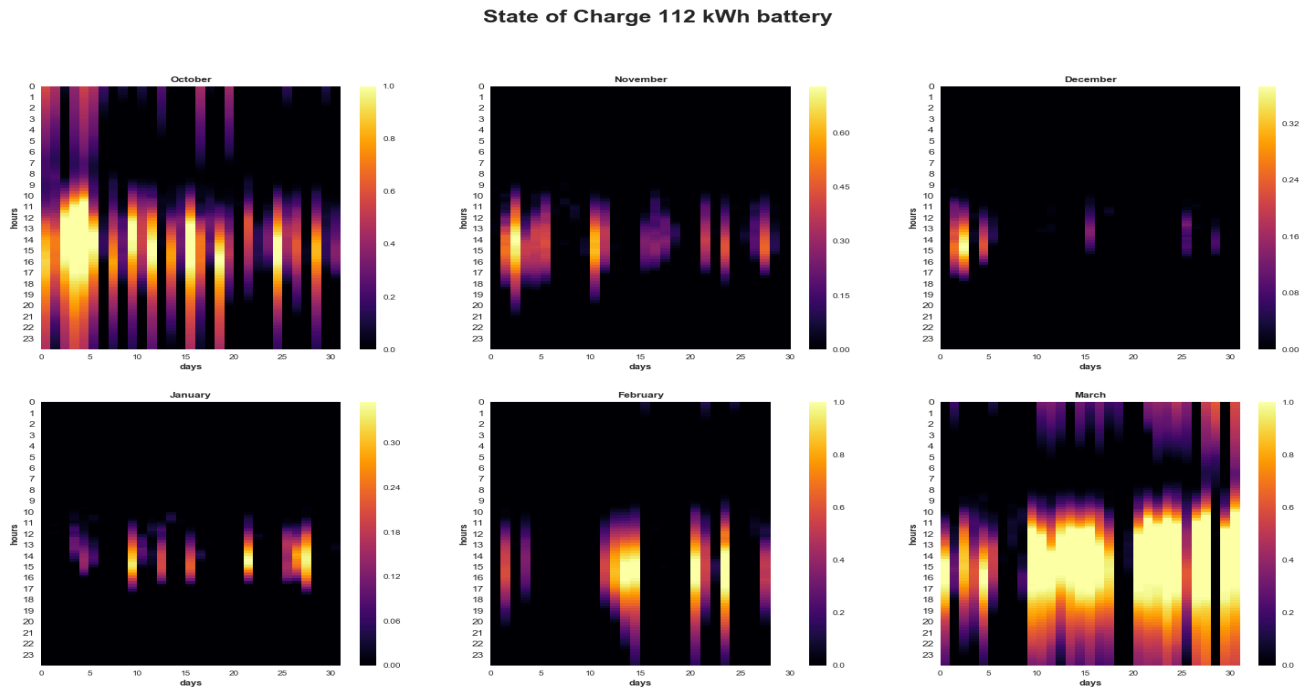

FIGURE 27: STATE OF CHARGE 112 KWH BATTERY FALL AND WINTER BAU SCENARIO

State of charge plots during winter months are shown, in this period batteries are used in the normal operating mode charging whenever generated energy is greater than the load. Large batteries struggle to reach the full storage capacity, particularly during December and January the state of charge of the devices is low due to the limited solar insolation and high load. While the main purpose of installing batteries in this study is to reduce the over-injection in the grid, some positive side-effects can be achieved as well. During Fall and partially Winter self-consumption can be boosted increasing the consumption of locally produced energy.

4.5.2 Battery and improved energy management

Similarly to the previous scenario, different battery sizes are considered, in addition to that the hot water production is shifted to the time range between 9AM and 3PM. Better results are expected compared to the BAU energy management, using energy to warm up water should reduce the amount of excess energy and helps to maintain some storage space in the batteries for longer.

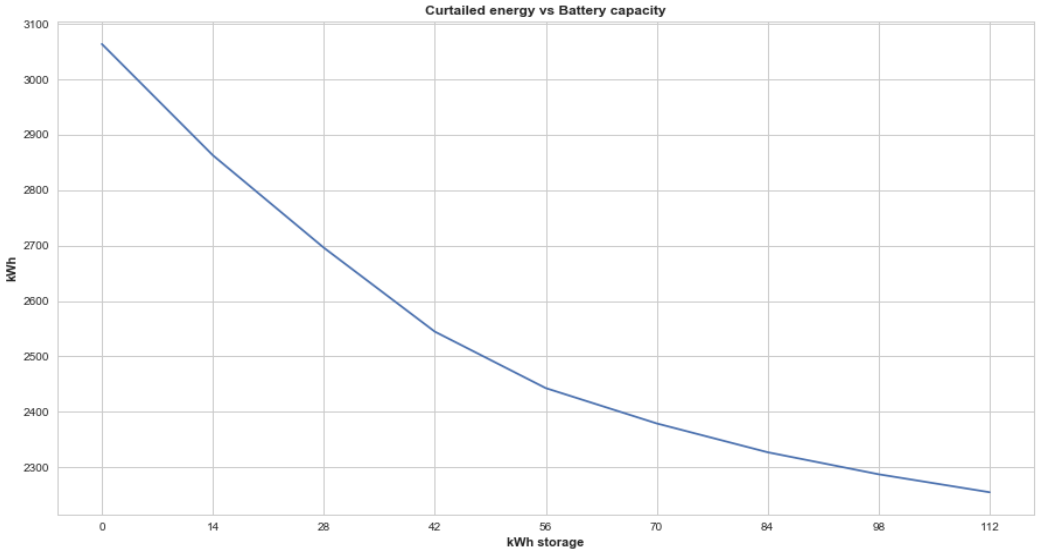

FIGURE 28: CURTAILED ENERGY AS A FUNCTION OF BATTERY SIZE, IMPROVED ENERGY MANAGEMENT

The hypothesis mentioned before is confirmed, the amount of curtailed energy is drastically reduced. That being said, the shape of the curve has not changed, meaning that batteries still get saturated during the day and part of the energy still gets lost.

FIGURE 29: STATE OF CHARGE 42 KWH BATTERY SPRING AND SUMMER WITH IMPROVED ENERGY MANAGEMENT

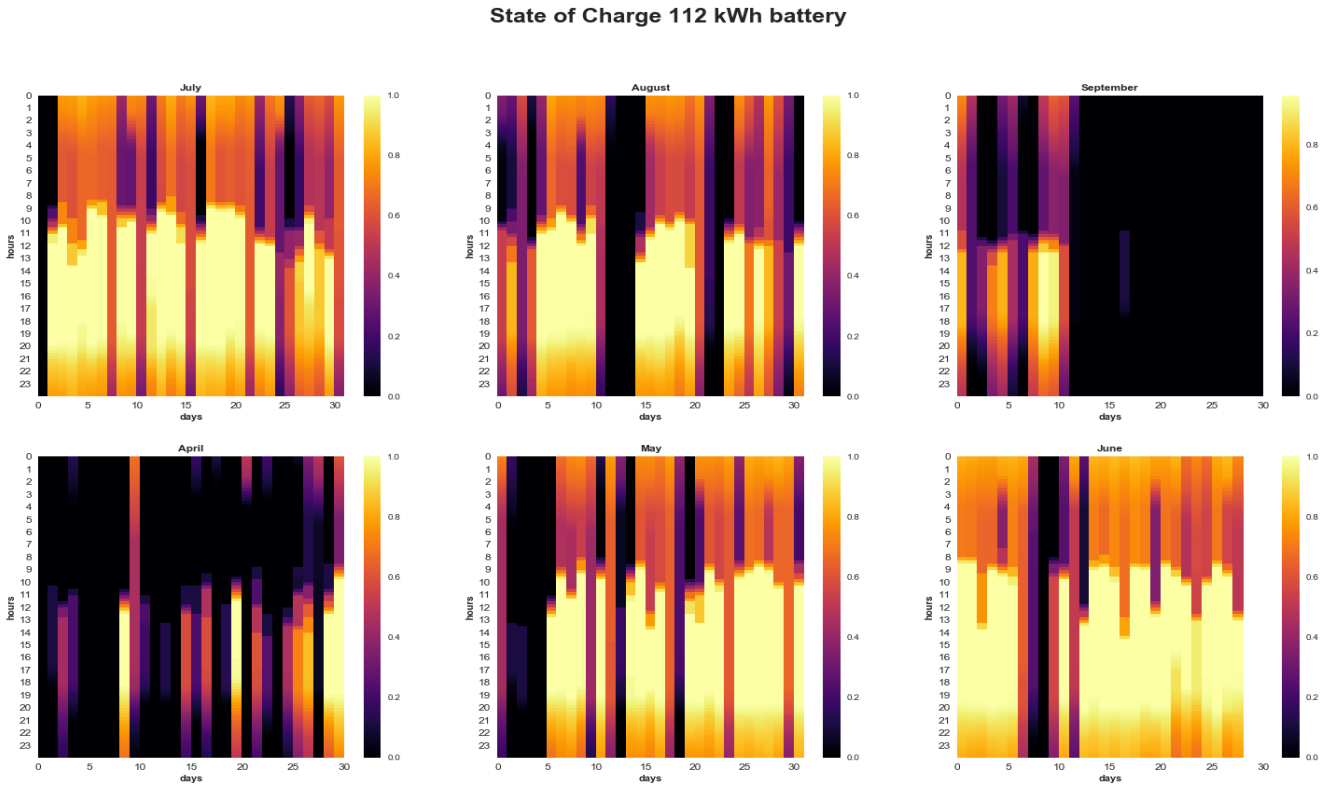

FIGURE 30: STATE OF CHARGE 112 BATTERY SPRING AND SUMMER WITH IMPROVED ENERGY MANAGEMENT

The state of charge heatmaps confirm that even though DSM is applied and part of the consumption is shifted to the central part of the day, batteries get full quickly and part of the energy needs to be curtailed. The heatmaps obtained using DSM are not very different from the one without it, because as shown before the controllable load is much lower than the fixed one, but looking at the yearly results it can be seen that shifting hot water production is highly beneficial.

The amount of produced energy is too high compared to the consumption, so no matter which solution is implemented the problem will not be solved entirely, but only reduced. A consideration would be to use smaller panels for the next installations or find some local consumer in the area that could make use of the extra energy. The houses analyzed are in the urban part of the city, so an example of big consumer could be some supermarket or some shops, another interesting alternative could be electric vehicles. The Netherlands has a very aggressive approach in terms of green mobility, electric cars offer an interesting possibility for energy storage through the big batteries that they are equipped with.

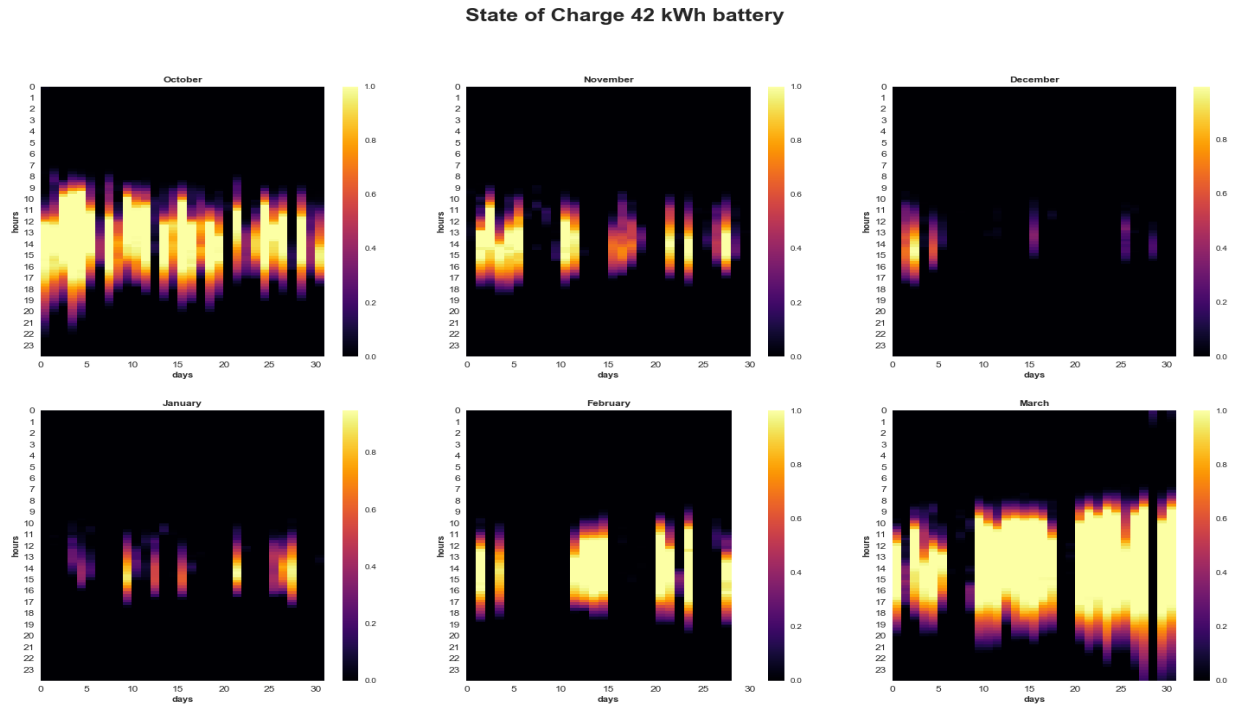

FIGURE 31: STATE OF CHARGE 42 KWH BATTERY FALL AND WINTER WITH IMPROVED ENERGY MANAGEMENT

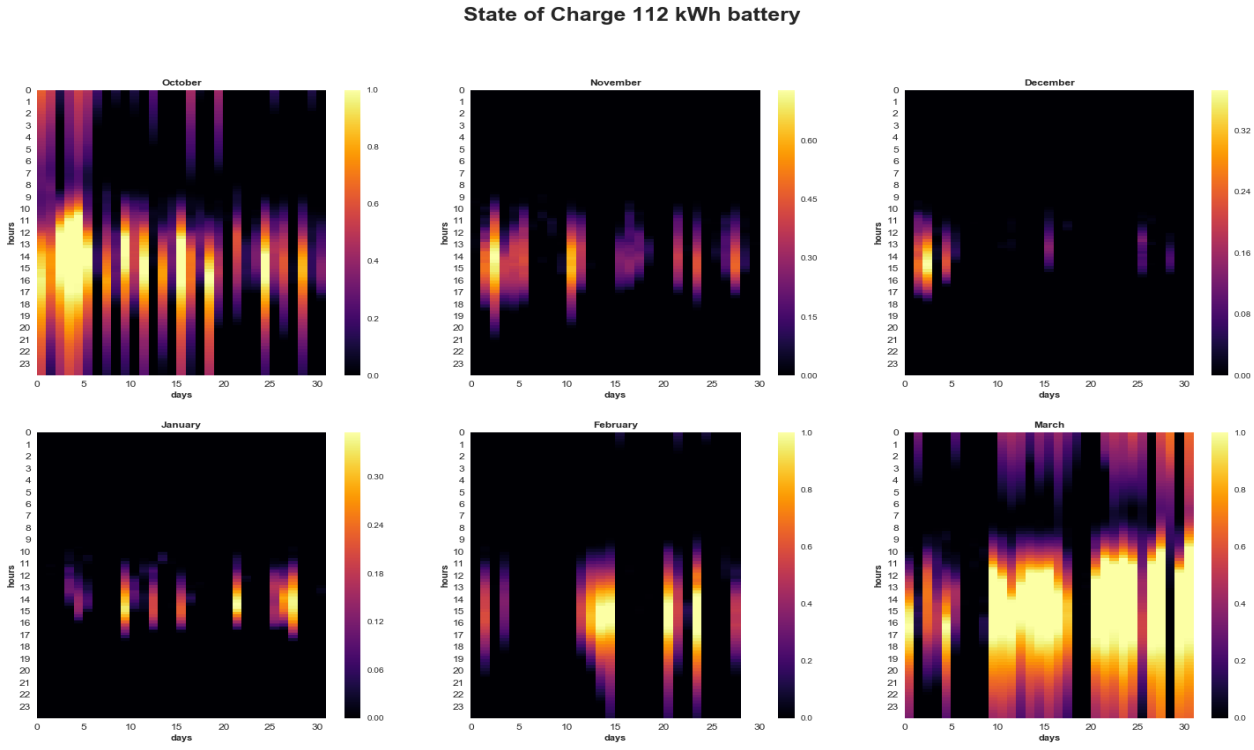

FIGURE 32: STATE OF CHARGE 112 KWH BATTERY FALL AND WINTER WITH IMPROVED ENERGY MANAGEMENT

As for the previous scenario, the state of charge of the batteries during Winter is analyzed. Big batteries still seem to struggle to reach full capacity during Winter months and they cannot be exploited completely.

4.5.3 Scenarios comparison

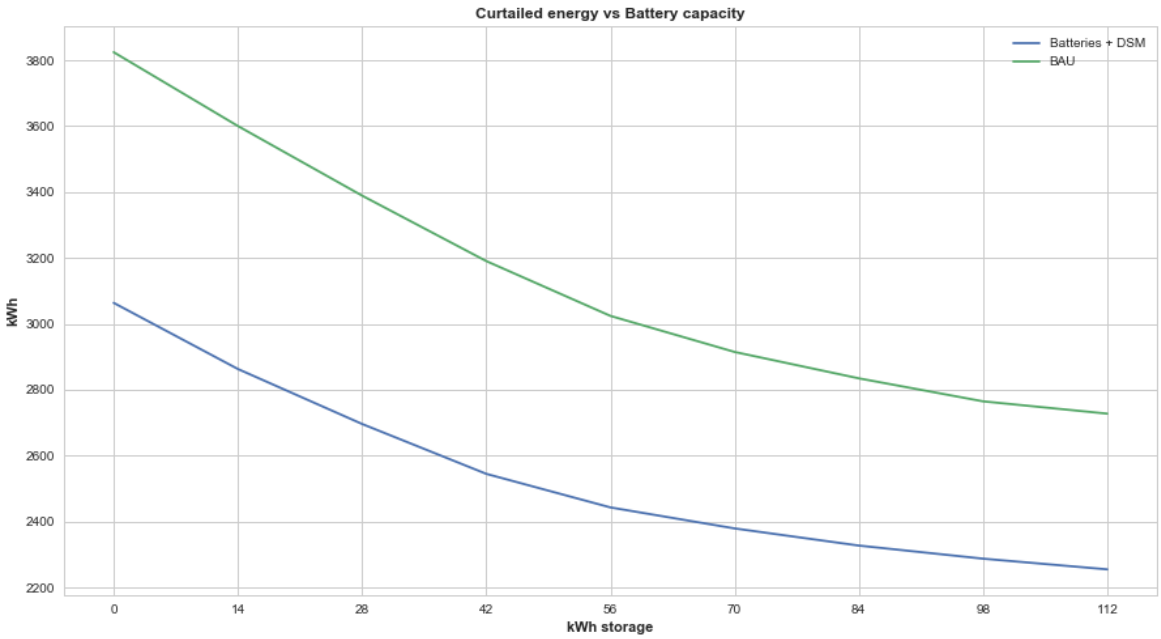

FIGURE 33: COMPARISON BAU AND IMPROVED ENERGY MANAGEMENT SCENARIO

Figure 33 helps to summarize the findings of the chapter, different scenarios are presented in the same picture. The green line shows the effect of installing batteries without modifying the energy consumption pattern. The blue line is the result of battery installation and load shifting. The effect of DSM can be observed measuring the distance between the two curves when the installed storage is zero, shifting hot water consumption saves between 700 – 800 kWh/year. As mentioned before, none of the strategies eliminates the over-injection problem, a quota of the energy is always curtailed. A very large amount of storage would be needed to eradicate the problem completely, but the solution would never be economical nor convenient from the practical point of view.

Chapter 5

Chapter 4 has proved that batteries and DSM are good solutions from the technical standpoint, now their economic feasibility is evaluated. Some assumptions regarding prices are needed to conduce the analysis. The main hypotheses, the calculation and the results are presented and explained in this chapter.

5.1 Key Economic Assumptions

Nowadays some storage solutions, like the Tesla Power-wall, cost around 500€/kWh [21]. Storage price is forecasted to decrease in the future, thanks to technological breakthroughs and mass production of the devices. To take into account future discounts on cost of batteries some estimates are used. Figure 34 shows estimates on the price of lithium-ion batteries, it comes from a study prepared for the Australian Energy Market Commission [5]. The price of electricity has been set to 0.2 €/kWh [22], of course this quantity is also subjected to variation through time, but no reliable estimates for the future have been found, hence it is kept constant.

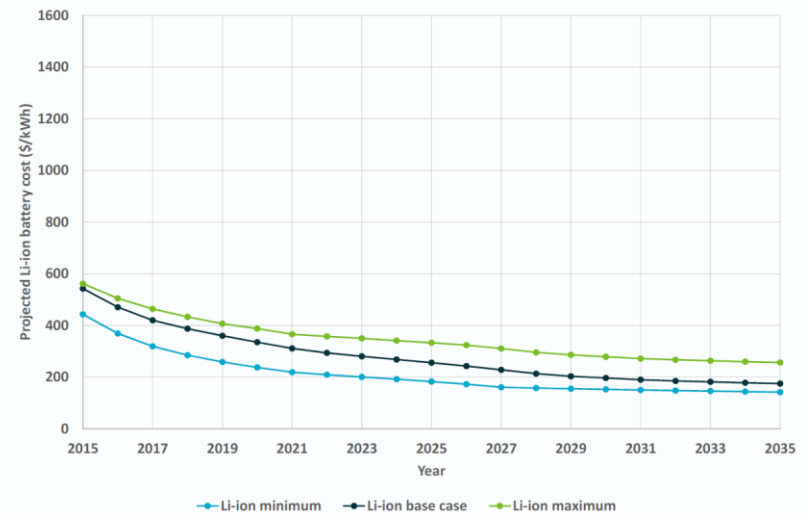

FIGURE 34: BATTERY PRICES FUTURE ESTIMATES [5]

The amount of curtailed energy is calculated using historical data, an underlying assumption is that this quantity will not vary significantly in the future. Considering the already advanced nature of the houses; each one of them has PV panels, heat pumps and hot water tank, it is reasonable to assume that things will not change greatly in the future. It is unlikely that more PV capacity will be installed in the coming years on these households. Nonetheless, the effect of increased/decreased curtailment is tested and compared to the reference scenario.

5.2 Economic calculation explanation

Different options are evaluated: installation of batteries plus load shifting, grid upgrade and no intervention. The fixed and variable costs of each scenario are considered and the total expenditures are actualized and compared on an equivalent time horizon.

Upgrading the grid is the most expensive solution in terms of capital cost. Due to their confidential nature, no public information about grid upgrade costs were found. Enervalis has no ongoing project with the DSO that manages the grid of the studied houses. An assumption on the cost of grid upgrade was formulated using the knowledge gained from the previous projects in which the company was involved. A cost ranging from 50 to 100 €/year per house for a time horizon of 40 years can serve as a good first guess. Further investigations should devote time to determine the exact value of grid upgrade cost.

Batteries useful lifetime is often set around 10 years thus, to compare it with grid upgrade it is necessary to evaluate the investment on a 40 years basis, batteries needs to be substituted every decade [9].

Grid upgrade advantage over batteries and curtailment is that reinforcing the grid, installing larger cables and bigger transformer, should allow to avoid curtailment entirely saving a lot of energy. Installing batteries as shown in chapter 4 helps reducing the amount of curtailed energy, but does not solve the problem entirely, some yearly losses due to wasted electricity will always be present.

Finally, doing nothing to solve the problem obviously does not cause any upfront costs but energy will be curtailed continuously and eventually the problem will become not negligible anymore.

Table 1 contains the electricity savings for different battery capacities and the related annual costs. The operator of the grid in fact is buying the electricity from the producers, but due to the limitation of the grid cannot transmit it entirely.

battery capacity [kWh]	electricity savings [kWh/year]	Avoided costs [€/year]
0	0	0
14	961	192.2
28	1128	225.6
42	1280	256
56	1381	276.2
70	1445	289
84	1497	299.4
98	1537	307.4
112	1570	314

TABLE 1: ENERGY CURTAILMENT REDUCTION

5.3 Economic calculation remainder

Scenarios with different capital and annual costs are compared. In order to build a fair comparison of the available alternatives all the costs need to be actualized using the following formula: [23]

$$NPV = \Sigma_{t-1}^{T} \frac{Cash\ Flow_t}{(1+i)^t} - \text{Initial Cash Investment}$$

Where "t" is the year at which the cash flow is earned, "i" is the interest rate of the investment this last value is set by the investor and influence significantly the results. More than one value for the interest rate is tested, to measure the sensitivity of the analysis to the parameter.

As mentioned before, installing batteries requires to buy and change the storage devices every ten years. Batteries are not bought and then stocked at year 0 of the simulation, since it is not a convenient solution from the financial point of view. It is much more economical to buy batteries only when they need to be changed, so that their actualized cost will result lower.

5.4 Economic results

5.4.1 Low interest rate scenario

Having all the necessary information it is possible to compute the actualized value of the investment. Figure 35 shows the actualized cash flows for all the available solutions over the 40 years' time horizon.

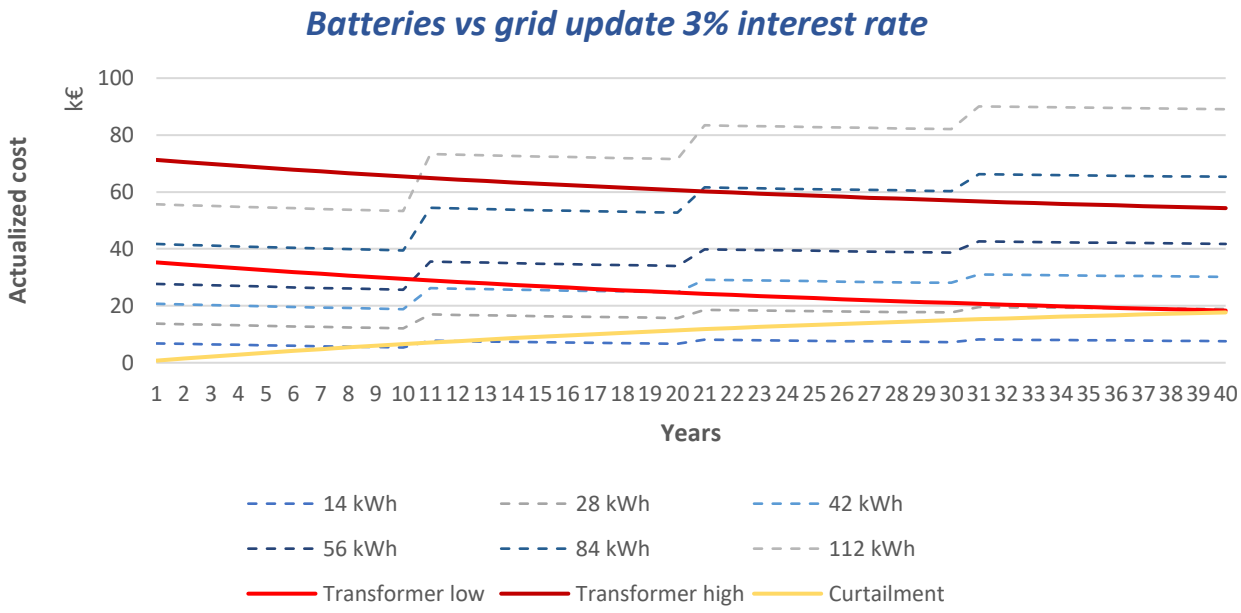

FIGURE 35: LOW INTEREST RATE ECONOMIC RESULTS

	[kWh]	NPC [€]		NPC [€]
battery cap	14	-7598.92087		
	28	-18868.4675	transformer low	-18321.8224
	42	-30207.3585		
	56	-41782.0202		
	70	-53527.7311	transformer high	-54321.8224
	84	-65328.9175		
	98	-77185.5794	curtailment	-17678.1776
	112	-89074.602		

TABLE 2: LOW INTEREST RATE ECONOMIC RESULTS

The x-axis represents years from 1 to 40, the y-axis is the cumulated actualized cost. To make the graph more readable only certain battery configuration are represented with the dashed lines. The two red lines are used for grid upgrades, the lower one refers to grid upgrade costs of 50€/year per house the other one to 100€/year, finally the yellow line is used for the option in which no measures are taken to face curtailment.

The bumps in the dashed lines correspond to the installation of new batteries, that is required every 10 years. The trend of the dashed and red lines is a downward one, since after the initial investment every year there are savings compared to the curtailment scenario. This last one on the other hand has an

upward trend, since yearly costs are associated to the constant waste of electricity due to saturation of the network.

A table with the final results is also provided, so that all the options can be compared more precisely. None of the option is able to reach a positive value, their actualized value is always negative. Of all the options, installing 14 kWh batteries seems as the most convenient one, 28 kWh batteries are slightly more expensive than grid upgrading when the lower cost range is applied.

5.4.2 Medium interest rate scenario

The interest rate is now set to 5%, the results are analyzed again to check the differences with the previous ones and investigate if some solutions have become convenient.

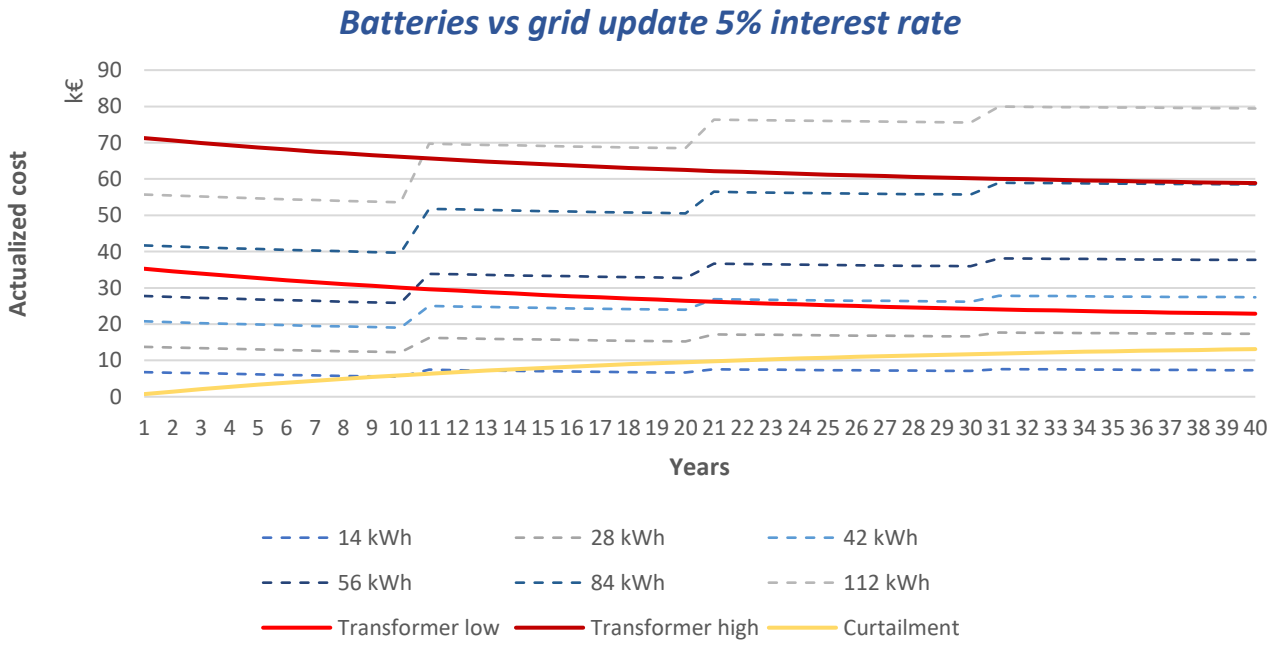

FIGURE 36: MEDIUM INTEREST RATE ECONOMIC RESULTS

	[kWh]	NPC [€]		NPC [€]
battery cap	14	-6484.25151		
	28	-16374.5817	transformer low	-19595.9134
	42	-26329.2584		
	56	-36502.7135		
	70	-46834.8901	transformer high	-55595.9134
	84	-57218.544		
	98	-67653.6751	curtailment	-16404.0866
	112	-78118.8347		

TABLE 3: MEDIUM INTEREST RATE ECONOMIC RESULTS

The final values of batteries and the no intervention strategies decreased, while grid upgrading costs substantially increased. Installing 14 kWh batteries is still the most convenient option, followed by non-intervention, 28 kWh are also more convenient than grid upgrading. It is also worth noticing that almost all the battery options are more convenient than grid upgrading if the highest price range is applied.

5.4.3 High interest rate scenario

One last value of 10% is tested for the interest rate, this value is quite unrealistic since it is fairly high, but to fully comprehend the effect of this parameter an extreme value needs to be observed.

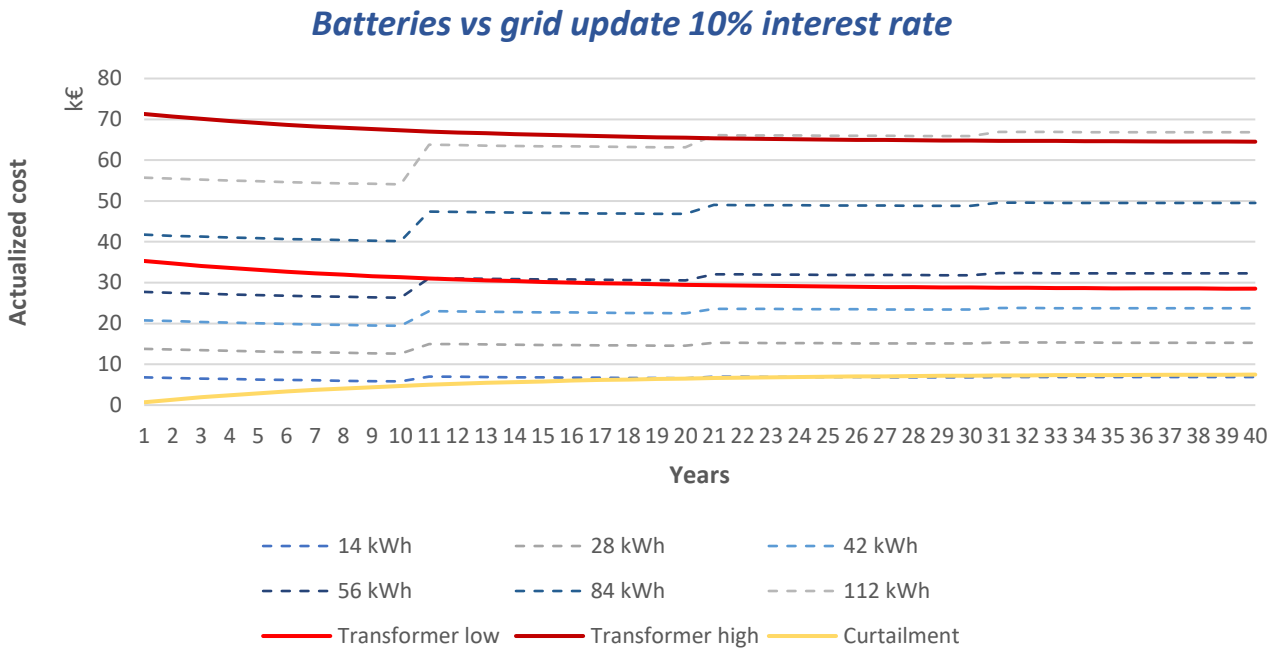

FIGURE 37: HIGH INTEREST RATE ECONOMIC RESULTS

	[kWh]	NPC [€]		NPC [€]
battery cap	14	-6386.966	transformer low	-26651.2275
	28	-14715.0736		
	42	-23079.8526	transformer high	-62651.2275
	56	-31569.3145		
	70	-40149.2326		
	84	-48758.4879	curtailment	-9348.77249
	98	-57397.0803		
	112	-66052.7861		

TABLE 4: HIGH INTEREST RATE ECONOMIC RESULTS

Battery NPC keep decreasing compared to the 3% and 5% scenarios, same behavior for the no-action strategy, while transformer values are getting higher. Once again 14 kWh storage seems like the most convenient option followed by curtailment and 28 kWh. It is interesting to notice that more and more storage can be installed and still remaining cheaper than upgrading the grid.

Besides varying the interest rate some other analyses are done, to verify the overall convenience of the investment and the importance of each factor. The uncertainty of each hypothesis need to assessed to understand how reliable the final results are and where more attention should be paid in order to improve the analysis.

5.5 Additional analysis

5.5.1 Battery price variation

The first parameter analyzed is the battery price, the goal of the calculation is to find the cost of batteries that results in a positive final value of the investment. In other words, the storage cost for which the NPC of each battery size turns to zero. *Table 5* shows the results:

	[kWh]	Battery price [€/kWh]
battery cap	14	136
	28	79.86
	42	60.41
	56	48.89
	70	40.92
	84	35.33
	98	31.09
	112	27.79

TABLE 5: BATTERY PRICE VARIATION SCENARIO

Batteries should be extremely cheap in order to turn the installation of batteries a convenient investment. The reason of such extreme results is related to the low value of electricity, while a considerable amount of energy can be saved using a small amount of batteries the economic return of this avoided cost is low compared to the high cost of the devices.

5.5.2 Electricity Price variation

Electricity price is the other main factor to investigate, the same procedure used for the batteries is followed and the results are shown below:

	[kWh]	Electricity price [€/kWh]
battery cap	14	0.5421
	28	0.9237
	42	1.2209
	56	1.5089
	70	1.8026
	84	2.0879
	98	2.3725
	112	2.6545

TABLE 6: ELECTRICITY PRICE VARIATION SCENARIO

The outcomes of this analysis are even more extreme than the one obtained in the battery study. Prices are well above reasonable values and they are unlikely to be seen in the future.

The previous two analyses clearly show that the investment in batteries for grid reinforcement is not convenient by itself in the analyzed situation. On the other hand, batteries are in certain condition the least expensive solution to cope with over-injection problems. When curtailment becomes excessive, actions need to be taken and batteries should definitely be taken into account.

5.5.3 Curtailed electricity variation

It is also important to evaluate what would happen in the case of decreased over-injection, that means less energy needs to be curtailed. This evaluation is needed because the data used is the historical and referred to a single year.

	[kWh]	NPC [€]		NPC [€]
battery cap	14	-7598.92087		
	28	-18868.4675	transformer low	-18321.8224
	42	-30207.3585		
	56	-41782.0202		
	70	-53527.7311	transformer high	-54321.8224
	84	-65328.9175		
	98	-77185.5794	curtailment	-17678.1776
	112	-89074.602		

TABLE 7: REFERENCE CASE CURTAILED ENERGY SCENARIO

	[kWh]	NPC [€]		NPC [€]
battery cap	14	-7154.65496		
	28	-18346.9983	transformer low	-16554.0046
	42	-29615.6204		
	56	-41143.5902		
	70	-52859.7142	transformer high	-52554.0046
	84	-64636.8613		
	98	-76475.0313	curtailment	-19445.9954
	112	-88348.7981		

TABLE 8: INCREASED CURTAILMENT (+10%) SCENARIO

	[kWh]	NPC [€]		NPC [€]
battery cap	14	-8043.18679		
	28	-19389.9368	transformer low	-20089.6401
	42	-30799.0967		
	56	-42420.4502		
	70	-54195.748	transformer high	-56089.6401
	84	-66020.9738		
	98	-77896.1275	curtailment	-15910.3598
	112	-89800.4058		

TABLE 9: DECREASED CURTAILMENT (-10%) SCENARIO

In the event of increased curtailment upgrading the grid and installing batteries become more convenient. The opposite happens when the amount of curtailed energy decreases, since the final cost of not acting to solve the problem is lower. Installing small batteries still seems the best course of action.

5.5.4 Non-economic considerations

The previous economic calculations help to determine the profitability of the different solutions, although some strategical considerations should be taken into account. Scalability, time needed to deploy the chosen solution, alignment with the future policies of the country and many other factors are to be considered.

Batteries are a very flexible solution, no major intervention on the network is needed. The main flaw of storage is the cost of the solution and its short lifetime compared to grid upgrade, but they are also easy and quick to install. Scalability is a point in favor of storage if more capacity is needed additional batteries can be installed without any major problem. Also, it is important to define what should be the final goal of the intervention on the grid, batteries are not able to eliminate curtailment entirely.

Grid upgrading is the solution to apply if the goal is to avoid curtailment entirely, but as the economic analysis revealed is a quite costly option. In addition to the high costs, grid reinforcement requires many resources (materials, working hours, etc.) and planning. Moreover, it is difficult to guarantee the continuity of the system during the interventions on the network. Finally, grid upgrading is not resilient to unexpected changes of the amount of energy to curtail in the future, the only way to cope with them is to oversize the system.

Chapter 6

Previous analyses were made using historical data only, but for an online running system estimates of the future generation and consumption are needed; so that the management of the grid can be planned in advance. The sixth chapter is devoted to the implementation of machine learning algorithms that can help in the task. First, a quick theoretical explanation of the algorithms is provided, then their usage is shown and the results analyzed.

6.1 Quick introduction to Machine Learning

Machine learning has been defined by Arthur Samuel as: “Field of study that gives computers the ability to learn without being explicitly programmed”.

Machine learning is part of the broader study field named as artificial intelligence (AI). The aim is to obtain accurate predictions from software that has not been programmed explicitly. Three main type of machine learning’s tasks can be identified:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

The desired kind of outcome determines the algorithm to use. A supervised learning problem is one in which the input data contains “labels”, the program receives a set containing input and output data. The algorithm is supposed to map the relationship between input variables and the output and then use the learned function to predict the value of new examples. Supervised learning tasks are further divided into:

- Regression
- Classification

The division is made on the type of output, in a regression problem a continuous output field is expected, for example predicting the price of a house given information such as its area, position, number of rooms, etc. is a regression task. Not all the supervised problems output a continuous outcome, many times the predicted variable is discrete and belongs to a finite number of classes. A typical example of classification problem is determining whether a mail is spam or not given some information about it.

Unsupervised learning uses unlabeled data, the algorithm receives only a set of input information, the objective is to identify clusters in the data.

Finally, in reinforcement learning the goal is to take a series of decisions over time, the algorithm needs to keep track of the past decisions and evaluate their effectiveness in reaching the final goal of the task.

Machine learning techniques are used in this thesis to predict the generated and consumed electricity one-day ahead using historical and weather information. It is a supervised regression task, labeled data is provided to the algorithm and the expected outcome is continuous. Different algorithms, such as ARIMA models, linear regression and random forest are used to verify which work best.

6.2 ARIMA models

ARIMA models are a class of mathematical models that can be used to analyze and predict future values of a data distribution. These models work as filters, that remove noise from the underlying information and use it to predict future values. Time-series data is required for this kind of analysis. Data should be collected regularly over time, information should be registered at a constant rate and no measurement should be missing. The sampling rate can be very short, fractions of seconds up to very large time interval such as years, but it has to remain constant over time.

Data needs to be stationary to fit an ARIMA model on it. A time-series is said stationary if its statistical properties, such as mean, variance, autocorrelation, etc. are constant over time, if not its future values would be unpredictable [24]. Quite often data is not stationary, thus before fitting an ARIMA models some transformation are needed. Sometimes data oscillates around a trend-line, identifying it and subtracting it may help stabilizing the series, this is the case for trend-stationary series.

Trend removal is not always enough, some additional differencing may be needed. Instead of studying the actual values of the series the difference between consecutive observations may be considered.

$$y'_t = y_t - y_{t-1}$$

$$y''_t = y'_t - y'_{t-1} = y_t - 2y_{t-1} + y_{t-2}$$

The two equations above show how to apply first and second order difference. Using first order difference, the variation between the current and the previous value is considered. Second order difference applies differentiation on the results of the first order method. It is possible to continue with the differentiation process, but rarely more than two orders are needed. The last differencing scheme worth to mention is seasonal difference. In this case the difference is not between subsequent values, but values separated by one season. Seasons within data are defined by the presence of cyclical patterns, that are not necessarily the calendar seasons. The equation below shows how to apply seasonal differencing; "*m*" is the number of seasons to consider.

$$y'_t = y_t - y_{t-m}$$

6.2.1 Autoregressive models

These models are described by the following equation:

$$y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + e_t$$

Where c is a constant, e_t is white noise: random variation with zero mean and finite variance characterized by absence of remaining correlation and the coefficients φ are the parameters to vary to fit the model on the data. Autoregressive models use a linear combination of the previous lags in the signal to predict the future ones, the prefix "auto" is added because the regression is on values of the time-series itself. The subscript "*p*" identifies the last relevant lag to consider, autoregressive models are often identified using the short notation *AR(p)* [25].

6.2.2 Moving average models

$$y_t = c + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}$$

Moving average models apply regression on past forecast errors, that in the formula above are indicated with the letter "e", the only exception is the first term e_t that is white noise. Similarly to autoregressive models, moving averages are identified by the number of lags to consider, indicated by the parameter "*q*". Short notation for this class of models is *MA(q)* [26].

6.2.3 ARIMA and model identification procedure

Many times, complex phenomena cannot be captured by pure AR, MA models. ARIMA is a combination of AR and MA with the addition of the integrated term "*I*". These models are defined by three parameters *(p, d, q)* the first and the last one are explained in the previous two paragraphs, "*d*" refers to the order of differencing that is needed to make the series stationary.

More complex models can be used, such as the so-called SARIMA. They work exactly like normal ARIMA models, with the difference that seasonal and non-seasonal components are treated separately. SARIMA models require the identification of six parameters *(p, d, q)* and *(P, D, Q)*. The first three refer to the more recent non-seasonal values, whereas the ones written in capital letters refer to the seasonal components. Finally, the effect of external factors can be included using ARIMAX models.

Determining all the necessary parameters is the task that the user of ARIMA models is required to accomplish, a standardized procedure can be followed. The "Box-Jenkins" method is an iterative procedure divided in three phases:

- Model identification and selection
- Parameter estimation
- Model checking

The first step makes sure that the time-series is stationary, if not differencing should be applied as much as necessary. Autocorrelation and partial autocorrelation plot are useful for this task.

Parameter estimation consists in the identification of the multipliers, φ and ϑ, that appear in the AR

and MA formulae. The model is applied on the given data and the associated error is minimized, different criterion can be applied, such as maximum likelihood or non-linear least-squares estimation.

Model checking is the last step in the process, once the model has been fit on the data it is necessary to control the residuals. They should be independent from each other, if not some information is still contained in them, moreover their mean and variance should not change in time. Being the procedure iterative, failing to respect the requirement of any phase means that another iteration is needed.

While the procedure is straightforward trial and error is part of the process, correlation plots helps a lot in the model identification, but they are not always easy to interpret. Given a set of data more than one model could correctly fit it, the criterion to choose between multiple valid models is to pick the simplest one, which is the one that utilizes the lower number of parameters [27].

6.3 Linear regression

Linear regression is probably the simplest machine learning algorithm available to extract information from data. A set of independent and dependent variables is analyzed, the goal of the procedure is to determine the underlying relation that links the independent to the dependent variables. Below is presented the typical equation used in a linear regression task:

$$Y_t = c + a_1 X_{1t} + a_2 X_{2t} + \cdots + a_k X_{kt}$$

Where "*c*" is a constant and "*a*" are the parameters to determine to fit the model on the input data.

An infinite number of lines could fit the data, in order to choose the best one an error function is defined and its value minimized iteratively, updating the parameters of the regression equation.

$$Q = \Sigma_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The expression above is the "cost function", the distance between each point and the value predicted by the linear model is squared and then summed. The cost function returns a scalar value, several optimization techniques can be used to minimize it, the most common one is gradient descent [28].

6.4 Random forest

Random forest is one of the most effective machine learning algorithm, it provides high quality results while maintaining a relatively simple implementation and good readability of the output. This algorithm belongs to the ensemble methods class. The predictions of several base estimators are combined to improve generalizability of the model [29].

Random forest is based on decision trees, a simpler machine learning algorithm. To understand how random forest works is first necessary to discuss decision trees. Figure 38 shows how the output of a decision tree may look like. The algorithm is named decision tree, because of its resemblance to a tree. The block at the top of the picture is called root, data is fed from the roots and subsequently split into interior nodes until when the bottom blocks are reached, these are called "leaves". In the case of regression trees the principle used for splitting the data is the variance reduction. The variable on which data is split is the one that guarantees the maximum reduction in the variance of the target variable at the current step [30].

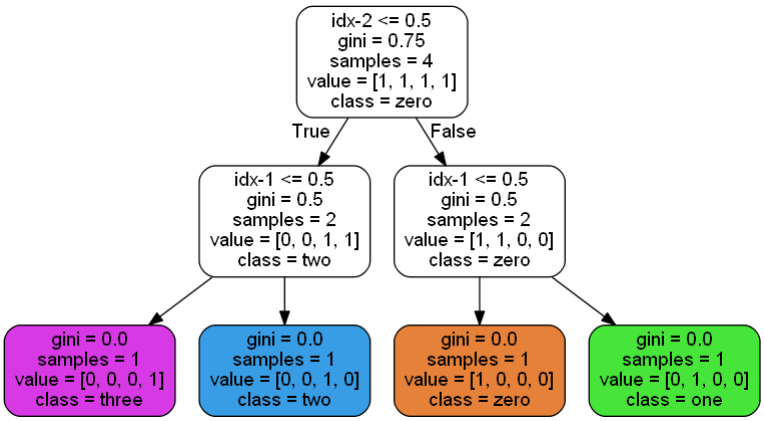

FIGURE 38: DECISION TREE STRUCTURE EXAMPLE [31]

Decision trees are a very convenient data mining technique, as the figure shows the outcome is highly readable. Many information is contained in a single picture, such as the importance of the variable and how the data is split. Another advantage is the possibility to perform very quick predictions on new data, following the flow-chart is easy to determine the output of new examples. The drawbacks of the algorithms are numerous, the accuracy is not as high as the one of more complex techniques, the robustness is limited, small changes in the training data can alter significantly the structure of the tree and its predictions, but the main problem is the tendency to overfit. Overfit happens when a learning algorithm generates an excessively complex model to fit the training data, resulting in a loss of the generalization power of the model and lower quality predictions on unseen data.

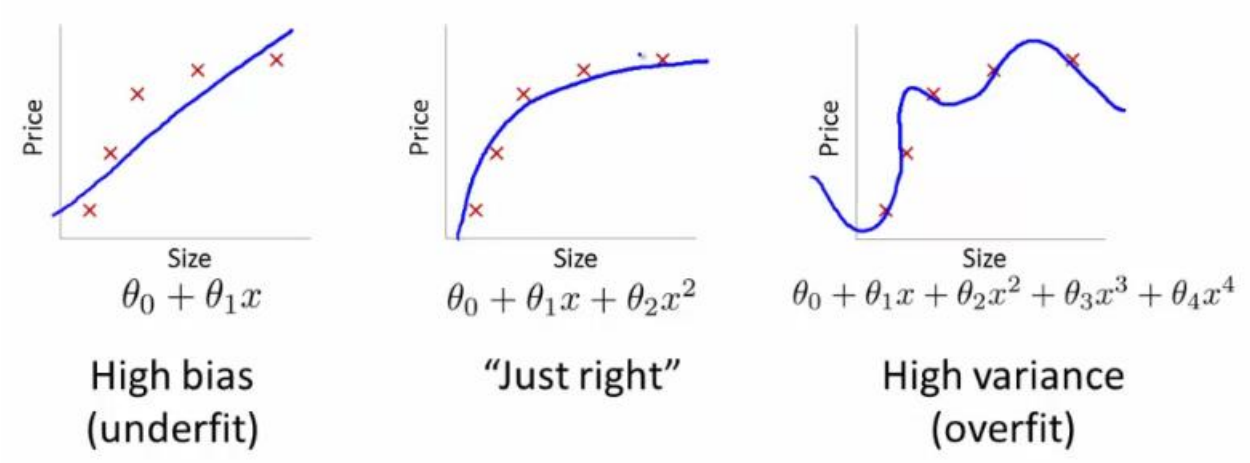

FIGURE 39: UNDERFITTING AND OVERFITTING EXAMPLES [32]

Random forest is an attempt to solve the problems that affect decision trees, while preserving their advantages. The idea behind random forest is to use a high number of decision trees and then average their outcomes, single trees have the tendency to overfit, growing very deep in the attempt to explain the behavior of some irregular patterns. When training a random forest just a part of the training data is used for each tree, a random subset is utilized and the results of all the different trees are aggregated by averaging their outcomes. While a single tree is sensitive to noise in the training data, the average of several trees is more resilient to noise disturbance. The random selection of the training subset for

each tree is a crucial step, because it ensures low correlation between the trees, if they were highly correlated the averaging process would be ineffective. In addition to this procedure, another degree of randomness is introduced in the algorithm: the feature subspace is selected randomly to ensure that very significant features are not over represented in the model. One last randomness step can be also applied by making the splitting process random as well, this is the case for Extra randomized trees, but this technique is not used in this thesis [33].

6.5 Input data

To predict load consumption, the only data necessary is the historical consumption for a sufficient long time. Good quality predictions can be obtained only using high quality data, that means that the sole consumption is not enough to feed the model, augmentation is needed.

The augmentation process is typically referred to as “feature engineering”, the objective is to use empirical knowledge of the data to extract additional information manipulating the raw data. Here are the added inputs for the model:

- Previous four lags for load consumption
- Previous four days’ lags for load consumption
- Mean of the previous three days
- Mean of the previous three weeks
- Mean of the previous day

These added features aim to capture the cyclic behavior of load consumption. Checking the correlation degree between the input data and the output is the best way to verify the quality of the chosen variables.

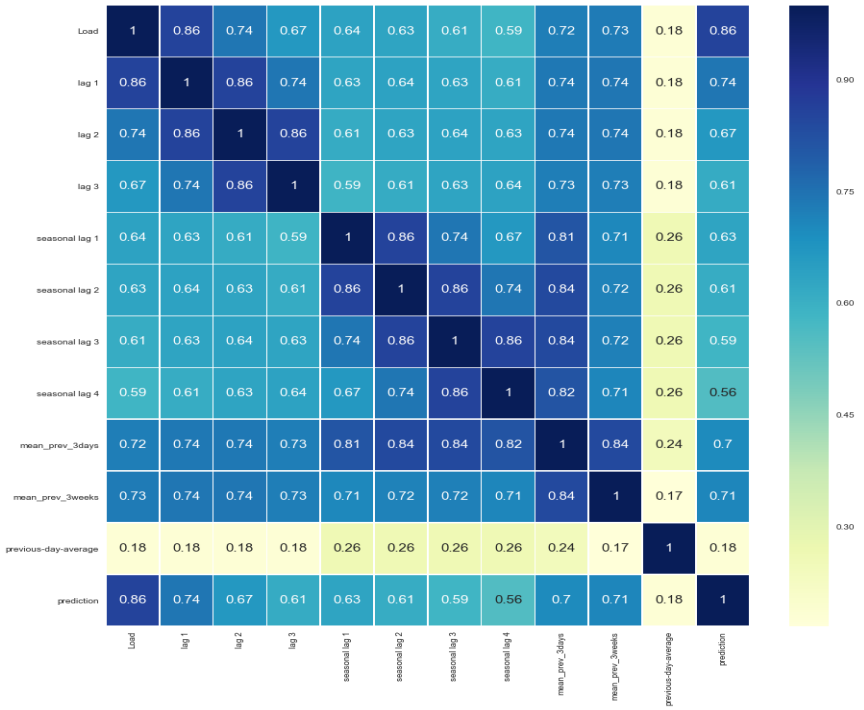

FIGURE 40: LOAD CONSUMPTION CORRELATION MATRIX

The correlation matrix shows how much features are related to each other, the range of the values is comprised between zero and one, the closer the value of a cell is to one, the darker the color. The cells named as “prediction” contain the value that the algorithm is supposed to predict using all the others. A significant feature has a high correlation with the “prediction”. The last column and the last row are the ones to focus on. As it is clear from figure 41, most of the features have a high degree of correlation with the predicted value. It is important to remember that correlation does not imply causation. That means that two variables might be highly correlated due to chance, but there is no causal link. Here, it is safe to assume that correlation also implies causation, the features are created from historical data of the load consumption and the old patterns are expected to be seen again in the future.

A similar procedure is applied in the preparation of PV forecasts input data, historical data is gathered, but due to the influence of meteorological factors on the PV production some additional information are used. Weather data is downloaded from “forecast.io”, a website that offers access to open -data.

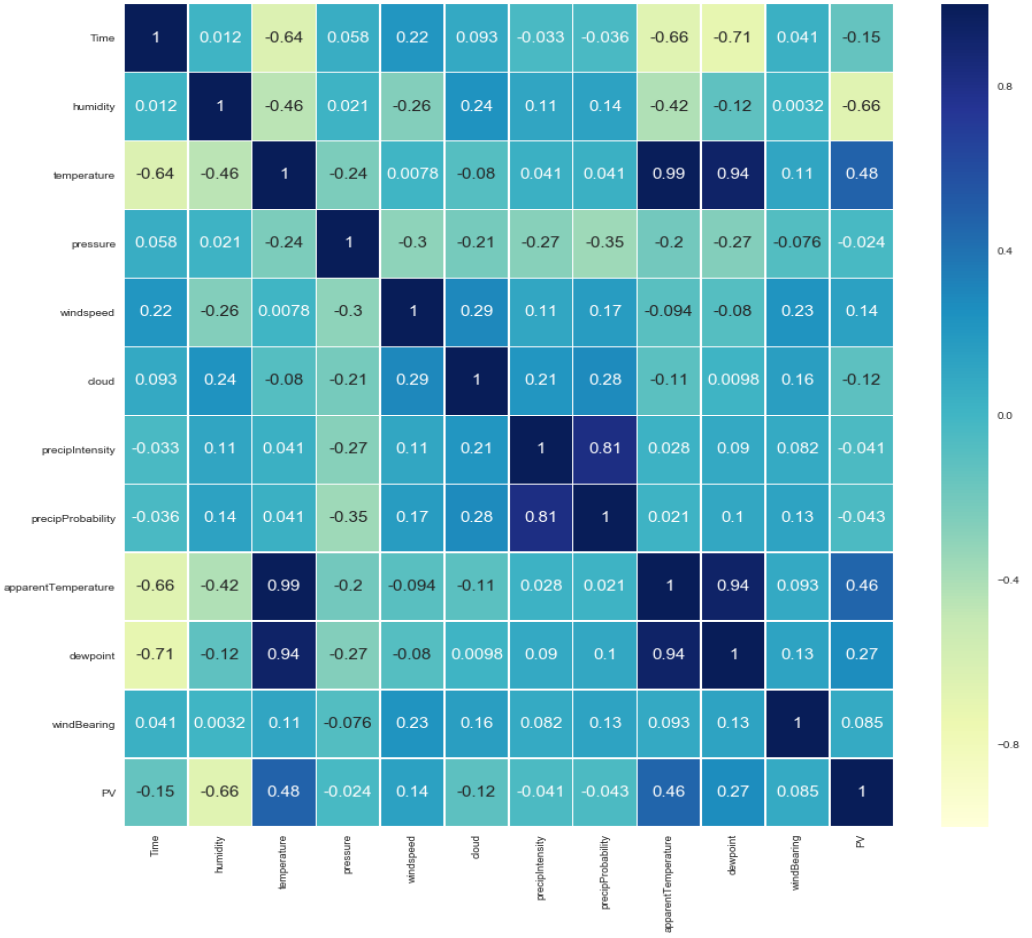

FIGURE 41: WEATHER DATA CORRELATION MATRIX

The selection of weather variable contains many features that are not so highly correlated to the PV production, moreover some of them are repetition of other variables, such as apparent temperature, dew-point and temperature or precipitation intensity and probability. It is important to remove repeated variables as well as the one that are not highly correlated to the PV production, otherwise the performance of the model could deteriorate.

Variables are filtered, keeping only the relevant ones: humidity, temperature and cloud coverage. It is well known that PV production is highly influenced by the irradiance, thus this is included in the model. Data about solar irradiance is collected from the Dutch Weather Service, the closest station to Delfzijl is situated 23 km south of it, in Nieuw Beertha.

FIGURE 42: MAP OF DELFZJIL AND NIEUW BEERTHA

As for the load input data, the historical measurements are included as additional features. A correlation matrix is shown below, all the chosen features are higly correlated to PV production and considering that they are all past production and weather data they can be considered good predictors of the photovoltaique generation.

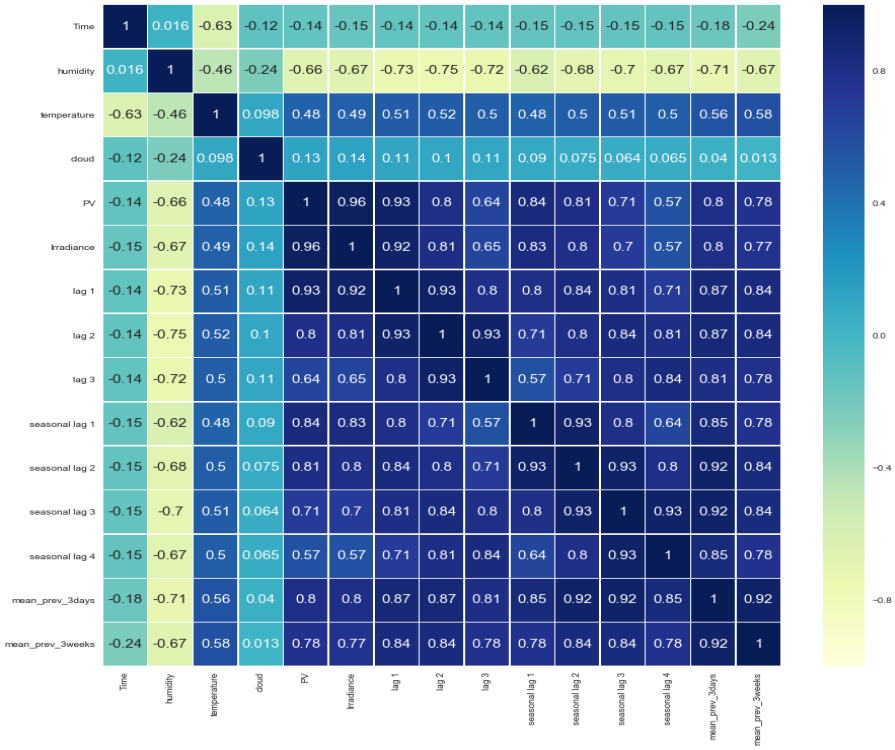

FIGURE 43: PV DATA CORRELATION MATRIX

6.6 Model implementation

The first action needed for the implementation of the models is to divide the dataset into two parts: the training and the test set. The purpose of having a test set is to verify the performance of the trained model, collecting information on its performances. Data from July 2016 to April 2017 is used for training the algorithms, performances are tested on the entire month of May. It is important to have a sufficiently long test period, to capture sunny and rainy days.

The goal of this implementation is to generate forecasts one-day ahead with quarterly hour resolution, for load consumption and hourly prediction for PV generation. Random forest and linear regression can output only one predicted value for every run, that means that the forecasts for an entire day need to be generated using a "rolling window" procedure. The output of the first run is used as starting point for the prediction of the very next time-step and repeating it until the forecast for an entire day are generated. This kind of technique is reliable only for short-term predictions, since the error of one forecast propagates, influencing the following ones.

Linear regression and random forest algorithms are pre-built in the "scikit-learn" repository, written in Python language; ARIMA models are available on different libraries implemented in R. The user is not requested to write the algorithms from scratch, their open-source version can be used. The tasks to take care off are preparing the input data, through feature engineering and selection, testing the different algorithms, tinkering with their regulation parameters and organize and analyze the outputs. Each model has different values, called hyperparameters, that need regulation to make the algorithm work at the best of its possibilities. A separate discussion about every model is presented below.

6.6.1 Linear regression

This is the easiest model to prepare, the algorithm is straightforward, the only type of regulation from the user is normalization of the data. Linear regression typically needs regularization of the input data, when the order of magnitude of the input numbers is not the same, the coefficients of the equation are heavily distorted. To avoid this problem, all the input values are taken and divided by the maximum value that the variable registered in the training data, so that all the measures are comprised between zero and one. Once the predictions are generated it is possible to invert the normalization and obtain the actual value of the forecast.

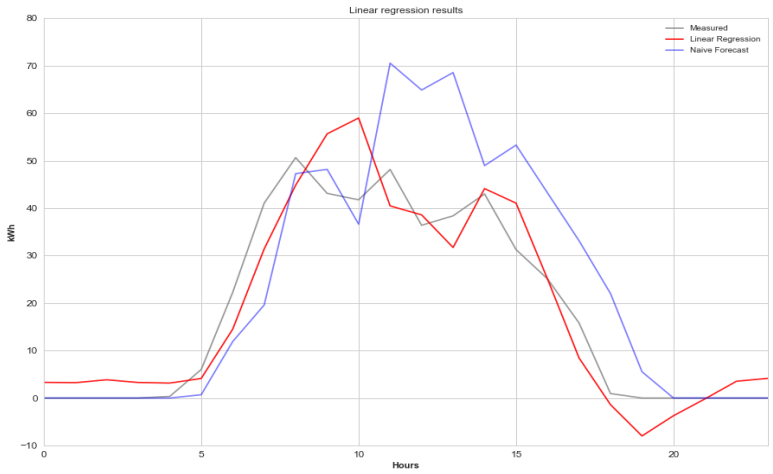

FIGURE 44: LINEAR REGRESSION OUTPUT PV PRODUCTION

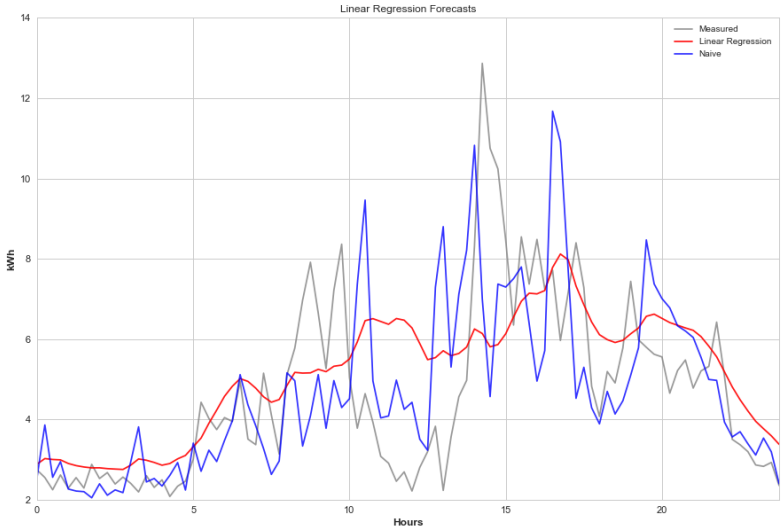

FIGURE 45: LINEAR REGRESSION OUTPUT LOAD CONSUMPTION

The pictures above are a graphic example of how the output of the algorithm looks like. Linear regression can capture the general trends in the data, but it is evident that the forecasts are not completely satisfactory, especially around noon.

6.6.2 Random forest

For this model adjustment of some hyperparameters is required: the number of trees to train and the maximum number of features fed to each tree. The higher the number of trees the better the results of the model, but computation time will also increase. The maximum number of features used should always be lower than the total amount of features, otherwise the influence of the variables with a high correlation index would affect excessively the output of the trees. A practical rule is to take the total number of features and use only a third of it.

Hyperparameter optimization is a trade-off problem, accuracy of the model and its complexity have to be considered. A large set of possible value was tested and the results compared and lead to choose 100 trees and 6 features per tree.

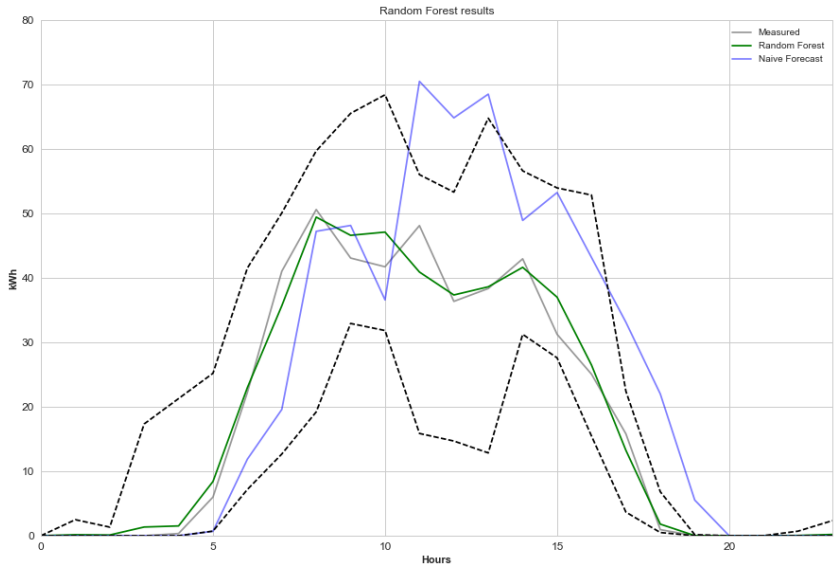

FIGURE 46: RANDOM FOREST OUTPUT PV PRODUCTION

The output of the random forest includes two black lines that delimit the confidence interval of the results. The prediction of each tree is gathered in a distribution, of which the 95 quantile is selected to create the confidence interval.

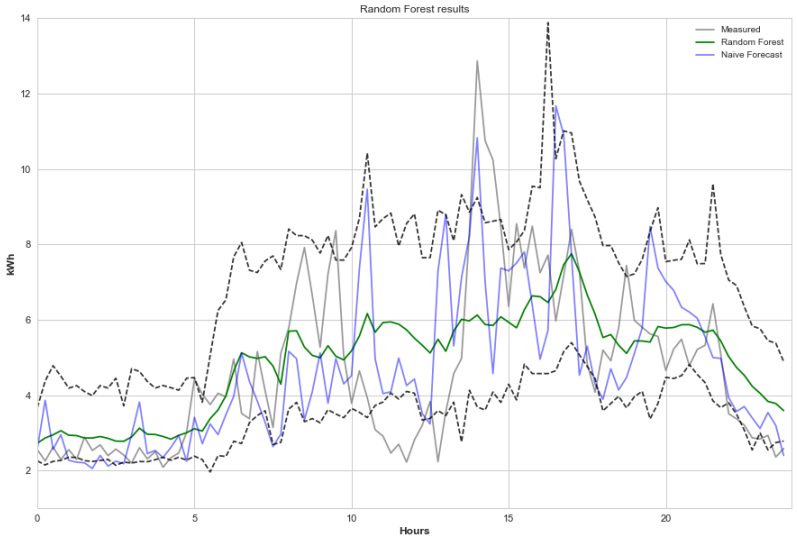

FIGURE 47: RANDOM FOREST OUTPUT LOAD CONSUMPTION

A first visual analysis suggests that random forest's predictions are better than the linear regression ones. The model is much more complex and its ability to represent non-linear relationship is higher than the one of a purely linear algorithm. Also, load is difficult to predict accurately due to the high sampling rate that causes large variability in the data, the lack of strong predictors, such as irradiance for PV production and the intrinsic unpredictability of the phenomenon. The usage of many furniture does not follow precise patterns and their consumption can be significant, an example is the utilization of a microwave or a toaster.

An interesting feature of random forest is the possibility to visualize the importance of the features, the number of times a certain feature is used in the first split of a tree is an indication of its relevance.

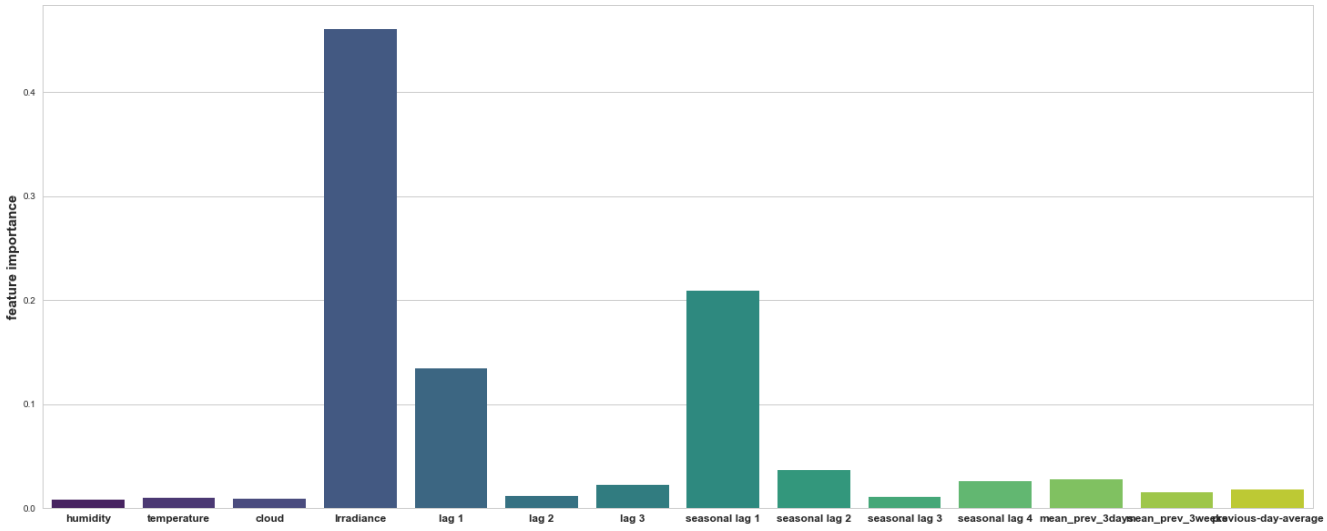

FIGURE 48: FEATURE IMPORTANCE PV PRODUCTION

Feature importance is linked to the correlation matrix, that was showed previously. Features with a high correlation index are the one that most likely have more importance in the model. Irradiance, and some historical information dominate the graph, whereas the influence of other weahter information is limited. The quality of the solar irradiance estimates greatly influences the accuracy of the predictions.

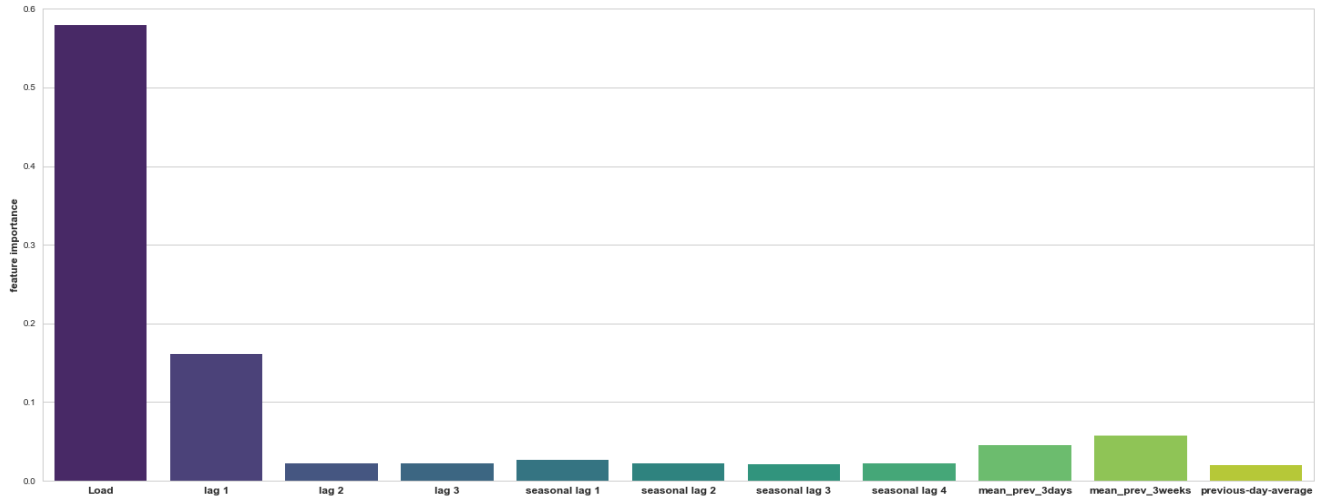

FIGURE 49: FEATURE IMPORTANCE LOAD CONSUMPTION

In the case of load forecasting only time-series data is available. The most recent lags are the more important ones, that being said it is necessary to include in the model some seasonal lags to ensure that the output does not diverge and is able to capture cyclical patterns in the data.

6.6.3 ARIMA

In the theoretical explanation of the algorithm it has been said that the first requirement to fit an ARIMA model to some data is to make sure that the time-series is stationary. Figure 50 shows the raw PV production time-series on the left and its stabilized version on the right.

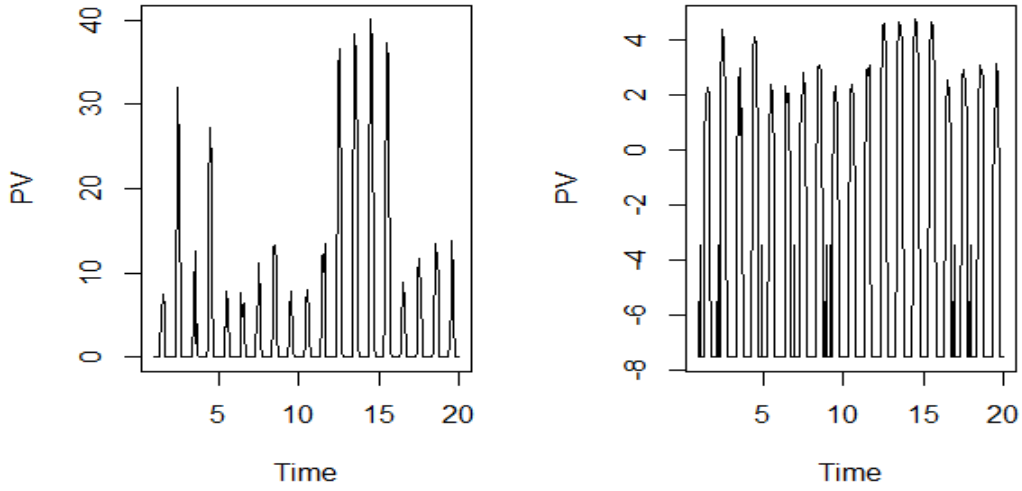

FIGURE 50: RAW AND TRANSFORMED PV TIME-SERIES

Once the series is stationary, the autocorrelation (ACF) and partial autocorrelation (PACF) plots can be analyzed to understand which model can better fit the data.

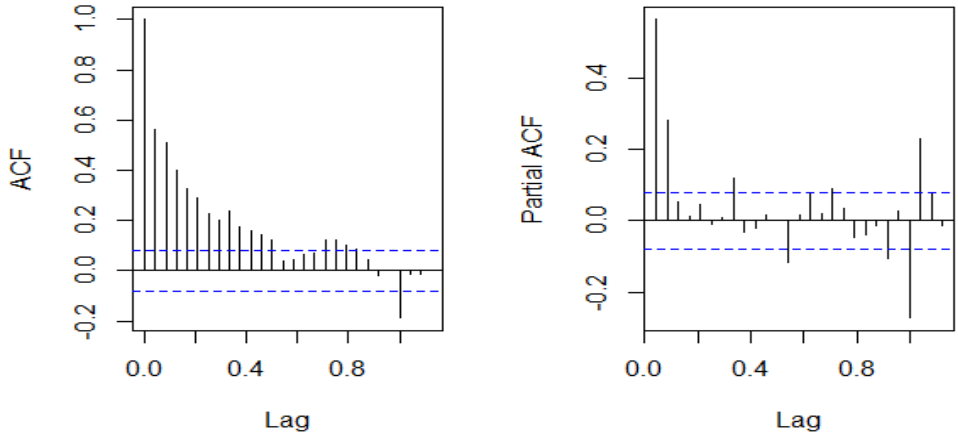

FIGURE 51: ACF & PACF PLOT PV PRODUCTION

On the horizontal axis are represented the lags, which are distanced once from each other by an hour of time. The vertical axis shows the correlation between the selected lag and the current time-step value. Two blue dotted lines are visible, they delimitate the non-relevance band, lags which value is contained in this area are not to be considered. The shape of the significant lags is used to identify which kind of model to choose and how many lags should be considered.

The plots show a slowly decreasing ACF and a sharp cut-off in the PACF after the second lag. The model should be an autoregressive of parameter "*p*" equal to two. It is known that the phenomenon has daily seasonality, so it is necessary to observe the shape the plot after one entire season (value 1 on the x-axis). Both ACF and PACF plots cut-off sharply with one and two significant lags.

The indications from the two plots above have been used and several models have been tested using the Box-Jenkins procedure, a seasonal ARIMA (3,0,1)(2,1,2) leads to the best results.

Three significant lags for the autoregressive and one for the moving average part have been used for the non-seasonal component, whereas two autoregressive and two moving average lags are needed for the seasonal component of the model, that additionally required one order of differentiation.

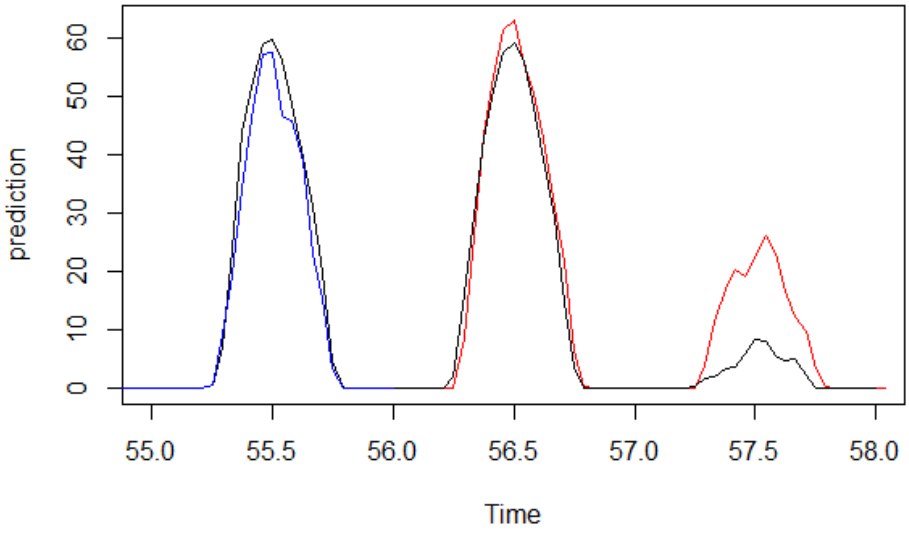

FIGURE 52: ARIMA MODEL OUTPUT PV PRODUCTION

The results for two consecutive days can be seen in the picture, a sunny and a cloudy day were forecasted. The black lines is drawn using the measured values, in blue the output of the ARIMA model on the training data and in red the forecasted values. A sunny and a cloudy day are used for the predictions, the difference in the model performance is clear. Sunny days' profile is much more regular and predictable; hence forecasts are more likely to be accurate.

The same procedure is applied to load consumption time-series. Data needs to be stabilized, figure 53 shows the transformation from the original data to its stationary version.

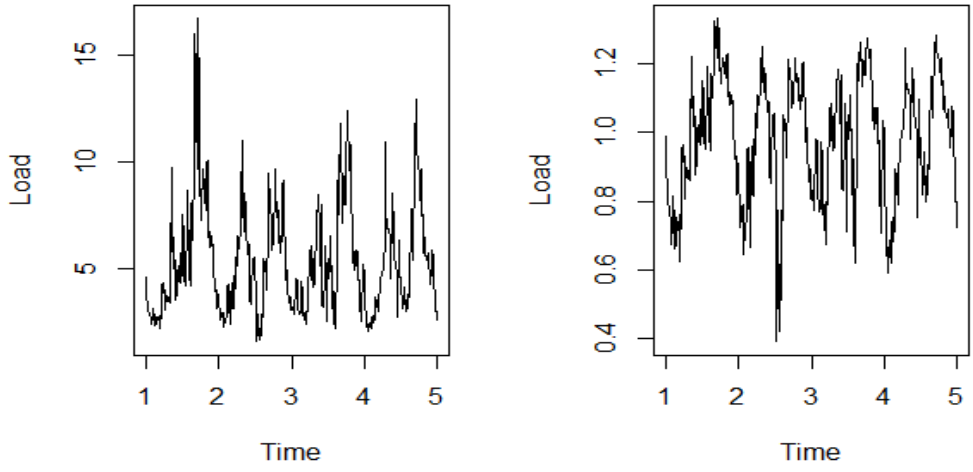

FIGURE 53: RAW AND TRANSFORMED LOAD TIME-SERIES

Load forecasts are generated on a quarterly hour basis, 96 values for each day need to be calculated. To stabilize the series one order of differentiation is needed both for seasonal and non-seasonal component. Once the data is stationary ACF and PACF plots can be analyzed.

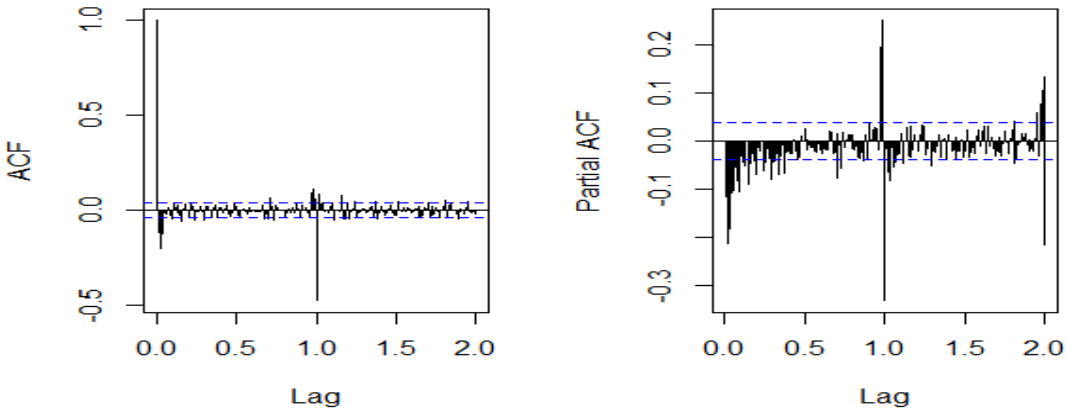

FIGURE 54: ACF&PACF PLOT LOAD CONSUMPTION

The data is extremely spicky, due to the high sample rate of the series. It is not possible to understand which model can better fit the data from the correlation plots, so a trial and error process is needed. Of the many models tested, a seasonal arima (3,1,2)(0,1,1) is the one that yields the best results. The chosen model is complex, several autoregressive and moving average orders are needed both for the seasonal and non-seasonal component.

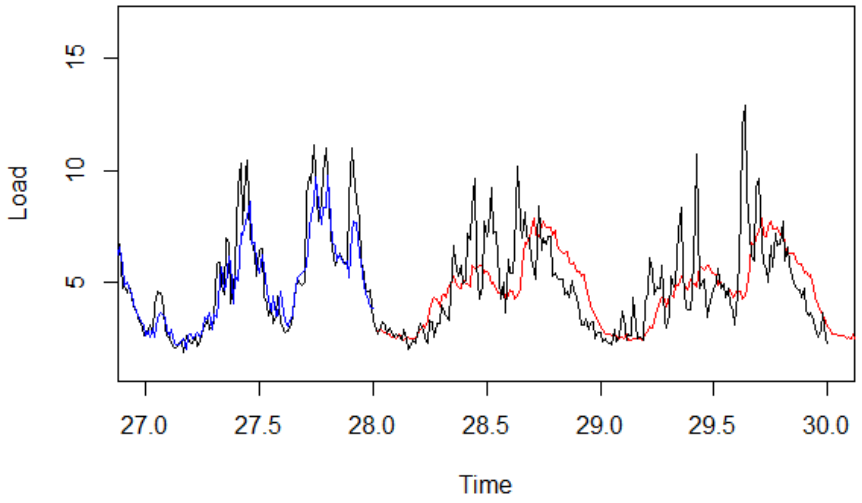

FIGURE 55: ARIIMA MODEL LOAD CONSUMPTION OUTPUT

Figure 55 shows that the model is able to fit the training data particularly well and capture the overall trend for future values. Unpredictable spikes are the main source of error.

A closing note on ARIMA models is related to the size of the training set. Since parameter estimation works minimizing the overall error on the training data a very large dataset can lead to a deterioration of the performances. Old data is not particularly relevant for the predictions and considering the high variability of the time-series, it is easy to understand why the training data should not be too large. Several tries were made for this analysis and the optimal size appears to be around one month of data.

6.7 Error metrics

The evaluation of the algorithms' performances is done using a large collection of error metrics. The chosen estimators are presented and explained.

Root mean square error

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

It is a scale-dependent error metric, its units are the same as the target variable one so it is very easy to interpret. The lower the value of RMSE the better the forecasts are. A normalized version is available, the normalized root mean square error. The normalized version in which RMSE is divided by the range of the target variable.

$$nRMSE = \frac{RMSE}{y_{max} - y_{min}}$$

A high value of nRMSE suggests that the predictions' error is significant and the model need improvements [34].

Mean absolute error

$$MAE = \frac{\Sigma_{i=1}^{n}|y_i - \widehat{y_i}|}{n}$$

MAE like RMSE shares the same units of the target variable, making it an easy metric to understand. In comparison with RMSE, MAE penalizes less large error taking the absolute value instead of squaring the quantity [34].

Mean bias error

$$MBE = \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)}{n}$$

The mean of the difference between the real value and the one predicted by the model is taken. The final result is not easy to interpret, because differences with opposite sign counterbalance each other, that is why RMSE and MAE are usually preferred to this metric [34].

Mean absolute percentage error

Contrarily to the previous error metrics, MAPE is a percentage error, its unit are not the same of the predicted variable. Typically, MAPE is defined by the following formula:

$$\boldsymbol{MAPE = \frac{100}{n}\Sigma_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|}$$

When "y_i" approaches or is equal to zero the division tends to infinity or is indefinite, in these cases an alternative formulation is needed.

$$\boldsymbol{MAPE^{*} = \frac{100}{n}\left|\frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)}{\Sigma_{i=1}^{n} y_i}\right|}$$

The second formulation is needed in the case of PV forecast, during the night the production is always nihil, whereas no major problems appears for load predictions some consumption is always present throughout the day [34].

6.8 Results analysis

Once the models and the error metrics are defined, it is possible to calculate the results and analyze them, so that the best algorithm can be found and used. The month of May is used as test set, daily predictions are prepared and compared to the actual values storing the results. Boxplot are used to showcase the performances of the models. To better understand the forecasting power of the algorithms their performances are compared to a naïve predicting technique. The PV production and load consumption profile are predicted transposing the profiles of the previous day. The models are supposed to do better than the naïve model.

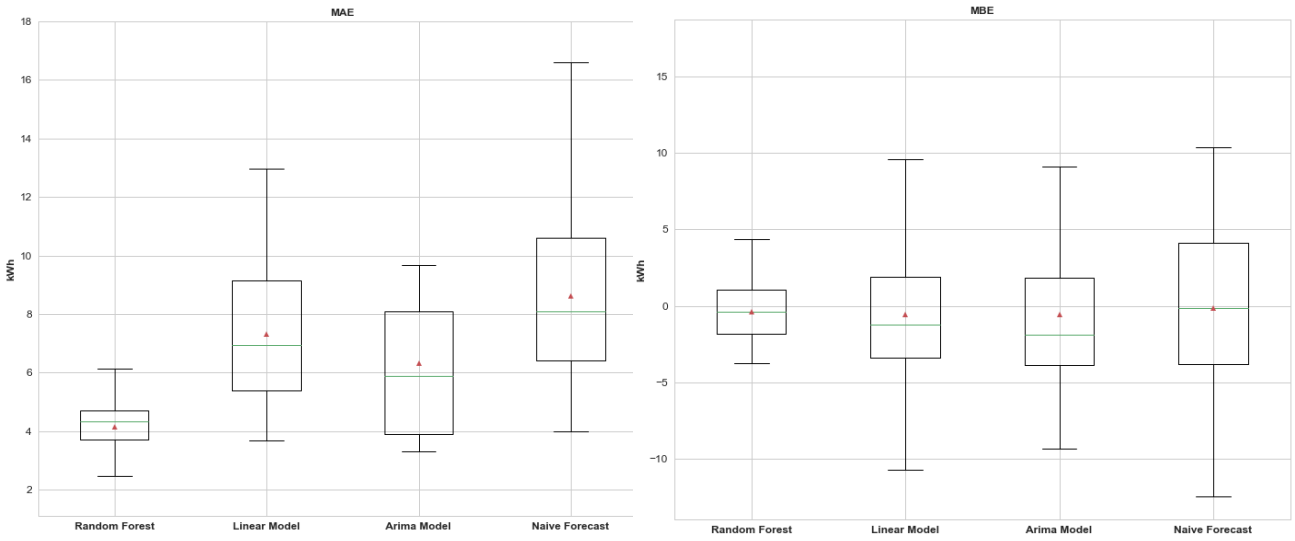

FIGURE 56: MAE & MBE BOXPLOTS PV PRODUCTION

MAE and MBE results for PV forecast are shown above, in both cases the closer are the results to zero the better it is. Random forest is the model that has better perfomances, the average error is low and the error distribution is contained in a relatively small range.

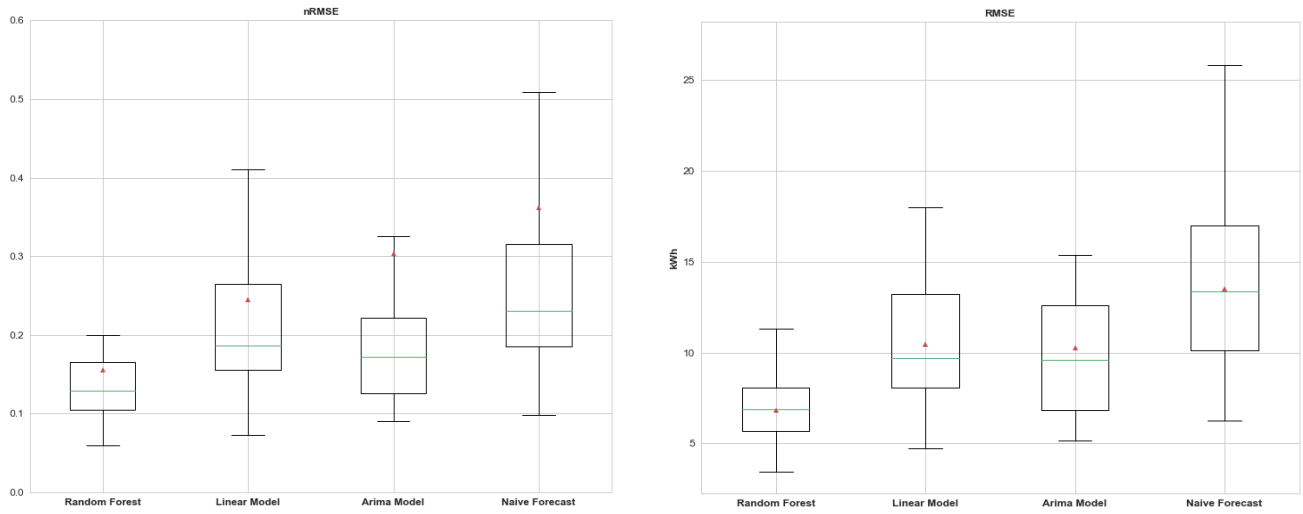

FIGURE 57: NRMSE &RMSE BOXPLOTS PV PRODUCTION

RMSE and nRMSE are presented, the models that are closer to zero are the ones that perform better. Random forest outperforms the other algorithms, its average daily cumulated error is only 7 kWh.

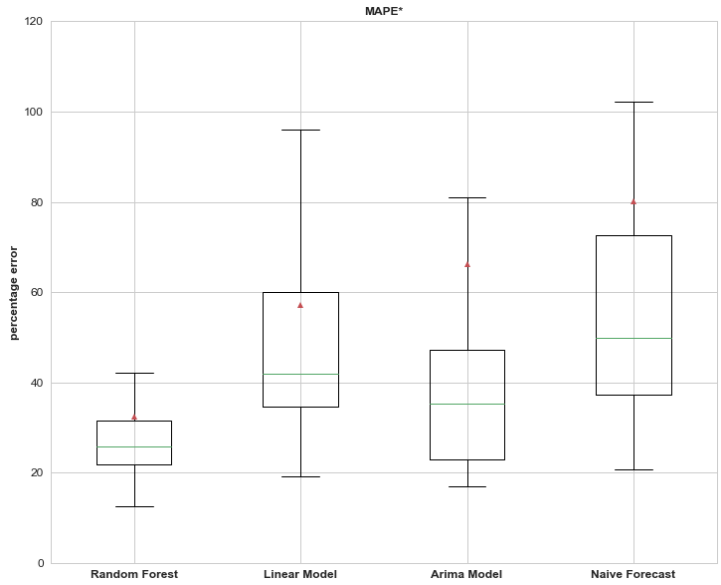

FIGURE 58: MAPE* BOXPLOTS PV PRODUCTION

Random forest is confirmed as the best model by MAPE* as well, the average error is around 30% that is a reasonable value for this group of houses, similar results were obtained by the study in reference [17].

The following plots shows the load forecasts' results.

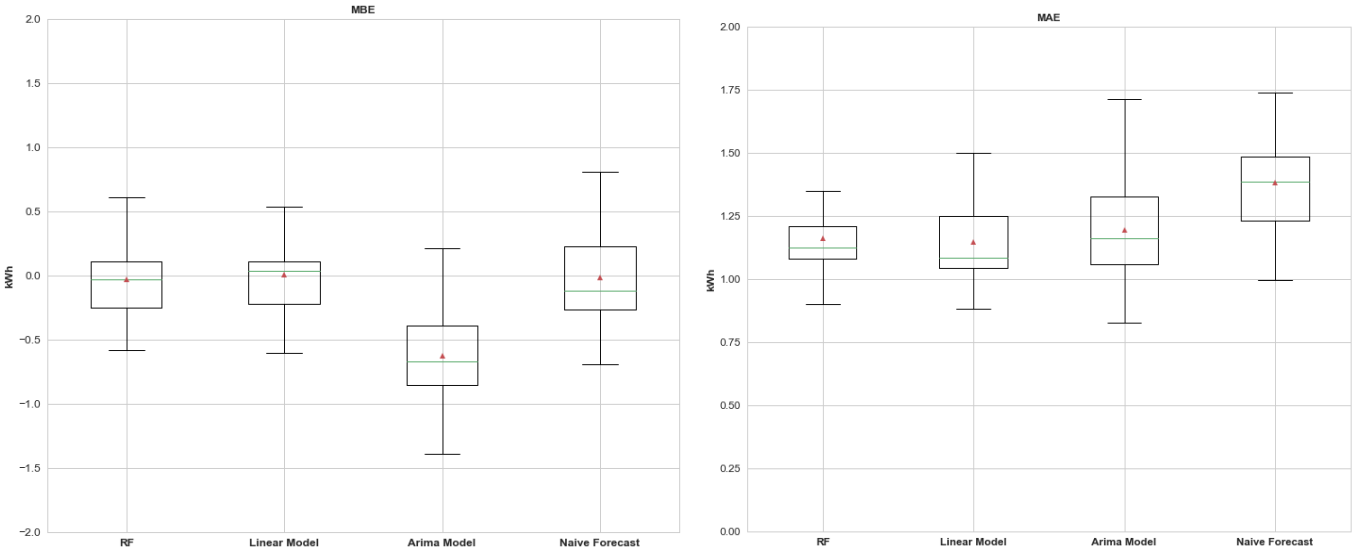

FIGURE 59: MAE & MBE BOXPLOTS LOAD CONSUMPTION

Performances of random forest and linear regression are very similar in terms of MBE, while the first algorithm is better in the MAE, because the spread of its error is smaller.

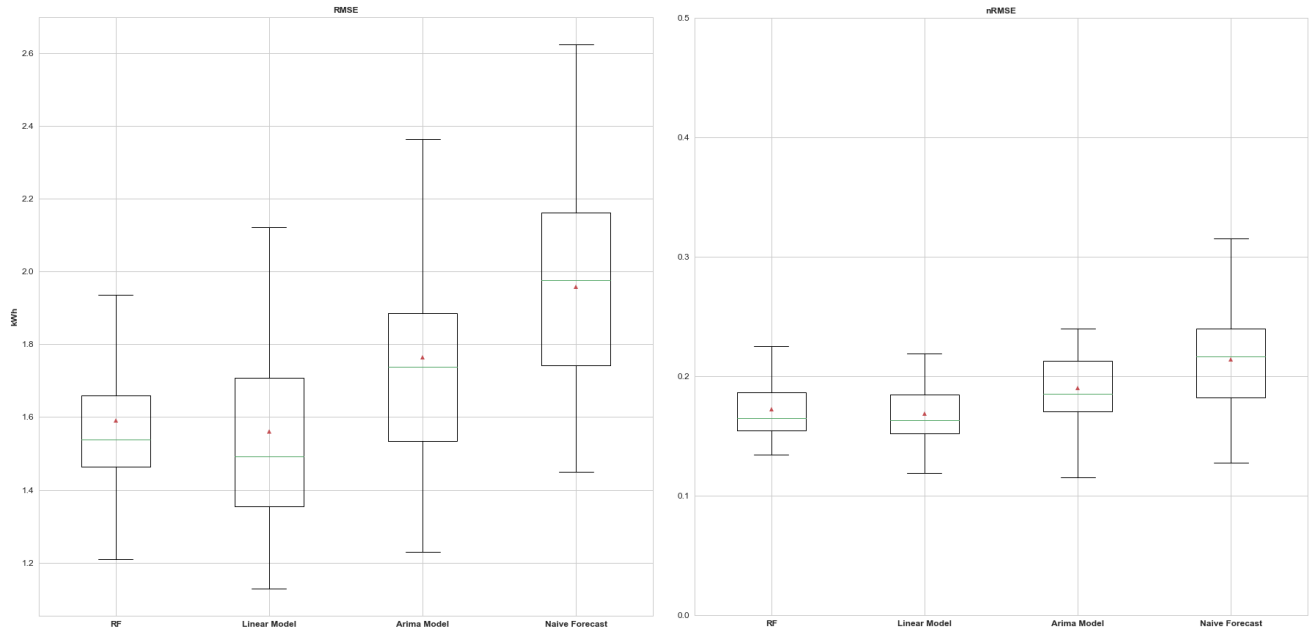

FIGURE 60: NRMSE & RMSE BOXPLOTS LOAD CONSUMPTION

As before random forest and linear regression are the two models showing better results, while the average error of the linear model is lower the spread of the distribution is larger.

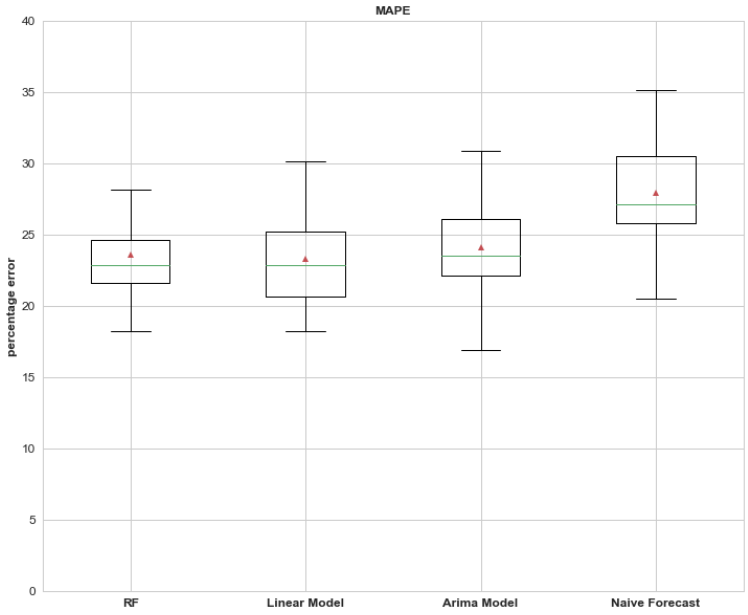

FIGURE 61: MAPE BOXPLOTS LOAD CONSUMPTION

Finally, the MAPE is analyzed, all the models perform better than the naïve forecast technique. An argument regarding which is the best model can be done, as the previous graphs show linear regression and random forest are very similar in terms of performances.

Chapter 7

The most important results of the thesis are here recapped. Some considerations about what are the limits of the analysis are explained and improvement for future works on the topic are suggested. A general consideration about the "lesson learnt" and the reason why this kind of studies are helpful is included in this chapter.

7.1 Summary of the results

The first part of the thesis focused on understanding the data from the houses. The most relevant findings are that solar rooftop are generally oversized for the needs of a common household. When Sun shines the installed PV panels produce a large amount of energy that is rarely met by the load demand, this can lead to problems in the management of the grid. Considering the desire of many governments to support the transition toward a decentralized renewable production of energy, over-injection problems are bound to be seen more and more.

Assessed the over-production problem the attention is shifted toward its solution. Different options are tested, installation of batteries, load shifting and combinations of the two. The simulation on historical data proves that shifting hot water production alone can save up to 700 kWh/year and installation of batteries can help even more saving additional energy. An important conclusion of the simulation is that the over-injection problem cannot be solved entirely by the two previous techniques, due to the large difference between production and consumption.

The economic feasibility of the mentioned solutions is studied and compared to the traditional scenario in which grid is upgraded and the one in which no intervention is done and energy is simply curtailed. The results show that installing up to 28 kWh of storage is usually the most convenient solution available.

Finally, in the sixth chapter are presented forecasting tools to predict load consumption and PV production one-day ahead. It is crucial to have some indications on future consumption and generation, to guarantee a more accurate grid management. The goal is finding a simple model that is still able to produce accurate predictions. Currently, deep neural networks are used for the task, but their

development, training and maintenance is not trivial. Literature review suggested that ARIMA models and random forests are good alternatives to neural networks for online systems. The results of the implementation of the mentioned models shows that random forest is the most promising algorithm. The MAPE of random forest is around 30% and 25% for PV and load forecasts respectively, the implementation is straightforward and fast, making it a good solution for online systems.

7.2 Future works

While an attempt to address all the relevant aspects to the over-injection problem is done in the thesis, some aspects are overlooked and not completely treated.

Optimization of the battery operating strategy should be addressed, in the simulation storage is managed naively, discharging batteries when needed without taking actions that prioritize their useful life. Additionally, battery placement should be investigated by studying the topology of the grid and preparing simulations to evaluate the effect of installing storage in different nodes of the network. A promising solution to this problem is the implementation of genetic algorithms to the optimal placing problem.

Future works should also devote some time in the verification of the assumptions of the physical and economic model. More detailed information regarding the grid upgrade costs and the current and future regulations for energy curtailment should be found.

Forecasting algorithms also deserve some attention, machine learning is a very popular research topic, innovative models are published frequently. New algorithms should be tested to verify if they can provide better forecasts than the ones presented in this thesis.

Finally, to unify all the aspects of the thesis it would be useful to design the energy management system that using the PV and load forecast and the storage provided by the batteries plans the energy flows.

7.3 Lesson learnt

Working on the thesis has been the perfect occasion to interact with real-life data, manipulating and visualizing it to better understand what a high renewable penetration system can face. Different competencies are required to accomplish this kind of studies, a good knowledge of how an energy system works is crucial such as learning how to write scripts to interact remotely with smart-meters and website API to access data. Learning how to implement machine learning algorithm is also an invaluable lesson, this technology is on the rise and it is the solution to many contemporary problems.

References

[1] A. S. Anees, “Grid Integration of Renewable Energy Sources : Challenges , Issues and Possible Solutions.”

[2] R. Fu, D. Chung, T. Lowder, D. Feldman, K. Ardani, and R. Margolis, “U.S. Solar Photovoltaic System Cost Benchmark: Q1 2016,” 2009.

[3] E. M. Sandhu and T. Thakur, “Issues , Challenges , Causes , Impacts and Utilization of Renewable Energy Sources - Grid Integration,” vol. 4, no. 3, pp. 636–643, 2014.

[4] JRC Smart Electricity Systems and Interoperability, “Smart Metering deployment in the European Union.” [Online]. Available: http://ses.jrc.ec.europa.eu/smart-metering-deployment-european-union. [Accessed: 29-Jul-2017].

[5] J. A. Hayward, P. W. Graham, E. L. Ratnam, and L. Reedman, “Future energy storage trends,” 2015.

[6] K. Baes, R. Francis, A. Merhaba, F. Carlot, and C. Nagal, “Battery Storage : Still Too Early ?,” 2017.

[7] P. Palensky and D. Dietrich, “Demand Side Management : Demand Response , Intelligent Energy Systems , and Smart Loads,” no. September 2011, 2011.

[8] M. Bragard, N. Soltau, S. Thomas, and R. W. De Doncker, “The balance of renewable sources and user demands in grids: Power electronics for modular battery energy storage systems,” *IEEE Trans. Power Electron.*, vol. 25, no. 12, pp. 3049–3056, 2010.

[9] International Renewable Energy Agency, “BATTERY STORAGE FOR RENEWABLES : MARKET STATUS AND TECHNOLOGY OUTLOOK,” 2015.

[10] G. Delille, B. Francois, G. Malarange, and J.-L. Fraisse, “ENERGY STORAGE SYSTEMS IN DISTRIBUTION GRIDS: NEW ASSETS TO UPGRADE DISTRIBUTION NETWORKS ABILITIES,” 2009, no. 524, pp. 8–11.

[11] “The Case for Distributed Energy Storage - Renewable Energy World.” [Online]. Available: http://www.renewableenergyworld.com/articles/print/volume-16/issue-4/storage/the-case-for-distributed-energy-storage.html. [Accessed: 30-Aug-2017].

[12] D. Fischer, M.-A. Triebel, T. Erge, and R. Hollinger, “Business Models Using the Flexibility of Heat Pumps - A Discourse,” *12th IEA Heat Pump Conf.*, no. May, 2017.

[13] S. Pelland, J. Redmund, T. Oozeki, and K. De Brabendere, “Photovoltaic and Solar Forecasting : State of the Art,” 2013.

[14] R. H. Inman, H. T. C. Pedro, and C. F. M. Coimbra, “Solar forecasting methods for renewable energy integration,” *Prog. Energy Combust. Sci.*, vol. 39, no. 6, pp. 535–576, 2013.

[15] C. Voyant *et al.*, “Machine learning methods for solar radiation forecasting : A review,” *Renew. Energy*, vol. 105, pp. 569–582, 2017.

[16] M. Bouzerdoum, A. Mellit, and A. M. Pavan, “ScienceDirect A hybrid model (SARIMA – SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant,” *Sol. Energy*, vol. 98, pp. 226–235, 2013.

[17] L. Narvarte, M. P. Almeida, and O. Perpin, “ScienceDirect PV power forecast using a nonparametric PV model,” vol. 115, pp. 354–368, 2015.

[18] T. Zufferey, A. Ulbig, S. Koch, and G. Hug, “Forecasting of Smart Meter Time Series Based on Neural

Networks."

[19] Government of the Netherlands, "Energy Agenda," 2016.

[20] Government of the Netherlands, "Energy report," 2016.

[21] "Powerwall | The Tesla Home Battery." [Online]. Available: https://www.tesla.com/nl_NL/powerwall. [Accessed: 17-Aug-2017].

[22] "Electricity price statistics - Statistics Explained." [Online]. Available: http://ec.europa.eu/eurostat/statistics-explained/index.php/Electricity_price_statistics. [Accessed: 17-Aug-2017].

[23] "Time Value Of Money: Determining Your Future Worth." [Online]. Available: http://www.investopedia.com/articles/fundamental-analysis/09/net-present-value.asp. [Accessed: 18-Aug-2017].

[24] "Stationarity and differencing of time series data." [Online]. Available: https://people.duke.edu/~rnau/411diff.htm. [Accessed: 29-Aug-2017].

[25] "8.3 Autoregressive models | OTexts." [Online]. Available: https://www.otexts.org/fpp/8/3. [Accessed: 29-Aug-2017].

[26] "8.4 Moving average models | OTexts." [Online]. Available: https://www.otexts.org/fpp/8/4. [Accessed: 29-Aug-2017].

[27] "6.4.4.6. Box-Jenkins Model Identification." [Online]. Available: http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc446.htm. [Accessed: 29-Aug-2017].

[28] "Lesson 1: Simple Linear Regression | STAT 501." [Online]. Available: https://onlinecourses.science.psu.edu/stat501/node/250. [Accessed: 29-Aug-2017].

[29] "1.11. Ensemble methods — scikit-learn 0.19.0 documentation." [Online]. Available: http://scikit-learn.org/stable/modules/ensemble.html. [Accessed: 29-Aug-2017].

[30] "1.10. Decision Trees — scikit-learn 0.19.0 documentation." [Online]. Available: http://scikit-learn.org/stable/modules/tree.html. [Accessed: 29-Aug-2017].

[31] "scikit learn decision tree export graphviz - wrong class names in the decision tree - Stack Overflow." [Online]. Available: https://stackoverflow.com/questions/41207923/scikit-learn-decision-tree-export-graphviz-wrong-class-names-in-the-decision-t]. [Accessed: 30-Aug-2017].

[32] "efficiency - When is a Model Underfitted? - Data Science Stack Exchange." [Online]. Available: https://datascience.stackexchange.com/questions/361/when-is-a-model-underfitted. [Accessed: 30-Aug-2017].

[33] L. Breiman, "Random forests," pp. 1–33, 2001.

[34] "2.5 Evaluating forecast accuracy | OTexts." [Online]. Available: https://www.otexts.org/fpp/2/5. [Accessed: 29-Aug-2017].