Data-driven forecasting models for electricity consumption and solar power generation to assess possible demand-response strategies

The case study of Instituto Superior Técnico

Fabia Miorelli

Thesis to obtain the Master of Science Degree in

Energy Engineering and Management

Supervisor: Carlos Augusto Santos Silva

Examination Committee

Chairperson: Prof. Luís Filipe Moreira Mendes
Supervisor: Prof. Carlos Augusto Santos Silva
Member of the Committee: Prof. Paulo José da Costa Branco

July 2019

≪The sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work.≫

John von Neumann

(1903-1957)

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Declaração

Declaro que o presente documento é um trabalho original da minha autoria e que cumpre todos os requisitos do Código de Conduta e Boas Práticas da Universidade de Lisboa.

Lisboa, 15-05-2019
Fabia Miorelli

Acknowledgments

First, I would like to express my sincere gratitude to my advisor prof. Carlos Silva for his insightful comments, continuous encouragement and for all the time he dedicated me. Thank you for giving me the possibility to work on such interesting topics, while at the same time, allowing me to widen my research the way I liked most. Above all, this thesis has been a priceless learning experience, which allowed me to broaden my knowledge and my skills.

Besides my advisor, I would like to thank eng. Mário Matos of the Campus Sustentável project for providing me with the consumption data and uncountable information during the development of my work. A great thanks also goes to the people of the IST DSI (Direção de Serviços de Informática) office and to Jorge Palma from the Técnico Meteo Station for sharing with me Wi-Fi connected users and weather related data respectively, which were essential to develop more accurate models. When working on solar modelling, I received valuable inputs from prof. Filipe Mendes and Lisa, which always managed to clear all my doubts. I highly appreciated all the machine learning insights from Mirko and Rui, which directed me in the right direction when it was not clear to me in which way to go. I owe a special thanks to Davide, Max, Olga and Guido for the discussions concerning the modelling of the battery, which in most cases went far beyond the topic of this thesis. I would also like to thank Jonathan for helping me with the abstract in Swedish.

I am very grateful for the contribution of all these people. The unconditional willingness to critically discuss, explain, and share ideas without any self-interest other than for the sake of spreading knowledge is probably one of the most essential virtues that our society needs. I am also immensely thankful to InnoEnergy for the priceless opportunities offered, and to Royal Institute of Technology and Instituto Superior Técnico for hosting me throughout these years. I would like to thank Nele, Peter and Duarte for the kind assistance in the most diverse occasions. I am extremely grateful to the SELECT family, that filled the past years with incredible experiences, authentic friendships and too many hours of travelling which contributed to create a once-in-a-lifetime Master's program.

An important acknowledgement to my most recent Lisbon flatmates Elisa, Maria, Victoria and Elisabeth that made me feel at home, filled every day with happiness, laughs and fun, even in the darkest moments and bore my sophisticated time estimations on how much longer I would need for my thesis. Thanks to all my close friends who could forgive my lack of attention in these past months.

This wonderful experience would not have been possible without the support of my family. A huge thank you to my exceptional mum, that was always caring and encouraging and to which I am unreservedly grateful. Mami, questa tesi é dedicata a te.

Abstract

To understand the optimal scenario which would allow to reduce building energy consumption and, as a result achieve economic savings, it is of the utmost importance to be able to quantify how much electricity can be produced in a decentralised manner, and how much the consumption will be in a specific time in the future. The development of energy forecasting models is therefore paramount to the achievement of higher energy efficiency standards, especially if coupled with implementations that enable the automatic control of the energy system.

The aim of this work is to develop electricity consumption and production forecasting models to suggest possible smart energy management measures for the main campus of Instituto Superior Técnico in Lisbon. First, the performance of different forecasting methods for the energy production of rooftop photovoltaic solar modules and of the energy consumption of selected buildings on campus is simulated and analysed by means of real data. The results show that, among all tested supervised learning methods, artificial neural networks can predict the building energy consumption and the rooftop solar production with good accuracy. An evaluation of possible demand-response strategies exploiting a battery energy storage system is then carried out. Using mixed-integer linear programming the scheduling of the battery system is optimized to shift the consumption from peak hour to off-peak hour. Exploiting time-of-use energy tariffs, the optimized schedule resulted in annual net savings of about 2% including the initial investment for the battery.

The work closes with an outlook towards possible improvements which would potentially allow the real-time implementation of the suggested measures.

Keywords: Energy Management, Demand-Response, Building Energy System Optimization, Artificial Neural Networks, Machine Learning, Battery Energy Storage System.

Sammanfattning

För att förstår det optimala scenariot som skulle leda till reducerad energiförbrukning inom byggnader och som resultat ger ekonomiska besparingar, det är viktigt att kunna kvantifiera hur mycket el can produceras på ett decentraliserat sätt och hur stort förbrukningen kommer blir till specifika tider i framtiden. Utvecklingen av modeller till energi prognostisering är därmed särskilt viktigt för att nå högre standarder i energieffektivitet, speciellt om implementationen kopplas till automatiska energistyrningssystem.

Syftet med detta arbete är att utveckla elförbruknings och produktionsprognosmodeller för att föreslå möjliga smarta energihanteringslösningar till Lissabons tekniska universitet Instituto Superior Técnico. Först simuleras och analyseras prestandan av olika prognosmetoder för energiproduktion med solcellsmoduler på huvudbyggnadens tak och energiförbrukningen av utvalda byggnader på universitets område med hjälp av reell data. Resultaten visar att bland alla testade övervakade inlärningsmetoder kan konstgjorda neurala nät förutsäga byggnadens energiförbrukning och soltaksproduktionen med god noggrannhet. En utvärdering av möjliga strategier for efterfrågan och respons som utnyttjar ett batterilagringssystem utfördes efteråt. Genom att använda linjärprogrammering med blandade-heltal optimeras schemaläggningen av batterisystemet för att förskjuta konsumtionen från topptimmar till lågtider. Med utnyttjande av tids relaterade energi priser, resulterade den optimerade schema i årliga nettosparande på ca 2% inklusive den ursprungliga investeringen för batteriet.

Arbetet avslutas med en syn på förbättringar som möjligen skulle tillåta en real tids implementering av de diskuterade åtgärderna.

Nyckelord: Energi Management, Demand-Response, Energisystemoptimering för Byggnader, Artificiellt Neuronnät, Maskininlärning, Batteri Energilatersystem.

Resumo

Para compreender o cenário ideal que permitiria reduzir o consumo de energia nos edifícios e, consequentemente, obter poupanças económicas, é extremamente importante quantificar quanta electricidade pode ser produzida de forma descentralizada e quanto será o consumo num momento específico no futuro. O desenvolvimento de modelos de previsão de energia é, portanto, fundamental para a obtenção de padrões mais elevados de eficiência energética, especialmente se associados à implementação que permite o controlo automático do sistema de energia.

O objetivo deste trabalho é desenvolver modelos de previsão de consumo e produção de eletricidade para sugerir possíveis medidas de gestão de energia inteligente para o campus principal do Instituto Superior Técnico em Lisboa. Em primeiro lugar, o desempenho de diferentes métodos de previsão para a produção de energia de módulos solares fotovoltaicos no telhado e do consumo de energia de edifícios selecionados no campus é simulado e analisado por meio de dados reais. Os resultados mostram que, entre todos os métodos de aprendizagem supervisionada testados, as redes neurais artificiais podem prever o consumo de energia do prédio e a produção solar no telhado com boa precisão. Uma avaliação das possíveis estratégias de resposta à demanda que exploram um sistema de armazenamento de energia da bateria é então realizada. Utilizando programação linear inteira mista, o agendamento do sistema de bateria é optimizado para mudar o consumo da hora de ponta para a hora de vazio. Explorando as tarifas de energia em função do tempo de uso, o cronograma otimizado resultou numa economia líquida anual de cerca de 2%, incluindo o investimento inicial para a bateria.

O trabalho termina com uma perspectiva de possíveis melhorias que potencialmente permitiriam a implementação em tempo real das medidas sugeridas.

Palavras-chave: Gestão da Energia, Gestão da Procura, Optimização de Sistemas Energéticos de Edifícios, Redes Neuronais Artificiais, Aprendizagem Máquina, Sistemas de Armazenamento de Energia Eléctrica.

Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

- ANFIS - Adaptive Neuro-Fuzzy Inference System
- ANN - Artificial Neural Network
- AP - Access Point
- ARIMA - Auto-Regressive Integrated Moving Average
- BEMS - Building Energy Management System
- BESS - Battery Energy Storage System
- BPNN - Back Propagation Neural Network
- BRP - Balance Responsible Party
- CBS - Case Base Reasoning
- CEA - City Energy Analyst
- CPP - Critical Peak Price
- CO_2 - Carbon Dioxide
- CT - Temperature Coefficient
- CV - Cross Validation
- DA - Deterministic Annealing
- DLS - Daylight Saving
- DNN - Deep Neural Network
- DOP - Day Of Prediction
- DR - Demand-Response
- DSM - Demand Side Management
- DSO - Distribution System Operator
- EDP - Energias de Portugal
- EPSO - Evolutionary Particle Swarm Optimisation
- EU - European Union
- FCM-FFNN - Fuzzy C-Means with Feed Forward Neural Network
- FTL - Follow The Leader
- FFNN - Feed Forward Neural Network
- GA - Genetic Algorithm
- GA-ANFIS - Genetic Algorithm Adaptive Neuro-Fuzzy Inference System
- GBRT - Gradient Boosted Regression Trees
- GHG - Green House Gas
- GLPK - GNU Linear Programming Kit
- GP - Genetic Programming
- GRBFN - Generalized Radial Basis Function Network
- HEMS - Home Energy Management System
- HDKR - Hay Davis Klucher Reindl
- HME-FFNN - Hierarchical Mixture of Experts with Feed Forward Neural Network

- HME-REG - Hierarchical Mixture of Experts with Linear Regression Experts
- HVAC - Heating, Ventilation, Air Conditioning
- IEA - International Energy Agency
- IRENA - International Renewable Energy Agency
- IST - Instituto Superior Técnico
- kNN - K-Nearest Neighbours
- LAP - Labour Activity Parameter
- LSSVM - Least Squares Support Vector Machine
- MAE - Mean Absolute Error
- MAPE - Mean Absolute Percentage Error
- MILP - Mixed Integer Linear Programming
- ML - Machine Learning
- MLP - Multi-Layer Perceptron
- MPP - Maximum Power Point
- MR - Multiple Regression
- MRE - Mean Relative Error
- MV - Medium Voltage
- NN - Neural Network
- NOCT - Normal Operating Cell Temperature
- nRMSE - Normalised Root Mean Square Error
- PCA - Principal Component Analysis
- PSO - Particle Swarm Optimisation
- PV - Photovoltaic
- RBFN - Radial Basis Function Network
- RTP - Real Time Price
- RMSE - Root Mean Square Error
- SOM - Self Organising Map
- SNMP - Simple Network Management Protocol
- SoC - State of Charge
- SSE - Sum of Squares Error
- SSR - Regression Sum of Squares Error
- SST - Total Sum of Squares Error
- STLF - Short Term Load Forecasting
- SVM - Support Vector Machine
- SVR - Support Vector Regression
- ToU - Time of Use
- TSO - Transmission System Operator
- UN - United Nations
- VAF - Variance Accounted For
- wKNN - Weighted K-Nearest Neighbours
- wSVM - Weighted Support Vector Machine

Symbols

- A - Anisotropy index $[-]$
- $C_{battery}$ - Battery capacity $[kWh]$
- D - Diode
- E_{grid} - Energy from the grid $[kWh]$
- $E_{outBattery}$ - Discharged energy from the battery $[kWh]$
- F' - Clearness index $[-]$
- F_1 - Circumsolar coefficient $[-]$

- F_2 - Horizon brightening coefficient $[-]$
- $f_{11}, f_{12}, f_{13}, f_{21}, f_{22}, f_{23}$ - Perez coefficients
- G - Solar irradiance $[W/m^2]$
- G_T - Total solar irradiance on the tilted surface $[W/m^2]$
- $G_{T,b}$ - Beam irradiance on the tilted surface $[W/m^2]$
- $G_{T,r}$ - Reflected irradiance on the tilted surface $[W/m^2]$
- $G_{T,d}$ - Diffused irradiance on the tilted surface $[W/m^2]$
- $G_{T,d,c}$ - Circumsolar component of the diffused irradiance on the tilted surface $[W/m^2]$
- $G_{T,d,i}$ - Isotropic component of the diffused irradiance on the tilted surface $[W/m^2]$
- $G_{T,d,h}$ - Horizon Brightening component of the diffused irradiance on the tilted surface $[W/m^2]$
- I - Solar radiation $[J/m^2]$
- I_b - Beam radiation $[J/m^2]$
- I_d - Diffused radiation $[J/m^2]$
- I_o - Extraterrestrial radiation $[J/m^2]$ or
- I_o^{ref} - Reverse saturation current $[A]$
- I_T - Total solar radiation on the tilted surface $[J/m^2]$
- I_s - Photoelectric current $[A]$
- I_{MPP} - Maximum Power Point current $[A]$
- I_{sc} - Short circuit current $[A]$
- k - Boltzmann constant $[J//K]$
- m - Air mass
- m' - Modified ideality factor $[-]$
- N - Day number in the year
- P_{MPP} - Maximum Power Point power $[W]$
- q - Charge $[C]$
- R^2 - Coefficient of determination $[-]$
- R_b - Ratio of the beam radiation on the tilted surface to that on the horizontal surface $[-]$
- T_{amb} - Ambient temperature $[\circ C]$
- T_{cell} - Solar cell temperature $[\circ C]$
- V_{MPP} - Maximum Power Point voltage $[V]$
- V_{OC} - Open circuit voltage $[V]$

Greek Symbols

- β - Tilt angle $[\circ]$
- Δ - Sky brightness index $[-]$
- ε - Sky clearness index $[-]$
- θ - Angle of incidence for a surface facing the equator $[\circ]$
- θ_z - Zenith angle $[\circ]$
- ρ_g - Reflectance of the ground $[-]$

Chapter 1

Introduction

Cities are where most European energy is consumed and also the origin of most greenhouse gas (GHG) emissions [1]. In 2007, for the first time in human history, the number of global urban dwellers outnumbered those living in rural settings. By 2050, urbanization will become one of the 21st century's most transformative trends putting cities right at the epicentre of a global shift from rural to urban areas, as the world's urban population is expected to nearly double. The latest UN estimates suggest that this trend is likely to lead to a total of 6.3 billion urban residents by 2050, approximately 70% of the predicted total global population.

Massive sustainability challenges in terms of housing, infrastructure, natural resources, services and health will have to be faced as most social and cultural interactions, economic activities as well as environmental and humanitarian impacts will be increasingly concentrated in cities. This global context is not unknown to Europe either. Several leading and progressive cities and towns across the European continent have already taken innovative steps to enhance the deployment and use of renewable energy resources within their geographic boundaries. However, European cities need to make an effort to further implement new measures to meet the ever-increasing demands for energy services and adapt to the changing energy landscape.

The building sector accounts for 40% of primary energy use and 40% of total GHG emissions, becoming one of the largest energy consuming sectors in the world. The electricity consumption in building is continuously increasing and if no action is taken towards more effective energy efficiency measures, it is set to increase by 50% by 2050 [2]. As a consequence, buildings have become the primary focus of energy efficiency policies, which depend heavily on understanding and modelling energy consumption to evaluate the impact of energy efficiency measures.

There are several ways to attempt to model and simulate a building in order to optimize the operation of its systems, evaluate audit retrofit actions, or forecast energy consumption. Different techniques, varying from simple regression to models that are based on physical principles to data-driven models, can be used for simulation. A frequent hypothesis for all these models is that the input variables should be based on realistic data when they are available, otherwise the evaluation of energy consumption might be highly under or over estimated. Electricity consumption patterns are though difficult to estimate as they depends on various seasonal, monthly, daily and hourly complex variations. As such, building energy consumption plays an important role in energy efficiency strategies and accurate energy forecasting models have numerous implications in energy planning and optimization of buildings and campuses. For new buildings, where past recorded data is unavailable, computer simulation methods are used for energy analysis and forecasting future scenarios. However, for existing buildings with historically recorded time series energy data, statistical and machine learning techniques have proved to be more accurate and quicker.

Knowing how much electricity can be produced in a decentralised manner, and how much the consumption will be in a specific time in the future, allows to understand the optimal scenario to achieve energy or economics savings. In particular, coupling forecasting model with implementations that enable the automatic control can help achieving optimised solutions and higher energy efficiency measures. Buildings that are equipped with such active solutions are called intelligent building. The starting step to reach this status, is the ability to simulate and predict energy consumption and production in an accurate manner. The development of energy predictions is therefore paramount to the achievement of higher energy efficiency standards and consequent energy savings.

1.1 Objectives and Research Questions

The main objective of this work is to analyse and suggest possible smart energy management measures for the main campus of Instituto Superior Técnico (IST), the engineering school of the University of Lisbon. In particular, the performance of different forecasting methods for the energy production of photovoltaic solar modules and of the energy consumption of the campus are simulated and analysed by means of real data. Moreover, the possibility of implementing demand response strategies is analysed. Pursuant to this main objective, this work provides solutions to the following research tasks:

- Perform an extensive literature review in order to discover any previous works which share relevance with this work's main objective. The literature review should also examine the current state of the technologies under consideration, investigate forecasting methods that have been employed in similar tasks and analyse possible demand-response (DR) strategy options;
- Develop two forecasting models based on historical data in Python: one to forecast the solar photovoltaic energy production, and the other one to predict the energy consumption of each of the analysed buildings on campus, using data science and machine learning approaches;
- Assess the modelling approach which best describes how energy is produced and consumed, and assess its performance and accuracy for each of the forecasting models mentioned above;
- Create a model to assess possible demand-response strategies, including its economical aspects to validate its feasibility in a real-life environment;
- Suggest future improvements to the proposed models to enhance the study of the implementation of energy efficiency measures even further.

1.2 Methodology and Thesis Outline

Considering the diverse research questions outlined above, in order to effectively carry out this work, the present work is structured by chapters as follows:

- In Chapter 1 (the current chapter), a general introduction to the topic is given, and the main research question and the outline of this thesis are defined;
- In Chapter 2, the background information and current state-of-the-art of the key system models (i.e. solar photovoltaic panels, electrochemical batteries, demand response management systems) and forecasting and optimisation methods being considered (machine learning methods and optimisation techniques) are examined and the respective economical aspects are presented;
- In Chapter 3, the methodology which is applied to structure, carry out and test the developed machine learning methods is presented together with the specific case study of Instituto Superior Técnico. The economic aspect of the proposed forecast-based solution is also investigated resulting in a project-based assessment of the proposed demand response management system;
- In Chapter 4, the results of the analysis presented in the previous chapter are highlighted, and by means of a sensitivity and accuracy analysis validated. This leads into the choice of the best modelling approach for the forecasting of energy consumption and generation. The results of the demand-response optimisation are laid out and its economical feasibility analysed.
- In Chapter 5, the main conclusions of this work are drawn and possible improvements are suggested.

Chapter 2

Background

This chapter looks to provide a clearer understanding of all concepts and technologies that are key to create the framework of this analysis. Therefore, this chapter is divided as follows: Initially, the concepts of energy management and flexibility are addressed in order to frame this work into the more general energy management topic. Thereafter key technologies and methods such as solar photovoltaic energy, data-driven prediction methods and Battery Energy Storage Systems (BESS) which enable the implementation of energy management flexibility strategies are defined. Eventually, the chapter concludes with the presentation of methods to evaluate the financial assessment of possible flexibility strategy.

2.1 Energy Management

Energy management is a widely diffused term in literature but no cohesive definition of this term is yet established. Different authors agree though that the focus of energy management practises is to implement continuous improvements to achieve higher energy efficiency standards and its importance has been demonstrated in several empirical studies [3], [4], [5]. Aside from the implementation of energy efficient technologies, energy management also deals with the maintenance of such technologies and it is often described as the combination of engineering, management and operation [6]. It usually does not involve large capital investments or particularly increased operating costs, but even simple measures have proved to have large overall savings. It usually mainly consists in an accurate data and system analysis, goal setting and continuous performance assessment and improvements [7].

The international standard ISO 50001 about energy management systems defines it as a ”set of interrelated or interacting elements to establish an energy policy and energy objectives, and processes and procedures to achieve these objectives” and helps organizations to develop and implement a policy to identify significant areas of energy consumption and commit to energy reductions [8]. Aside form facilitating investments, continuous data collection and analysis can also help detecting malfunctioning equipment, optimising the energy system and evaluating its performance.

2.1.1 Flexibility in the Energy Field

According to the International Energy Agency (IEA), the flexibility of a power system refers to ”the extent to which a power system can modify electricity production or consumption in response to variability, expected or otherwise. In other words, it expresses the capability of a power system to maintain reliable supply in the face of rapid and large imbalances [...]" [9].

Another source described it as "...the modification of generation injection and/or consumption patterns in reaction to an external signal (price signal or activation) in order to provide a service within the energy system. The parameters used to characterize flexibility in electricity include: the amount of power modulation, the duration, the rate of change, the response time, the location etc." [10].

Recently published literature provides a wide range of definitions and detailed discussions, like in Ela et all. [11], of the flexibility in the energy sector, but all refer to the same general concept: the extent to which an energy system can modify its electricity production and consumption in response to variability, expected or not [12]. The energy value chain has two clear sides: generation and consumption. From the definitions of flexibility reported previously, it is clear that flexibility strategies can be implemented on both sides of the energy value chain: what is defined as upward regulation, i.e. an increase in generation or a decrease in demand, or applying what is called downward regulation, which implies a reduction in generation or an increase in demand [10]. As a result, different types of resources excel at different forms of flexibility, and they also incur in different costs when providing flexibility [13].

The most common examples of flexibility can be achieved thanks to energy storage options, electric vehicles (EVs) or from the generation or consumption side. Flexibility can be characterized as well depending on its source, which can be classified as uncontrollable (which do not provide flexibility), curtailable, shiftable, buffered or freely controllable (Figure 2.1).

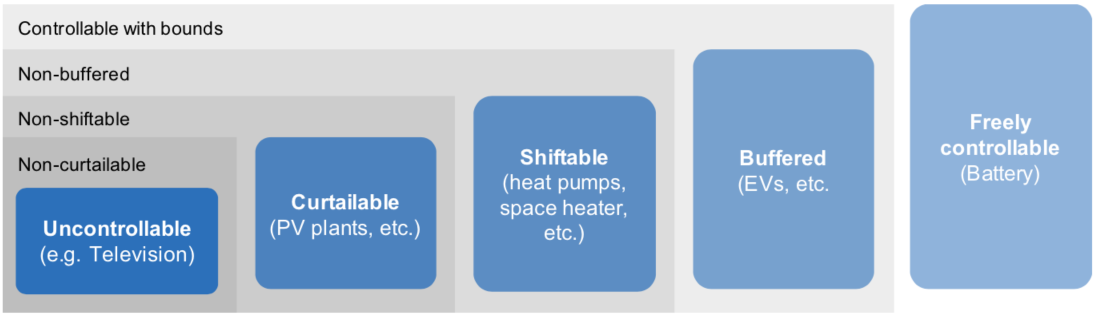

Controllable with bounds
Non-buffered
Non-shiftable
Non-curtailable
Uncontrollable (e.g. Television)
Curtailable (PV plants, etc.)
Shiftable (heat pumps, space heater, etc.)
Buffered (EVs, etc.
Freely controllable (Battery)

Figure 2.1: Categorization of flexibility sources [14].

As it can be sensed from Figure 2.1, curtailable loads, like a PV plant, are a type of load that does not need to recover the curtailed energy once reconnected to the power source. In contrast, in a shiftable load the amount of electricity consumed does not increase or decrease because of a flexibility strategy but it gets shifted in time. Different categories of loads can have different flexibility properties, which might depend on the way they are controlled, as for example EVs, which can be considered curtailable, shiftable or buffered depending on the chosen operational strategy [14].

Besides from its regulation potential and controllability, flexibility options can be classified according to their scale, which is to say whether they provide low, medium or high flexibility, and their position within the electricity grid: behind-the-meter, in the distribution network or at the generation side. As a result, different customer segments of flexibility options arise as for example, a Distribution System Operator (DSO) or a Transmission System Operator (TSO) might be interested in flexibility options to manage grid congestions in their network, perform voltage/reactive power control or controlled islanding and reduce in this

way costs resulting from grid upgrading interventions. Balance Responsible Parties (BRP) could as well exploit flexible resources to manage their portfolio and reduce deviation penalties and operation costs.
On the other end of the energy chain, prosumers, i.e. consumers who both consume and produce electricity, will be empowered by their own flexibility in energy consumption, production and possibly storage to reduce their electricity bill. Since a prosumer is capable of producing and consuming electricity and its interaction with the electrical grid is by definition behind-the-meter, this is where prosumers can take advantage the most from flexibility. Examples of flexibility options for prosumers are:

- Self-balancing: includes the optimum usage of production, self-consumption and selling electricity to the grid based on divergence of prices;
- Demand Charge Reduction: involves reducing the maximum load, i.e. perform what is called peak-shaving, which could result in a smaller contracted power and consequent cost reduction;
- Time-of-Use optimisation: exploits load shifting from high-price intervals to low-price intervals, when tariff schemes favour off-peak consumption;
- Controlled islanding: during grid outages controlled islanding allows to maintain electricity supply behind the meter. This option is largely practised by buildings' complexes, such as higher education facilities, or hospitals where uninterrupted power supply is necessary.

Flexible buildings are an example of prosumer that could help the transition to a low carbon energy system. In the context of building, or building clusters, energy flexibility is defined as the potential for using a building or a set of buildings to perform demand-response [15]. Junker identifies several factors that influence the building's ability to provide energy flexibility [16], among which: the technologies the building is equipped with, such as ventilation, heating and storage, its physical characteristics (insulation, architectural layout), its occupants' behaviour and the associated comfort requirements and lastly its control system, that allows it to respond to external signals such as carbon dioxide (CO_2) emissions or electricity price.

The energy flexibility potential of a building can be quantified deductively, which implies modelling the building with the help of building simulation tools such as City Energy Analyst (CEA) [17], or inductively, exploiting experimental data and time series analysis. Predicting the energy flexibility of a building's energy system involves a lot of challenges related to the prediction of the consumption and generation, as well as the occupancy behaviour and technical and feasibility constraints [18].

2.2 Solar Photovoltaic Energy

A solar photovoltaic system is a technology to generate electric power by using solar cells to convert energy into a flow of electrons exploiting the photovoltaic effect. The basic solar photovoltaic cell is indeed made up by joining a p-type and a n-type semiconductive layers to form a p-n junction diode. The p-type region is composed by a semiconductive material, usually silicon, that has been doped with acceptor impurities and therefore has one fewer valence electron; on the other end, the n-type region is the layer that has been doped with donor impurities and is thus able to cede one additional valence electron to the other layer. When a p-n junction is formed, this concentration difference in electrons results in the creation of a flow between the two layers, which is exactly the working principle of the basic solar cell [19].

The preliminary step to calculate the output of a solar cell consists of implementing a model to compute irradiance on its surface through one of the existing solar radiation model and, in a second moment, implement a solar cell model to compute the power output.

2.2.1 Solar Radiation Models

In most building energy simulation applications involving solar energy, the solar irradiation is usually measured horizontally which leads to the need of calculating the effective solar radiation on the tilted surface of interest. Besides this, the available data usually only includes the horizontal solar radiation, while the measurement of the direct normal and/or the diffuse radiation are not as common but contribute to add an additional level of accuracy. In the absence of direct measurements of direct normal and diffuse irradiance, models become even more important and have to be used to split global irradiance into direct and diffuse irradiance [20].

To this purpose, there exist multiple radiation models to translate the solar irradiation measured on the horizontal plane to the tilted one, given a solar radiation input and appropriate correction factors and clear sky models. The most widely used models used in building energy simulation include the Isotropic Sky, the Klucher, the Hay and Davies, the Hay-Davis-Klucher-Reindl (HDKR) and the Perez model.

2.2.1.1 Models to Compute Global Irradiance on a Tilted Panel

The incident solar radiation on a tilted surface is composed by the sum of the set of radiation streams that include direct or beam radiation, reflected radiation and diffused radiation. The total incident solar radiation on the tilted surface G_T can thus be written in the following form:

$$G_T = G_{T,b} + G_{T,d} + G_{T,r} \tag{2.1}$$

where G_T is the total incident solar irradiance on the tilted surface, $G_{T,b}$ is the beam irradiance, $G_{T,d}$ is the diffused irradiance and $G_{T,r}$ is the ground reflected irradiance. In particular, the beam irradiance $G_{T,b}$ refers to the quantity directly received without any reflection or refraction from the sun in the form of light per surface unit, and it is also known as direct solar irradiance. The reflected irradiance includes instead the ratio of irradiance received from the sun under the form of light after it has been reflected from the surroundings and the various surfaces seen by the panel. Finally, the diffuse irradiance, also known as sky irradiance or solar sky irradiance, is the fraction of total solar radiation which is received from the sun when its direction has been changed by atmospheric scattering. Its direction is highly variable and mainly depends on cloudiness and atmospheric clearness. The diffused radiation is also the combination of three components namely isotropic $G_{T,d,i}$, circumsolar $G_{T,d,c}$ and horizon brightening $G_{T,d,h}$. The isotropic diffuse radiation component is received evenly from the entire sky dome. The circumsolar diffuse part is received from onward dispersion of solar radiation and concentrated in the section of the sky around the sun, whereas the horizon brightening component is concentrated near the horizon and it is most obvious in the clear skies. As a result, the total incident solar radiation received by a tilted surface can be written as:

$$G_T = G_{T,b} + G_{T,r} + G_{T,d,i} + G_{T,d,h} + G_{T,d,c} \tag{2.2}$$

with $G_{T,b}$ the beam radiation, $G_{T,d,c}$ the circumsolar diffuse radiation, $G_{T,d,i}$ the isotropic diffuse radiation, $G_{T,d,h}$ the horizon diffuse radiation and $G_{T,r}$ the reflected radiation. Figure 2.2 shows the different irradiance components.

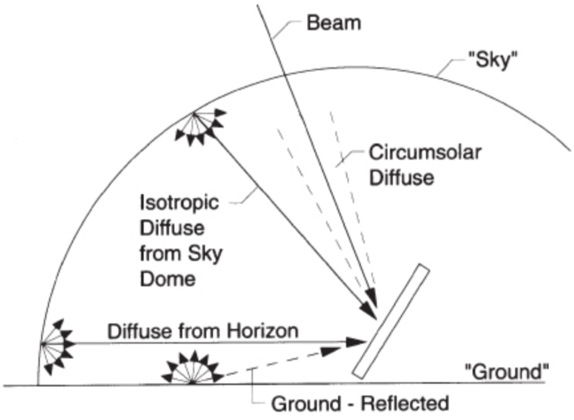


Figure 2.2: Solar radiation components [21].

The following paragraphs present the most common models which can be used to compute the global radiation on a tilted panel, given the global radiation on the horizontal plane. All described models handle beam and reflected radiation in the same way so the major modelling differences are in the calculation of the diffuse radiation. If the time period of the measurements is small enough compared to the typical time constant of the irradiance variations, then the following assumption can be made: $I = G\Delta t$ where I is the radiation in J/m^2, G is the irradiance in W/m^2 and Δt is the time period [21].

- **Isotropic Sky** The simplest possible model to calculate the total irradiance on a tilted panel is the isotropic sky model presented by Liu and Jordan. It relies on the assumption that the circumsolar and horizon brightening components are negligible compared to the others and that ground reflected radiation is diffuse. To put it simple, it assumes all diffuse radiation is uniformly distributed over the sky dome. As a result, the values of these components are only adjusted based on geometrically derived factors and the model usually results in lower numerical values than the real ones. The equation for the isotropic sky model is the following:

$$I_T = I_b R_b + (\frac{1+cos\beta}{2}) + I\rho_g(\frac{1-cos\beta}{2}) \tag{2.3}$$

where R_b is the ratio of the beam radiation on the tilted surface to that on the horizontal surface. R_b is a function of the transmittance of the atmosphere, and for the surfaces sloped towards the equator in the northern hemisphere can be computed as the ratio between the cosine of θ divided by the cosine of θ_z:

$$R_b = \frac{cos\theta}{cos\theta_z} \tag{2.4}$$

- **Anisotropic Sky** The anisotropic sky models are improvements made to the isotropic diffuse model, which take into account the terms that the isotropic model does not account for.

  – *Hay and Davies Model*

  According to the Hay and Davies model, the diffuse radiation is only composed of an isotropic and a circumsolar component, no horizon brightening is present and the reflection from the ground is dealt with like in the isotropic model. To this purpose an anisotropy index A is defined which expresses the quantity of the diffuse radiation treated as circumsolar with the remaining portion of diffuse radiation assumed isotropic. If the anisotropy index is null, then the model reduces back to the isotropic model.

  $$A = \frac{I_b}{I_o} \tag{2.5}$$

  where is I_b the beam radiation and I_o is the extraterrestrial radiation.

  The total irradiance on the tilted surface can therefore be computed as:

  $$I_T = (I_b + I_d A)R_b + I_d(1 - A)(\frac{1+cos\beta}{2}) + I\rho_g(\frac{1-cos\beta}{2}) \tag{2.6}$$

  – *Klucher Model*

  Klucher modified the Hay and Davies model by multiplying the isotropic diffuse irradiance by a clearness index F' accounting for the cloudiness.

  $$F' = 1 - (\frac{I_d}{I})^2 \tag{2.7}$$

  The Hay and Davies performs well in the case of overcast skies but it underestimates the irradiance under clear skies and partly overcast conditions. Under overcast skies, the clearness index F' becomes zero and the model reduces to the isotropic model. The Klucher model calculates the total irradiance as:

  $$I_T = I_b R_b + I_d(\frac{1+cos\beta}{2})\left[1 + F'\sqrt{\frac{I_b}{I}} sin^3(\frac{\beta}{2})\right] \times \left[1 + F'cos^2\theta cos^3\theta_z\right] + I\rho_g(\frac{1-cos\beta}{2}) \tag{2.8}$$

  The modified terms in the diffuse component try to account for the horizon brightening and for the effect of the circumsolar radiation.

  – *HDKR Model*

  Always employing the anisotropy index, in order to account for the horizon brightening as well, Reindl combined the Hay and Davies and the Klucher model, resulting in the HDKR model. The total irradiance on a tilted surface can then be calculated through:

$$I_T = (I_b + I_d A)R_b + I_d(1-A)(\frac{1+cos\beta}{2})\left[1+\sqrt{\frac{I_b}{I}}sin^3(\frac{\beta}{2})\right] + I\rho_g(\frac{1-cos\beta}{2}) \tag{2.9}$$

Because of the additional term in the horizon brightening component, the HDKR model provides slightly higher diffuse irradiances than the Hay–Davies model or the Klucher model.

- **Perez** The more complex and computationally intensive model was developed by Perez. Perez tries to account both for the isotropic diffuse, for the circumsolar and for the horizon brightening radiation by using empirically derived coefficients [22]. The total irradiance on a tilted surface is given by:

$$\frac{I_T}{I} = \left(1-\frac{I_d}{I}\right)R_b + \frac{I_d}{I}F_1\frac{a}{b} + \frac{I_d}{I}(1-F_1)\left(\frac{1+cos\beta}{2}\right) + \frac{I_d}{I}F_2 sin\beta + \rho_g\left(\frac{1-cos\beta}{2}\right) \tag{2.10}$$

The coefficients F_1 and F_2 are the circumsolar and horizon brightness coefficients that depend on the sky condition parameters clearness ε and brightness Δ, whereas the terms a and b take into account the incidence angle of the sun on tilted surface. These terms can be computed according to the following formulas:

$$a = max\,(0,\, cos\theta)\,;\, b = max\,(cos85,\, cos\theta_z) \tag{2.11}$$

$$\Delta = m\frac{I_d}{I_{on}} \tag{2.12}$$

$$\varepsilon = \frac{\frac{I_d+I_{b,n}}{I_d} + 5.53510^{-6}\theta_z^3}{1+5.53510^{-6}t\theta_z^3} \tag{2.13}$$

$$F_1 = max\left[0,\, \left(f_{11}+f_{12}\Delta+\frac{\pi\theta_z}{180}f_{13}\right)\right];\, F_2 = \left(f_{21}+f_{22}\Delta+\frac{\pi\theta_z}{180}f_{23}\right) \tag{2.14}$$

The coefficients f_{11}, f_{12}, f_{13}, f_{21}, f_{22} and f_{23} were derived based on a statistical analysis of empirical data [22], [23] and are reported in Table 2.1.

Table 2.1: Perez coefficients.

[-0.0080	0.5880	-0.0620	-0.0600	0.0720	-0.0220]
[0.1300	0.6830	-0.1510	-0.0190	0.0660	-0.0290]
[0.3300	0.4870	-0.2210	0.0550	-0.0640	-0.0260]
[0.5680	0.1870	-0.2950	0.1090	-0.1520	-0.0140]
[0.8730	-0.3920	-0.3620	0.2260	-0.4620	0.0010]
[1.1320	-1.2370	-0.4120	0.2880	-0.8230	0.0560]
[1.0600	-1.6000	-0.3590	0.2640	-1.1270	0.1310]
[0.6780	-0.3270	-0.2500	0.1560	-1.3770	0.2510]

2.2.2 Solar Cell Model

There are various solar cell models available in literature. For the modelling of the solar cell the the three-diode-one-parameter has been chosen, although more accurate models may exist. When the sun's rays reach the surface of the solar cell, the photon's energy creates free charge carriers. A solar cell under illuminated conditions can be represented by an equivalent circuit that consists of a diode (D) and a power source (see Figure 2.3). The power source (sun) generates a photo electric current (I_s) which has a direct correlation to the level of irradiance (G) [19].

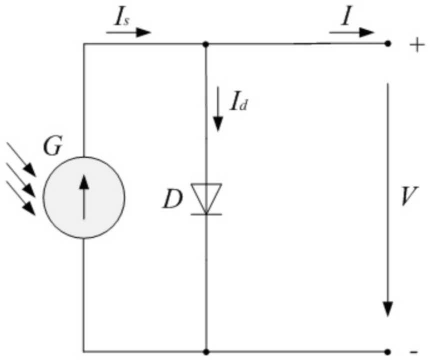


Figure 2.3: Depiction of an ideal model equivalent circuit for an illuminated solar PV cell [21].

2.2.2.1 Three Diode - One Parameter Models

The solar cell has inherent current-voltage $I(V)$ characteristics that are dened by its cell structure, material properties, and the operating conditions. The following equations can be evaluated to determine the $I(V)$ characteristics of a solar cell. The PV panel main standard parameters have been calculated according to the following formulas:

$$V_T^{ref} = \frac{k \cdot T_{cell}}{q} \tag{2.15}$$

$$I_O^{ref} = I_{SC}^{ref} \cdot \frac{e^{m' V_T^{ref}} - 1}{V_{OC}^{ref}} \tag{2.16}$$

$$m\prime = \frac{V_{MPP}^{ref} - V_{OC}^{ref}}{V_T^{ref} \cdot ln\left(1 - \frac{I_{MPP}^{ref}}{I_{SC}^{ref}}\right)} \tag{2.17}$$

$$m = \frac{m'}{N} \tag{2.18}$$

Using the one diode and three parameters' model it is possible to calculate all the data of the simplified circuit at different irradiances and temperature. From these values, the temperature of the cell can be calculated as:

$$T_{cell} = \frac{T_{amb} + (NOCT - 20)}{0.8} \cdot G \tag{2.19}$$

From the cell temperature it is then possible to calculate all the other parameters of the circuit such as the short circuit current, the open circuit voltage, the maximum point voltage and current, the power produced and the efficiency of the cell, according to the following formulas. The short circuit current calculation includes both the effects of temperature and of irradiance.

$$\frac{E_g}{E_g^{ref}} = 1 - C \cdot \left(T_{cell} - T^{ref}\right) \tag{2.20}$$

$$I_{SC_{Tcorrected}} = I_{SC} \cdot \left[1 + \alpha_{I_{SC}} \cdot \left(T_{cell} - T^{ref}\right)\right] \tag{2.21}$$

$$I_O = I_O^{ref} \left(\frac{T_{cell}}{T^{ref}}\right)^3 \cdot e^{\frac{qN}{m'} \cdot \left(\frac{E_g^{ref}}{kT^{ref}} - \frac{E_g}{kT_{cell}}\right)} \tag{2.22}$$

$$e^{\frac{V_{MP}}{m' V_T}} = \frac{\frac{I_{SC}}{I_o} + 1}{1 + \frac{V_{MPP}}{m' V_T}} \tag{2.23}$$

$$V_{OC} = m' V_T ln\left(\frac{I_{SC}}{I_O} + 1\right) \tag{2.24}$$

Having found the maximum power point voltage, it is immediate to find the maximum power point current at the previously calculated I_o at that irradiance and temperature. The maximum power produced in these conditions is then simply found by multiplying the MPP current and the MPP voltage.

$$P_{MPP} = V_{MPP} \cdot I_{MPP} \tag{2.25}$$

2.3 Forecasting using Machine Learning

Nowadays, thanks to technology, observations can be collected about any phenomenon and stored efficiently. This immense amount of data can then be analysed to extract useful underlying information. In particular, although they differ in the approach, there are two main fields that deal with data to extract knowledge, namely statistics and machine learning. Statistics deals with stochastic models to develop parameters to create a model that fits the data well. The goal of machine learning on the other hand, is to build a computer system that automatically learns from the data, disregarding its internal working principles [24]. In machine learning present data can be used to predict future data and the ability of the system to adapt and its predictive accuracy is more important than the model itself. In contrast to statistics, that primarily deals with models to understand the data generating process, the prime machine learning goal is to be able to learn algorithms and replicate the data, especially when considering very large data sets. Although these two fields are different, both can be used to uncover patterns, extract knowledge from data and predict the future from historical time series. These problems are known as forecasting problems, which can be found in areas such as load forecasting, economics and meteorology [24].

2.3.1 Machine Learning Types of Problems

The problems that can be modelled thanks to machine learning mainly fall into three main categories, supervised learning, unsupervised learning and reinforcement learning, although there exist also hybrids between these learning types, like semi-supervised learning [24], [25]. In supervised learning the goal is to model the relationship between features of the input data and a label associated with it. Once the model has learned the relationship between the features and the label, it can be applied to unknown data to predict the corresponding labels. Supervised learning can be further classified into classification and regression tasks; classification involves assigning discrete categories to data, while in regression tasks the output is continuous.

Unsupervised learning, on the other hand, tries to model the features of a dataset without reference to any label, i.e. lets the data create an output without referencing any previous label. Among others, clustering and dimensionality reduction are types of unsupervised learning in which the algorithms try to create distinct groups of data or, more in general, a more succinct representation of it.

The last type of learning, called reinforcement learning, does not rely on existing data, but thanks to a feedback from the environment, the machine's output is evaluated, positively or negatively, about its decisions. The goal of the machine is to maximise its positive decisions, taking advantage of the created prior knowledge. Such type of learning is well suited for situations in which no historical data are available or when such data varies significantly through time [25]. Figure 2.4 gives an overview of the types of machine learning, their methods, output data and an example of field of application in the energy sector.

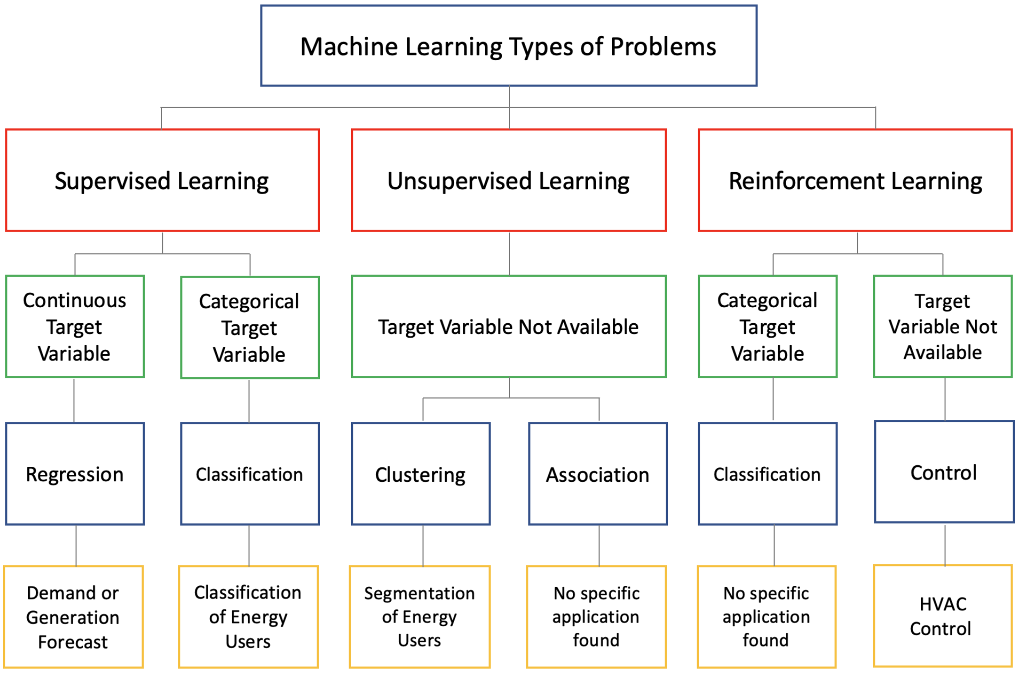

Machine Learning Types of Problems
Supervised Learning
Unsupervised Learning
Reinforcement Learning
Continuous Target Variable
Categorical Target Variable
Target Variable Not Available
Categorical Target Variable
Target Variable Not Available
Regression
Classification
Clustering
Association
Classification
Control
Demand or Generation Forecast
Classification of Energy Users
Segmentation of Energy Users
No specific application found
No specific application found
HVAC Control

Figure 2.4: Machine Learning Types of Problems [own creation based on [24]].

2.3.2 Machine Learning Algorithms

Machine learning algorithms can be classified into twelve groups based on their learning method, approach, and typical applications. Table 2.2 shows a summary of the groups of ML algorithms [24].

Table 2.2: Groups of machine learning algorithms [24].

Group	Learning Method	Learning Approach	Application	Examples
Regression Analysis	Supervised, Reinforcement	Use of relationship between dependent and independent variables estimated through probabilistic method or error function minimization		Linear Regression, Polynomial Regression, Least Squares Regression
Distance-Based Algorithms	Unsupervised, Supervised	Use of the distance between the observed features	Classification, Regression	K-Nearest Neighbours, Learning Vector Quantization, Self-Organising Maps
Regularization Algorithms	Supervised, Reinforcement	Extension of the regression analysis that introduces a penalty term to balance complexity and precision		
Decision Tree Algorithms	Supervised	Use of sequential conditional rules	Decision Making, Classification, Regression	Decision Tree Regression, Random Forest, Conditional Decision Tree
Bayesian Algorithms		Use of inference from distribution of variables	Classification, Inference Testing	Naive Bayes, Gaussian Naive Bayes, Bayesian Network
Clustering Algorithms	Unsupervised	Maximisation intracluster while minimising intercluster similarities		K-Means, K-Medians, Hierarchical Clustering
Association Rule Mining Algorithms		Relationship among variables is quantified for predictive and exploratory objectives		Apriori Algorithm, Context Based Rule Mining
Artificial Neural Network Algorithms	Supervised, Unsupervised	Inspired by the biological neural networks	Powerful enough to model non-linear relationships	Perceptron, Back-Propagation, Hopfield Network, Radial Basis Function Network
Deep Learning Algorithms		Complex neural structures	Capable of abstracting higher level information from huge datasets	Deep Boltzmann Machine, Deep Belief Networks, Convolutional Neural Networks, Stacked Auto Encoders
Dimensionality Reduction Algorithms	Supervised	Apply transformations to data	Amplification of the signal contained in data prior to modeling	Principal Component Analysis, Principal Component Regression, Partial Least Squares Regression, Multidimensional Scaling, Linear Discriminant Analysis
Ensemble Algorithms	Supervised, Unsupervised, Reinforcement	Formed by a combination of multiple machine learning algorithms	Ability to create superior results and possibility to break into independent models to train over a distributed network.	Boosting, Bagging, AdaBoost, Stacked Generalization (blending), Gradient Boosting Machines
Text Mining	Supervised	Subfield of Natural Language Processing	Creation of insights of ML features from unstructured textual data	Sentiment Analysis, Speech Recognition, Topic Modeling

2.3.3 Machine Learning for Electricity Consumption Forecasting

Electricity consumption forecasting in buildings is of the utmost importance when it comes to reducing energy consumption and reaching predefined energy targets. Wei [26] reviews the most common methods to forecast building energy consumption. Among the most common ones in literature artificial neural networks (ANNs) and support vector machines (SVMs) seem to be prevalent because of their capability of modelling non-linear relationships.

The complexity of building energy consumption forecasting is due mainly to how the consumption can be structured: aside from a base load, which is caused by appliances that are running all the time, and seasonal load, attributable to temperature changes and subsequent heating or cooling, the fraction of the consumption that causes difficulties in its prediction is the active consumption. The active consumption is due to the activities in a building and its occupancy patterns [27]. Edwards presents seven machine learning methods to predict residential electrical consumption and shows that Least Squares Support Vector Machine (LSSVM) proves to be the best method among the tested ones [28]. SVMs are though generally considered too complex, especially for a real time implementation [29]. To avoid this drawback of SVMs, Pombeiro proposes low-complexity models to predict electricity consumption, including ANNs and fuzzy systems [30]. In particular fuzzy models prove to be the most reliable in their prediction overpassing ANNs. On the same line, Tso, together with ANNs proposes a traditional regression analysis and a decision tree model. It becomes clear that in an empirical application, the decision tree and neural network models appear to be viable alternatives and both overcome the traditional regression model [31]. Overall, neural networks seem to be the dominant tool applied, either alone or more often together with other optimization techniques, or using deep recurrent neural networks [32], [33], [34], [35].

In the case of complex tools, though, the learning methods and the tuning of the models, still seem to be hindering their complete real-time implementation as the computational impact would be too high. As a result, it may well be that more simple methods, such as decision tree or very simple neural networks, are to be preferred especially in cases where correlations patterns can be easily spotted or a real-time implementation is desired.

2.3.3.1 Summary of some of the reviewed works on electricity consumption forecasting

Table 2.3 summarises some of the reviewed works on electricity consumption forecasting, including the forecast method and the main findings.

Table 2.3: Summary of some of the reviewed works on electricity consumption forecasting.

Authors and Reference	Title	Year	Forecast Method	Main Contributions
Tso et all. [31]	Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks	2007	Regression analysis, decision tree and neural networks	This work compares three modelling techniques to predict electricity consumption, namely regression analysis, decision tree and neural networks. Different models are developed for the winter and summer period. Model selection is based on the square root of average squared error. When comparing the accuracy in the predictions, it is found that the decision tree model performs the best in summer and neural network model perform better than any other model in the winter period.
Edwards et all. [28]	Predicting future hourly residential electrical consumption: A machine learning case study	2012	Regression, FFNN, SVR, LS-SVM, HME-REG, HME-FFNN, FCM-FFNN	This work proposes seven ML methods to predict electrical consumption in commercial and residential buildings, including regression, Feed Forward Neural Networks (FFNN), Support Vector Regression (SVR), Least Squares Support Vector Machine (LS-SVM), Hierarchical Mixture of Experts with Linear Regression Experts (HME-REG), Hierarchical Mixture of Experts with Feed Forward Neural Networks,(HME-FFNN) and Fuzzy C-Means with Feed Forward Neural Networks (FCM-FFNN). Results confirmed that Neural Network-based methods perform best on commercial buildings but poorly on residential data. LS-SVM is the best performing predictor for residential consumption.
Pombeiro et all. [30]	Comparative assessment of low-complexity models to predict electricity consumption in an institutional building: Linear regression vs. fuzzy modeling vs. neural networks	2017	Linear Regression, Fuzzy C-Means, ANN	This work aims at demonstrating that low complexity non-linear models can be used to accurately describe energy consumption baselines. The work compares a linear regression model, a ANN model and a fuzzy model using simple predictor variables such as time-of-day, weather conditions, and occupancy. The developed fuzzy and NN models achieve considerably better performance and accuracy indexes than linear regression models, with the fuzzy model being the one with the highest variance accounted for (VAF).

Table 2.3: Summary of some of the reviewed works on electricity consumption forecasting (continued)

Authors and Reference	Title	Year	Forecast Method	Main Contributions
Escriva-Escriva et all. [36]	New artificial neural network prediction method for electrical consumption forecasting based on building end-uses	2011	ANN	This paper presents an artificial neural network (ANN) method for short-term prediction of total power consumption in buildings with several independent processes. A new prediction method has been presented using a versatile and adaptive algorithm based on artificial neural networks (ANNs) trained with the minimum possible number of days and with features as similar as possible to the day of prediction (DOP). The selection of the days is made with two parameters: labour activity parameter (LAP) to consider work patterns, and weather conditions using the temperature coefficient (CT). Validation of the method has been performed with the prediction of the whole consumption of the Universitat Politècnica de València.
Hassan et all. [37]	Examining performance of aggregation algorithms for neural network-based electricity demand forecasting	2015	ANN	This work examines the efficiency of different aggregation algorithms obtained from individual neural network (NN) models put in an ensemble. An ensemble of 100 NN models is constructed with a heterogeneous architecture and the outputs are combined by three different aggregation algorithms, like simple average, trimmed mean, and a Bayesian model averaging and are employed to obtain forecasts. A comparison of the results shows that the Bayesian model averaging approach has been found as the best combination method to predict electricity demand. The equal-weight is also a good method of combination, however, its result is greatly affected here by the number of NN models that were included in the combination.
Amber et all. [38]	Energy consumption rorecasting for university sector buildings	2017	Statistical analysis, ANN	This study proposed the forecast of electricity consumption of different university buildings at the Southwark Campus of London South Bank University in London through Multiple Regression (MR) technique. The results demonstrate that out of six explanatory variables, three variables, namely surrounding temperature, weekday index and building type, have significant influence on buildings energy consumption. The results showed that the chosen variables were found to be significant in the development of the model which showed a relative good accuracy although its simple mathematical formulation.

Table 2.3: Summary of some of the reviewed works on electricity consumption forecasting (continued)

Authors and Reference	Title	Year	Forecast Method	Main Contributions
Biswas et all. [39]	Prediction of residential building energy consumption: A neural network approach	2016	ANN	This study aims to train a ANN to model building electricity consumption of a heat pump. The used input variables include the number of days, the outdoor temperature and the solar radiation. The model is based on Levenberg-Marquardt algorithms which is able to address the non-linearity of the consumption pattern. The results showed a coefficient of determination above 0.9.
Deb et all. [40]	Forecasting energy consumption of institutional buildings in Singapore	2015	Artificial Neural Network (ANN), Adaptive Neuro Fuzzy Interface System (ANFIS)	The paper presents a methodology to forecast the load consumption used for cooling in three institutional buildings in Singapore. The model is developed using two machine learning tools, ANN and ANFIS, and the energy consumption data is divided into five classes to be used as inputs to the forecasting model. The results show that both ANN and ANFIS forecast the cooling load energy consumption of the three buildings with good accuracy. The ANFIS model needed though to be adapted depending on the considered building whereas in the ANN model there was no major difference in the model development methodology across the three buildings.
Singh et all. [32]	Integration of new evolutionary approach with artificial neural network for solving short term load forecast problem	2018	ANN + different optimisation algorithms	This paper proposes the integration of an ANN with an evolutionary algorithms based on the ‘follow the leader (FTL)’ behaviour of sheep. This algorithm is used in combination with ANN to forecast the electricity consumption thanks to its great feature extraction properties. To validate the proposed algorithm, the results are compared with the ones of other ANN optimised with genetic algorithm (GA), particle swarm optimisation (PSO) and back-propagation neural network (BPNN). The proposed algorithm based on FTL performs the best when compared to more traditional optimisation approaches for ANN, being able to overcome overfitting problems with a good generalisation ability.

Table 2.3: Summary of some of the reviewed works on electricity consumption forecasting (continued)

Authors and Reference	Title	Year	Forecast Method	Main Contributions
Platon et all. [41]	Hourly prediction of a building's electricity consumption using case-based reasoning, artificial neural networks and principal component analysis	2015	Artificial neural networks, Case-based reasoning (CBS)	The papers present two simplified models to predict the hourly electricity consumption of an institutional building. Measurements from a Canadian institutional facility, along with weather forecasts, were used to develop and validate this approach. Two artificial intelligence techniques, artificial neural networks (ANN) and case-based reasoning (CBR), are proposed to develop the forecasting models. Among 22 possible input variables, the 10 most significant ones have been selected thanks to PCA, although no strong difference in the results was visible between models developed with all variables and the ones using only the PCA-selected inputs. The results show that ANN models are more accurate than CBR models in predicting the consumption, although large amount of data are necessary for their accurate training.
Amber et all. [42]	Intelligent techniques for forecasting electricity consumption of buildings	2018	Multiple Regression (MR), Genetic Programming (GP), Artificial Neural Network (ANN), Deep Neural Network (DNN) and Support Vector Machine (SVM).	This paper compares five techniques (Multiple Regression (MR), Genetic Programming (GP), Artificial Neural Network (ANN), Deep Neural Network (DNN) and Support Vector Machine (SVM)) to predict the electricity consumption of a higher education building. The prediction models are based on different parameters such as solar radiation, temperature, wind speed, humidity and weekday index. Results demonstrate that ANN performs better than all other four techniques.
Li et all. [43]	Forecasting building energy consumption using neural networks and hybrid neuro-fuzzy system: A comparative study	2011	ANN, ANFIS-GA	This study compares the performance of traditional ANN with hybrid genetic algorithm-adaptive network-based fuzzy inference system (GA-ANFIS) to carry out building energy consumption prediction. A hierarchical structure of ANFIS is proposed and optimised trough GA. The results show that the performance of the ANFIS is slightly higher than traditional ANN whereas the modelling time is comparable between the two methods. Both models' performance is then tested on two different data sets from the library building of Zhejiang University, China.

Table 2.3: Summary of some of the reviewed works on electricity consumption forecasting (continued)

Authors and Reference	Title	Year	Forecast Method	Main Contributions
Gonzalez et all. [44]	Prediction of hourly energy consumption in buildings based on a feedback artificial neural network	2015	Hybrid ANN	This study proposes a method for short-term load forecast (STLF) in buildings based on ANN trains by means of a hybrid algorithm. The inputs to the model are the current and the forecasted values of temperature, the current load, and time related variables such as date and hour. The results show a comparable performance to similar methods presented in literature. The biggest advantage of such method lies in its simplicity and the reasonable sources needed to apply it to a STLF problem.
Neto et all. [45]	Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption	2008	Physical model, ANN	This paper compares two different approach to forecast building energy consumption, a physical model developed in EnergyPlus and a data-driven model based on ANN. The input features to both models are meteorological data and historical data of energy consumption of the administration building of the University of São Paulo. The main source of error in the EnergyPlus prediction is related to the correct evaluation of lighting, equipment and occupancy schedules. The ANN model shows similar average errors as the EnergyPlus model but manages overall to provide a slightly better prediction for the energy consumption than the EnergyPlus model. However both models could benefit from the integration of more accurate occupant's behaviour data that would probably lead to more reliable models.
Bento et all. [46]	Optimization of neural network with wavelet transform and improved data selection using bat algorithm for short-term load forecasting	2019	ANN, Elman NN, RBFN, SVM	The paper presents an enhanced ANN method to carry out one day ahead forecasts. The method includes data selection and features extraction through correlation and wavelet analysis and a combination of Bat and Scaled Conjugate Gradient Algorithms to improve neural network learning capability of the feed forward neural network. The testing is carried out on the load of the Portuguese national system, the load of the city of New York and the load of New England. The proposed ANN method is then compared to other forecasting techniques such as Elman NN, a Radial Basis Function Network and Support Vector Machine and its superior effectiveness acknowledges in all cases.

Table 2.3: Summary of some of the reviewed works on electricity consumption forecasting (continued)

Authors and Reference	Title	Year	Forecast Method	Main Contributions
Ribeiro et all. [47]	Enhanced ensemble structures using wavelet neural networks applied to short-term load forecasting	2019	MLP, regression trees, wavelet ensemble, naïve models	The paper proposes a pipeline to build wavenet ensembles to predict load from real data. An ensemble aggregation algorithms composed of wavenet learners is therefore trained with a subset of selected features with hourly load values. The results are then compare with other forecasting techniques such as MLP, regression trees, regression and naïve models. The results show that the proposed ensemble outperforms all other methods.
Zhang et all. [48]	A composite k-nearest neighbour model for day-ahead load forecasting with limited temperature forecasts	2016	kNN ensemble	This study developed an enhanced method to forecast one day ahead electricity consumption based on a k-nearest neighbour (kNN) model. Th model uses an input only very limited features such as the minimum and maximum temperature of the day. Three individual kNN models are combined into an ensemble to improve the prediction accuracy. The final model is tested with real-world consumption data and it shows a reasonable accuracy in its prediction confirming that it can be used as an alternative tool for day-ahead load forecasting when only limited information is available.
Fan et all. [49]	Application of the weighted K-Nearest Neighbour algorithm for short-term load forecasting	2019	kNN, ANN, ARIMA	This paper presents the forecasting of short-term electricity load based on based on the weighted k-nearest neighbour (wKNN) algorithm. The model was applied to develop the forecasts of the Australian load considering the inverse of Euclidean distance as the weight. The results are compared to the ones achieved with back-propagation neural network and the autoregressive moving average (ARIMA) models and show that the proposed wKNN achieves greater forecasting accuracy and effectiveness. The proposed wKNN method is able to reflect well the variation in the power and shows good fitting ability over a short forecast horizon.

2.3.4 Machine Learning for Solar Photovoltaic Production Forecasting

The forecasting of solar photovoltaic production becomes increasingly important to mitigate the impact of the intermittent nature of solar power and the increasing penetration of renewables in the electric grid thanks to new legislations favouring self-consumption and the increasing deployment of large-scale PV power plant. In particular, among machine learning solar photovoltaic power forecasting two main approaches prevail: indirect forecasting and direct forecasting. Indirect forecasting implies the prediction of solar irradiance or of meterological prediction of weather variables to then use them as inputs to a physical model of the considered PV plant. On the other hand, direct forecasting aims at the direct estimation of the solar power output.

In his work, Rana compares different direct methods of forecasting using an ensemble of neural networks and support vector regression algorithm, both using only previous power data and combining the previous power data with weather variables. The results show that the ensemble of neural networks performs better than the SVR for forecast horizons up to one hour, even using only previous power data and no weather data in the input features [50]. Shi tries to increase the perfomance of a one-day ahead predictive model by distinguishing different weather conditions. First a classification of the weather is performed (clear sky, cloudy, foggy or rainy) and then a SVM is modeled to forecast the power output. Results show that training different SVMs for different weather types increases the accuracy [51].

Similarly to Shi, Liu predicted the solar power output 24h ahead thanks to four different backpropagation NN based on a classification of the type of day with the addition of aereosol index data to PV power data and meteorological data [52]. Using only historical power output data of the panels, Pedro compared the performance of ARIMA, k-nearest neighbours, NN trained with a backpropagation algorithm and NN trained with a GA. The results showed that the two NN-methods outperformed all other proposals [53].

Like Pedro, Chen employed an ANN to carry out a 24h ahead forecast exploiting power data from the previous day and weather forecast for the next day. The results showed the importance of selecting meaningful features to improve the forecast [54]. The importance of feature management is even more evident from the work of Davo, who used principal component analysis (PCA) as feature selection method before implementing ANN. A comparison between a model including PCA and the one developed without showed that using PCA enhances the prediction accuracy [55].

Persson used Gradient Boosted Regression Trees (GBRT) to forecast solar energy power production from 1 to 6 hours ahead using historical power output as well as meteorological features. The GBRT model performed better than persistence models and climatology model on all forecast horizons [56]. Overall, the common result in literature concerning solar power forecasting is that independently of the technique employed, feature selection is the most impactful parameter on the accuracy of the prediction for all forecast horizons. Multiple authors propose the exploration of deep learning techniques in combination with proper feature management to try to overcome the limitations of the unpredictability of weather conditions that characterises solar power.

2.3.4.1 Summary of Some of the Reviewed Works on PV Power Forecasting

Table 2.4 present a summary of the consulted papers on solar power forecasting, specifying the main findings, the methods used, the forecast horizon and the metrics employed to measure the forecasting performance.

Table 2.4: Summary of Some of the Reviewed Works on PV Power Forecasting

Authors and Reference	Title	Year	Forecast Method	Forecast Horizon	Forecast Error	Main Contributions
Shi et all. [57]	Forecasting power output of photovoltaic systems based on weather classification and support vector machines	2012	Weather classification and SVM	1 day ahead	MRE 8.64%	The forecast of PV power generation is based on weather classification and SVM. Days are classified based on weather conditions (clear sky, foggy, cloudy, rainy) and subsequently a SVM model with 4 sub-models was developed and trained based on the input vectors of similar category data.
Khang et all. [58]	Development of algorithm for day ahead PV generation forecasting using data mining method	2011	K-means clustering method	24h ahead	MAPE 11%	Five years of data were analysed and classified according to rainfall probability patterns. The processed data were then used in the PV power forecasting model. The authors suggest to install digital camcorders to detect the status of the clouds and modify the algorithm to account for that so that the prediction error will decrease.
Rana et all. [59]	Forecasting solar power generated by grid connected PV systems using ensembles of neural networks	2015	NN ensemble methods, ANN	30 minutes ahead	Mean Relative Error (MRE) 16.9%-17.6%	This study compared three different approaches of ensembles of ANN to forecast solar power generation. One method was iterative and the other two are not. The results showed that the iterative approach is the most accurate one and that the ensemble of NN improved the forecasting compared to a single ANN.
Pedro at al. [53]	Assessment of forecasting techniques for solar power production with no exogenous inputs	2012	Persistent model, ARIMA, kNN, ANN, ANN with GA	1h to 2h ahead	nRMSE 13.07% (1h) and 18.71% (2h)	The study compared 5 different forecasting techniques (persistent model, ARIMA, kNN, ANN, ANN+GA) without the use of exogenous inputs. The ANN models outperformed all other models and their optimisation with GA improved the accuracy even further. The accuracy of all models highly depends though on the seasonal characteristics of solar variability.
Liu et all. [52]	An improved photovoltaic power forecasting model with the assistance of aerosol index data	2015	Back propagation ANN	24h ahead	MAPE 7.65%	A model to forecast PV power which takes into account aerosol index as an additional input parameter is proposed. A back-propopagation NN which exploits a previous seasonal weather classification is developed. The results showed that the proposed model improves the prediction accuracy when aerosol index is included as a feature.

Table 2.4: Summary of Some of the Reviewed Works on PV Power Forecasting (continued)

Authors and Reference	Title	Year	Forecast Method	Forecast Horizon	Forecast Error	Main Contributions
Rana et all. [50]	Univariate and multivariate methods for very short-term solar photovoltaic power forecasting	2016	Ensemble of ANN and SVM	5min to 60 min ahead	MRE 4.15%–9.34%	This paper analysed the forecasting of solar power for horizons from 5 to 60 min ahead, from previous PV power and meteorological data. A small set of informative variables are used as inputs for an ensemble of NN and SVM. Two types of models are developed: a univariate model, that uses only previous PV power data, and a multivariate model, that also uses previous weather data. Both model perform similarly, thus the PV power output for very short-term forecasting horizons of 5–60 min can be predicted accurately by using only previous PV power data, without considering weather information.
Chen et all. [54]	Online 24-h solar power forecasting based on weather type classification using artificial neural network	2011	SOM and RBFN	24h ahead	MAPE 6.36%-15.08% for sunny and cloudy days, MAPE 24.16%-54.46% for rainy days	The method uses as input past power measurements and meteorological forecasts of solar irradiance, relative humidity and temperature. First the self-organizing map (SOM) is trained to classify the local weather type to subsequently train a radial basis function network (RBFN). Results show that the model developed is very suitable to forecast sunny and cloudy days and it provides reasonably good results for rainy days.
Persson et all. [56]	Multi-site solar power forecasting using gradient boosted regression trees	2017	Gradient boosted regression tree (GBRT)	1h to 6h ahead	nRMSE 10%-15%	The power output of 42 PV rooftop installations in Japan is forecast. To this scope, historical power generation and relevant meteorological variables are used to train a gradient boosted regression tree (GBRT) model for forecast horizons from 1h to 6h ahead. When compared to single-site linear autoregressive and variations of GBRT models the multi-site model shows competitive results in terms of root mean squared error on all forecast horizons. Feature analysis shows that variables related to lagged observations are more important for shorter forecast horizons, whereas for longer horizons the importance of weather forecasts increases. The drawback of the model is that it does not present simple updating procedures.

Table 2.4: Summary of Some of the Reviewed Works on PV Power Forecasting (continued)

Authors and Reference	Title	Year	Forecast Method	Forecast Horizon	Forecast Error	Main Contributions
Yagli et all. [60]	Automatic hourly solar forecasting using machine learning models	2019	68 different methods	1h	nRMSE, nMBE (normalised mean bias error)	68 ML methods, such as tree-based, linear/nonlinear, kernel methods, boosting/bagging/Bayesian variants, quantile regression etc, are tested using irradiance data from 7 locations in 5 different climate zones in the USA. Forecast results are evaluated in both short- (daily) and long-term (over a period of two years) averages, under 3 different sky conditions (overcast, clear-sky, and all-sky), including the trade-off between training time and model performance in the performance assessment. Tree-based methods were found superior in long-term average nRMSE under all-sky conditions, whereas variants of MLP and SVR were the best performers under clear-sky conditions. Random forest quantile regression (RFqr) performed consistently well under overcast skies at all locations. None of the methods was found to be dominating in terms of nMBE, except for RFqr under overcast-sky conditions and daily results showed how forecast performance of a method could change with sky conditions.
Mellit et all. [61]	A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy	2010	ANN	24h ahead	MAE 2.75%-4.48% Co-efficient of correlation R 90.8%-94.14%	A MLP network is developed to forecast 24 h ahead solar irradiance for a PV power plant. The model accepts as input parameters the mean daily irradiance and the mean daily air temperature. A comparison between the power produced and the one forecasted shows a good predicted performance for sunny days (correlation coefficient >98%), slightly lower for cloudy days (95%). Results showed that the model can be easily improved by adding more input parameters such as cloud, pressure, wind speed, sunshine duration, geographical coordinates and etc. if available.
Mandal et all. [62]	Forecasting power output of solar photovoltaic system using wavelet transform and artificial intelligence techniques	2014	Wavelet transform and RBFNN, BPNN	1h ahead	MAPE 2.38% for sunny days and 4.08% for cloudy days	The model combines Wavelet Transform (WT) to filter the spikes and the changes in PV power and meteorological time series data and radial basis function neural network (RBFNN) to model the non-linear fluctuations of PV power production. The combination of WT and RBFNN outperforms the simple RBFNN. The model performance remains unsatisfactory for rainy days.

Table 2.4: Summary of Some of the Reviewed Works on PV Power Forecasting (continued)

Authors and Reference	Title	Year	Forecast Method	Forecast Horizon	Forecast Error	Main Contributions
De Giorgi et all. [63]	Photovoltaic power forecasting using statistical methods: impact of weather data	2014	Statistical methods based on MLR and Elman NN	1h to 24h ahead	nRMSE 10.91%-23.99% nMAE 6.5%-19.5%	This work proposes statistical methods based on multiregression (MR) analysis to analyse the impact of different input features, and the Elmann artificial neural network (ANN) to predict power production of a grid-connected PV plant in Italy. Different combinations of inputs to the ANN were tested and results showed that the best performance is found when all of the weather parameters, including PV power output data, are considered as the inputs to the model.
Dolara et all. [64]	Comparison of different physical models for PV power output prediction	2015	Physical Model	24h ahead	Normalized MAE (nMAE) <1% Weighted MAE (wMAE) <2%	This study investigates three physical models for forecasting the PV power generation by monocrystalline and polycrystalline PV panels were evaluated in this study. The three models are based on three, four and five parameters respectively. The comparison was performed using actual weather data measured by a meteorological station. The results showed that the accuracy of the model depends on the data used for the calibration and the calculation of the cell temperature. Moreover, a training period was not required in this approach unlike ANN-based forecasting models. This approach is suitable for the initial period of a PV plant.
Mori et all. [65]	Development of GRBFN with global structure for PV generation output forecasting	2012	GBRFN, DA and particle swarm optimization	24h ahead	Maximum error 0.228 pu	This work uses deterministic annealing (DA) to determine the center and width of the RBF in a generalised radial basis function network (GRBFN). Weight decay technique was employed to avoid overfitting the learning data of complicated non-linear time series and Evolutionary Particle Swarm Optimization (EPSO) was used to optimize weights between neurons in GRBFN. The simulation results presented that the proposed model significantly reduced the errors in comparison with conventional ANNs, such as MLP, RBFN, and GRBFN.

Table 2.4: Summary of Some of the Reviewed Works on PV Power Forecasting (continued)

Authors and Reference	Title	Year	Forecast Method	Forecast Horizon	Forecast Error	Main Contributions
Xu et all. [66]	Short-term photovoltaic power forecasting with weighted support vector machine	2012	Day selection algorithm and weighted SVM (wSVM)	1h ahead	Mean square error (MSE) 21.8	This work proposes a method based on weighted support vector machine (wSVM) to forecast short-term PV power generation. To train the algorithm, the data for the 5 most similar days to the one to be forecast were used as inputs to the machine. The weights in the wSVM are determined based on similarity measurements. Results showed that the proposed wSVM makes more accurate forecasts than ANN.
Silva Fonseca Junior et all. [67]	Forecasting regional photovoltaic power generation - A comparison of strategies to obtain one-day-ahead data	2014	Principal Component Analysis and SVR	24h ahead	RMSE 10.24%	This paper proposes three strategies to forecast one day ahead PV power generation using support vector regression (SVR) and past PV power data and weather data as input features. These three strategies differentiate themselves depending on the availability of the input data. The strategy that proved to be the best is the one whose input data are chosen based on a principal component analysis (PCA), which, if applied, can lead to meaningful improvements of regional forecast errors.
Leva et all. [68]	Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power	2017	ANN	24h ahead	nRMSE 12.5% - 36.9%	An ANN model was proposed to forecast the solar power output of a PV plant by assessing its performance during sunny, partially cloudy and cloudy days. A sensitivity with respect to the input data and the amount of data is carried out. The results shower that the accuracy of the model is strictly related to the pre-processing of the data and its accuracy.
Cococcioni et all. [69]	24-hour-ahead forecasting of energy production in solar PV systems	2011	Time series analysis and ANN	24h ahead	MAPE <5.0%	The model implemented a 1 day forecast model to forecast PV power based on ANN with tapped delay lines. A tool was proposed to configure the model correctly according to the installation characteristics.

2.3.5 Discussion of Main Findings

The main findings of the literature review presented in the previous sections are that there is no unique criteria to classify PV power generation and building electricity consumption forecasting, however different categories and methods can be identified depending on the forecasting time horizon, on the available data and on the methods used.

Classification based on Forecast Horizon The forecast horizon is the period of time in the future for which the power generation or consumption is forecast and it greatly influences the purpose and accuracy of the models. For example, Lipperheide proposes a method to analyse the spatio-temporal variability of solar power over different forecast horizons [70]. In his analysis, it is evident how the forecast horizon influences the accuracy of the proposed model. In a similar fashion, Lonij shows how, given the same model parameters for the same model, the forecasting accuracy changes with respect to the forecast horizon [71].

A number of other works state the importance of defining the appropriate forecast horizon to consider before the development of the forecast model, especially as far as the forecasting of photovoltaic solar energy is concerned. In buildings instead, the forecasting of consumption is less related to seasonal and unexpected variable effects but more related to occupancy and activities taking place in the building which could possibly be more easily known in advance. As a result, three main categories of forecasting horizons can be identified:

- **Short-Term Forecasts:** Short-term forecasting refers to the forecasting for the next hour, several hours, one day up to a week. In power production such type of forecasting is mostly useful when scheduling and dispatching electrical power and to enhance the security of the grid. In buildings instead, it is useful to apply energy management strategies and scheduling of potential load-shifting.
- **Medium-Term Forecasts:** Medium-Term forecast are more common for power generation than in the prediction of building energy consumption and are mainly used to predict the availability of electric power in a more distant future, i.e. more than a week up to a month. It is common to carry out a medium-term forecasting to plan the power system and its maintenance.
- **Long-Term Forecasts:** Forecasting horizons longer than a month both in electricity production and consumption fall into what is called long-term forecasting, whose main goal is the planning of electricity generation, consumption, transmission and distribution or the creation of more general renewable energy production scenarios.

Classification based on Forecasting Method and Available Data Depending on the used and available input data, it is possible to identify three main types of forecasting methods, namely engineering, statistical and machine-learning methods. Both in electricity production and consumption, engineering methods involve physical principles based on mathematical equations to carry out a forecast and are usually based on commercially available software, such as EnergyPlus or other open-source engineering software. These tools are quite commonly used and proved to be precise in prediction but they required a wide range of data as input, which might be difficult to collect, such as building layout, constructions, conditioning systems (lighting, HVAC, etc.), utility rates or weather data, and are in most cases time-consuming when it comes to their implementation.

Statistical methods instead are mostly used as benchmark models and predictions are made based on statistical analysis of the different input variables. Statistical methods though can be relatively simple methods which need much less data compared to engineering methods but often fail in predicting accurately. Among machine learning methods instead, ANNs and SVMs are the most common methods implemented and the most popular ones in research [26]. Both methods are based on historical data and are usually quite successful at modelling the power generation and consumption as they are able to capture the non-linearity in the data without any prior assumption. Table 2.5 summarises the most commonly used forecasting methods used for building energy consumption and solar photovoltaic power output prediction.

Table 2.5: Comparative analysis of the commonly used methods for the prediction of energy consumption and solar power production [72], [73].

Methods	Model complexity	Easy to use	Running Speed	Inputs needed	Accuracy
Elaborate engineering	Fairly high	No	Low	Detailed	Fairly high
Simplified engineering	High	Yes	High	Simplified	High
Statistical	Fair	Yes	Fairly high	Historical Data	Fair
ANNs	High	No	High	Historical Data	High
SVMs	Fairly high	No	Low	Historical Data	Fairly high

2.3.6 Selected Machine Learning Algorithms for Regression Problems

Following the main findings from literature (section 2.3.2 and 2.3.3), four ML algorithms for regression problems have been chosen to be tested to carry out the forecasting of building electricity consumption and solar power generation. The mathematical derivations for each algorithm are not presented fully because the intention is simply to give a basic understanding that will be relevant in the implementation of the chosen models for this thesis. The general theory and mathematics behind the algorithms that have been implemented in this project is briefly outlined below.

2.3.6.1 Multiple Linear Regression

Linear regression aims to minimize the error between the observations and the estimations by measuring a predefined error function (e.g. quadratic error, etc.). The term linear refers to a linear relationship between two or more variables, whose relationship in a two-dimensional space is represented by a straight line. In linear regression the task is to predict a dependent variable (y) based on a given independent variable (x) to undercover the coefficients of the linear model that explain such relationship. In the case of a univariate model the equation that represents such relationship is of the form:

$$y = mx + b \tag{2.26}$$

where y is the variable to predict, x is the input variable, b is the intercept and m is the slope of the straight line. The values to be optimised in a regression algorithm are therefore m and b. Multiple straight lines can exist, depending on the parameters of intercept and slope and in this case, the linear regression algorithm fits multiple lines to the data and returns the one resulting in the smaller error. Extending this concept to more than two variable results in a multiple linear regression, as the dependent variable, or target variable, depends on multiple independent variables. Such a model can be represented by:

$$y = b_0 + m_1 b_1 + m_2 b_2 + ... + m_n b_n \tag{2.27}$$

In particular, the multiple linear regression model proposed aims at minimising the total sum of squares (SST) resulting from the addition of the error of the sum of squares (SSE) and the regression sum of squares (SSR):

$$min\left(\sum_{i=1}^{n}(y_i - \overline{y_i})^2\right) = min\left(\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 + \sum_{i=1}^{n}(\widehat{y_i} - \overline{y_i})^2\right) \tag{2.28}$$

where y_{i} is the real value of the observation, $\overline{y_i}$ the mean value of y_{i} of n observations and $\widehat{y_i}$ the value modelled by the regression model.

Although linear models are quite widely used in baseline scenarios predictions, they have limited capacity to model non-linear relations between variables.

2.3.6.2 k-Nearest Neighbours Regression

The k-Nearest Neighbour algorithm working principle is different from other methods as it uses a local learning approach, while other methods use a global learning approach [74]. Global learning tries to map all possible input features to an output by creating a function, i.e. fitting a distribution over the data. This is possible because of the assumptions that the treated data was originally generated by a function. In contrast to this, the kNN algorithm, or more in general local learning, denies the existence of an underlying function and only exploits local data. A kNN algorithm is the perfect example of a lazy-learning algorithm, as no global model of the entire domain is kept, but the computation is deferred until an output is requested. At this point, single outputs are mapped by selecting data similar to the input feature.

As a result, the assumption the algorithm makes about the data are on average weaker than other algorithms but as a consequence, it adapts well to various data sets, provided they are quite small. The computational demand increases indeed linearly with the size of the data set because the algorithm tries to take into account a bigger number of training samples and of observations that need to be considered to find the k nearest neighbours.

The kNN algorithm uses a function to determine the similarity of points (the k neighbours) which are the closest to the input point according to some distance metric. In the case of regression, a prediction is made by averaging the output of the k neighbours nearest to the given input feature:

$$y = \frac{1}{k}\sum_{i=1}^{k} y_i \tag{2.29}$$

where y_i is the nearest neighbour. The k parameter k is a very sensitive parameters that control the fit of the algorithm (the bias-variance trade-off, see section 3.3.1.4). Higher values of k involve more neighbours contributing to the output and therefore a smoother fit to the training data, a lower variance and a high bias, and the opposite for smaller k values. To improve performance of the algorithm, two main parameters can be modified: the distance function and the function that averages the outputs of the k nearest neighbours, although the most common approaches include using the Euclidean distance and a weighted averaging function so that points close to each other contribute more to the prediction. Overall this algorithm is simple, versatile and easy to implement and it often performs fairly well, providing results that can be easily interpreted.

2.3.6.3 Decision Tree Regression

A decision tree is a non-parametric model governed by simple decision rules inferred from the input data. Decision trees present three types of nodes: root nodes, i.e. all features which are going to be split, decision nodes, which split the samples into other sub-trees or leaf nodes based on the chosen decision rule and leaf nodes which indicate the final region or class defined by the tree. Like other machine learning methods, depending on whether the variable is continuous or not, they can be used both for classification and regression (a regression tree). The decisions at the splitting point are usually taken to reduce the variance in the target value. When new data falls into a node, its predicted value is the mean of all the samples in that class. The main advantage of decision trees over other machine learning methods is their simple interpretability, which leaves decisions traceable along the tree and allows the formation of clear rules from them.

2.3.6.4 Artificial Neural Network

Artificial neural networks are non-linear computational models inspired by biological neural networks. Their working principle attempts indeed to mimic the human nervous system and its continuous dynamics. A typical ANN topology includes layers and neurons, the basic unit of the artificial nervous system. In the brain, neurons transfer continuous information between them and through the various layers of the cortex. In the same way, in artificial neural networks, there are typically three sequential layers, an input layer, a hidden layer and an output layer. Each layer has a specific number of neurons and each neuron possesses an activation function that triggers the exchange of information. The simplest type of ANN is the Perceptron. As depicted in figure 2.5, the Perceptron take several inputs (x), multiplies them by specific weights (w) to produce an output ($\hat{y}$).

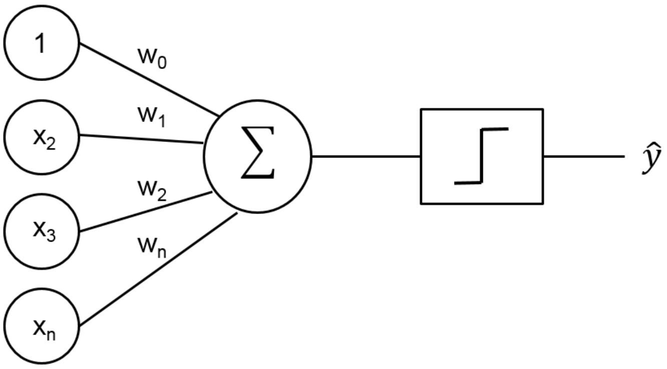


Figure 2.5: Perceptron scheme.

To characterise ANNs there are three parameters to be set: the interconnection pattern between the neurons of the different layers, the learning process of updating the weights of the interconnections, and the activation function that converts a neuron's weighted input to its output activation [75]. When an ANN presents multiple layers it forms a multilayer perceptron (MLP) (Figure 2.6).

By applying different weights to the neurons in the different layers, adaptive models can be developed and more complex functions can be modelled. In particular, in supervised learning Feedforward Neural Network with Back Propagation are a commonly used type of ANN. The term feedforward refers to the direction of the propagation of information. Once divergences are found between the input and the desired output, they are propagated back to the previous layers. The number of input and output neurons depends on the number of chosen input features to the model, whereas the number of output neurons corresponds to number or outputs of the model. Finding an optimal number of hidden layers and of neurons in the hidden layers is rather demanding. It is important to find a trade off between the network architecture and the accuracy of the task to be solved. A wrong number of neurons will either lead to overfitting or generalise the model too much at the point that it will not be capable of solving the task it is meant for. Higher numbers of neurons and layers allow to solve very complex non-linear tasks even on large datasets of theoretically any type of data.

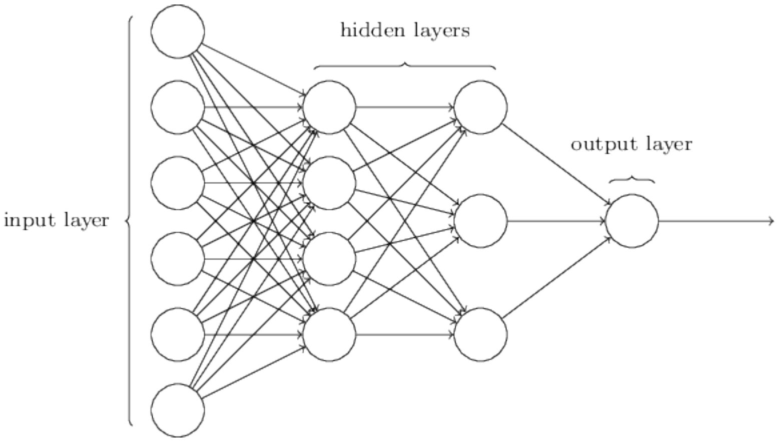


Figure 2.6: Feedforward neural network with back-propagation scheme.

2.4 Demand Response

Within the energy management measures, demand-side management (DSM) is the portfolio of measures that focuses on the consumer side of the energy network. It includes a variety of measures ranging from energy efficiency measures, to real-time control of distributed energy resource to incentive based tariff schemes [76]. To this purpose, the energy flexibility potential (i.e. potential for using a building to perform demand-response, see section 2.2.1) of a building can be assessed using building simulation tools or thanks to experimental data and time series analysis. Assessing the flexibility of a building based on price-responding conditions using time series analysis is not a new concept, as it was already proposed by Corradi [77]. The authors propose a dynamic model to control heating systems and reduce the peak consumption in response to time-varying prices. In a similar fashion, Dorini develops a chance-constrained optimization framework using real-life data to estimate the flexibility and shift the load from peak hours to off-peak hours [78]. Dynamic characterisation of flexibility, for example of a building, has been vastly researched as it has been highlight as one of the priorities in the EU Winter Package [79].

One mechanism to increase flexibility is demand-response. Demand response consists of a change in the consumption pattern of a customer which can either reduce or shift its peak consumption to off-peak hours, thus helping to balance power supply. Depending on the strategy used demand-response involves the shift of the consumption based on different electricity tariff schemes, as for example time of use (ToU), critical-peak price (CPP), and real-time price (RTP). Otzurk for example, proposed a home energy management system (HEMS) in which the appliances were scheduled exploiting the ToU pricing scheme using a controller based on the branch-and-bound algorithm to minimise overall costs [80]. Kinhekar on the other hand, tries to accommodate both benefits for the customer and the utility company implementing an integer genetic algorithm to shift the load by fitting the consumption power curves to the utility production curve at each time step [81]. Another example of load scheduling for demand-response was proposed by Setlhaolo [82]. Setlhaolo implemented a mixed-integer non-linear optimisation model which showed that the households were able to achieve electricity cost savings up to 25% thanks to the scheduling of home appliances in response to the varying prices of ToU.

Independently of the chosen strategy, it is clear that scheduling techniques and accurate forecasts of consumption and of generation, if present, help the customers to manage their load properly. Because of the difficulty in predicting the load consumption, forecasts are of great help in the assessment of demand-response strategies, especially if combined with load scheduling techniques. Demand-response schemes need a scheduling techniques to manage loads especially in the presence of non-shiftable loads that cannot be operated at different times. A solution to this is provided by battery energy storage system, as they allow to partially shift load for optimal cost.

2.4.1 Battery Energy Storage System

The importance of energy storage is to mitigate energy fluctuations, whether on the production or on the consumption side, is widely recognised [83], [84] and contributes to the increasing importance that energy storage systems are playing in the future of the smart grid. Energy storage options range widely in size, application, geographical applicability, energy conversion process and response timing. The most common ones include flywheels, compressed air energy storages, electro-chemical batteries, and large thermal storage tanks. Figure 2.7 summarises the most common storage systems based on their energy conversion process (electro-mechanical, electro-magnetic, electro-chemical and thermal) [85].

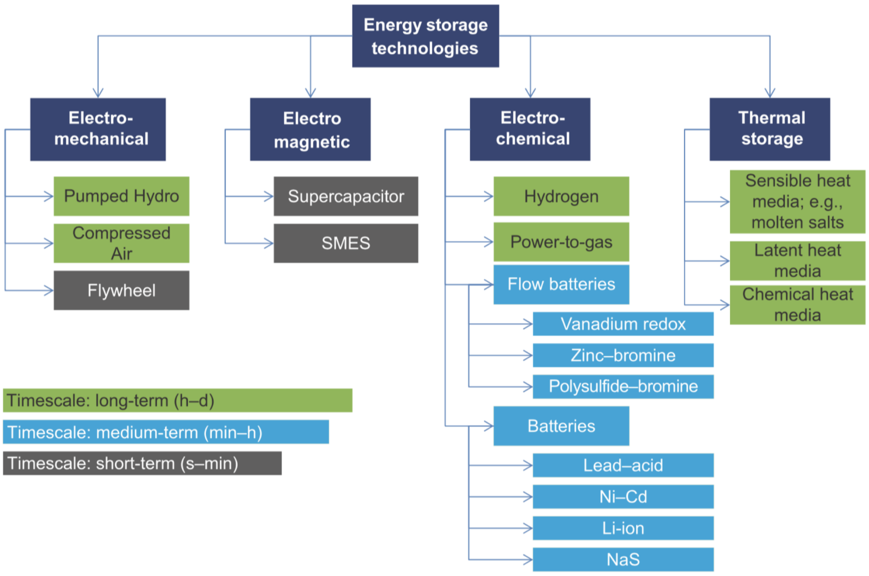


Figure 2.7: Classification of different energy storage systems [86].

The increasing employment of energy storage technology has been caused by the fast growing implementation of renewable energy sources and by the possibility to exploit financial incentives related to the services such a technology can offer. Figure 2.8 summarises the most common services provided by storage technologies, in particular by BESS, depending on whether their application is centralised and connected at the transmission or distribution level or decentralised and integrated behind the meter [86].

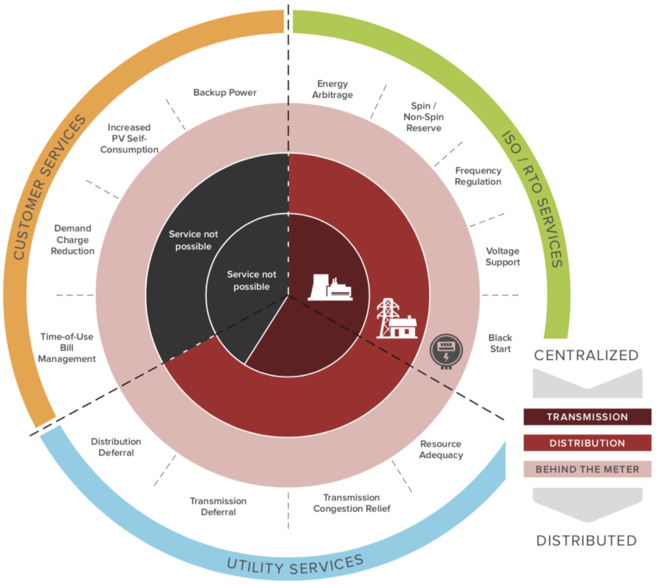


Figure 2.8: Services provided by BESS [86].

Among all storage technologies, the most common employed ones to perform load shifting at customer level are electrochemical battery energy storage systems. In particular, following a drop in their prices Lithium-Ion (Li-Ion) batteries have become the most popular technology in stationary and mobile applications and have been extensively studied as a form of DR.

Kishore provided an example of direct load control thanks to a home energy controller to take advantage of the two level pricing scheme of the utility company. The applied optimization scheme could further be extended to multiple buildings in the neighbourhood while reducing costs [87]. Prasatsap proposed an approach to determine the optimal capacity of battery energy storage system (BESS) for peak shaving at Naresuan University in Thailand. The results show that the optimal capacity for the BESS is successful in shaving the peaks of consumption during high-price periods and that an oversized BESS could further decrease peaks but would results in reduced savings [88]. Lorenzi compares the potential of demand-response exploiting storage in batteries and domestic-hot-water to reduce the bill in the residential sector. The results suggest that with the current market prices of batteries, demand-response using hot-water tanks should be preferred but a significant decrease in the batteries' price could make storage an interesting alternative [89].

Kim presented a method to size and integrate BESS to reduce a building's annual energy costs [90]. The results highlighted that a significant reduction can be achieved, although the compensation of the upfront investment is not always granted and is highly dependant on the battery size and the achievable profit. Semigran and Tsim (2014) quantified the savings derived from using a battery, showing that significant reduction in cost can be achieved to partly compensate for the upfront cost of buying a battery.

Optimal energy consumption schemes based on ToU BESS scheduling show vast potential to shave peak energy consumption and reduce the electricity bill. Such solutions become increasingly interesting especially if combined with decentralised energy generation such as photovoltaic energy to decrease consumption in peak hours, and learning algorithms that could increase the knowledge of the expected consumption to boost economic savings.

Chapter 3

Model Development

This chapter provides an explanation as to how the different models were developed and implemented and how their performance was assessed. All the simulations were carried out using Python and its available packages, which allowed for the calculation of different models such as the technical model developed to calculate the photovoltaic electricity production, the two machine learning models used to carry out the forecasting of demand and supply and the optimization and sensitivity analysis concerning the energy management system proposed, as well as different performance indicators.

3.1 Case Study: Instituto Superior Técnico

As stated in the objectives of this thesis, the developed models have been applied and tested to carry out the analysis of the main campus of Instituto Superior Técnico, an engineering higher education and research facility situated in Lisbon, Portugal. Due to the high number of buildings present on its main campus, the Alameda campus, the analysis focused on four main buildings, that correspond to 50% of the total demand, namely the civil building, the central building and the south and north towers. The exact position of these buildings is highlighted in Figure 3.1.

To be able to assess a possible energy management strategy and carry out the forecasting for both the electricity consumption and production, three main models have been developed:

- a technical model to assess the power output from a predefined number of rooftop photovoltaic panels, which are planned to be installed on the buildings;
- a machine learning model to carry out a short-term forecasting of both the electricity production and the electricity consumption of the above-mentioned buildings;
- a model to assess the profitability potential of different energy management strategies, in particular of DR strategies.

For the sake of consistency, the complete analysis focuses on the civil pavilion, since the two towers are not going to be equipped with PV panels and therefore there is no possible assessment of the PV production. As far as the central pavilion is concerned, its complete analysis could be carried out in a similar fashion to the civil pavilion and therefore only the results of the civil building are presented in detail in this thesis.

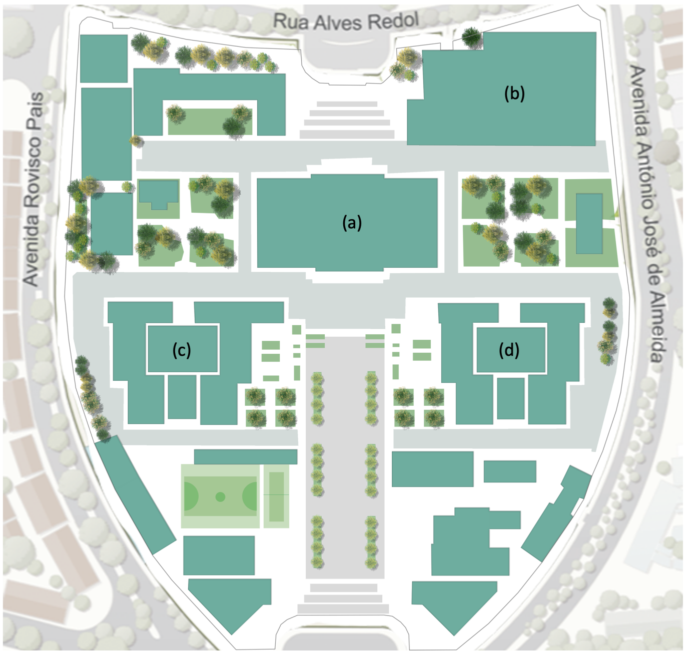


Figure 3.1: Map of the Alameda campus highlighting the analysed buildings: (a) central building, (b) civil building, (c) south and (d) north towers.

However, to further assess the feasibility of the consumption forecasting model proposed, the machine learning models related to the forecasts of the consumption have been tested on all four buildings and the results of the different models are summarised in section 4.1.7.

3.1.1 The Civil Pavillion

The civil pavillion is one of the main buildings on the Alameda campus, with a total area of about 25152 m^2 and is composed by seven floors, three below ground and four upper floors (Figure 3.2 shows the blueprint of the ground floor of the building). The upper floors are composed by two blocks, an eastern and a western block, separated by an inner patio. The building has a central backbone covered by a glass ceiling that allows natural lighting while the access to the upper floors is granted by three towers (north, central and south) inside the building, which extend themselves up to the third floor. The building hosts classrooms, teachers' and researchers' offices, laboratories, a library and an auditorium, and as such, it operates almost continuously throughout the year, with small exceptions during weekends and national holidays and during the month of August. The entire building is open during the week from 7am to 9pm and from 7am to 5pm on Saturdays while the area with some studying rooms is open 24/7. As a consequence, the peaks of activities correspond to the periods of classes, which are divided into two semesters and mostly during weekdays. It is indeed possible to notice a lower occupancy and lower electricity consumption both during the weekends and in the month of August, when the rate of the activities decreases.

The form of energy used in this building is mostly electric energy, and partially natural gas, especially in the rented spaces, such as the cafeteria on the ground floor of the building. The electric energy supplied to this building is used for multiple purposes, such as lightning, the HVAC system, electric devices and in the laboratories.

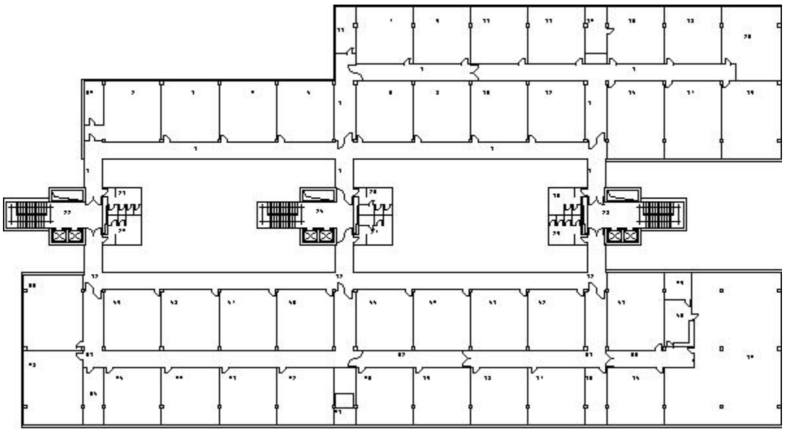


Figure 3.2: Blueprint of the ground floor of the civil building.

3.1.2 Available Data and Model Overview

The aforementioned models and their development process are introduced in greater detail in this sub-chapter and form the different steps to be able to assess the profitability potential of the suggested energy management strategy. Figure 3.3 shows a model overview and the main calculations performed in each of them. All models are developed, validated and tested using a dataset of different parameters which ranges from 01/01/2017 to 31/12/2018 for a total of 730 days. To this purpose, three data sources are available: electricity consumption data, weather data and occupancy data. A more detailed explanation of the input data is given below and in their respective sections.

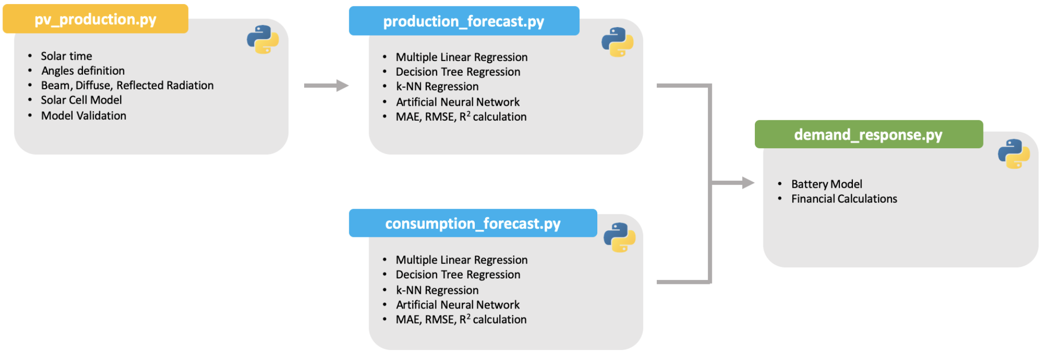

pv_production.py
• Solar time
• Angles definition
• Beam, Diffuse, Reflected Radiation
• Solar Cell Model
• Model Validation
production_forecast.py
• Multiple Linear Regression
• Decision Tree Regression
• k-NN Regression
• Artificial Neural Network
• MAE, RMSE, R² calculation
consumption_forecast.py
• Multiple Linear Regression
• Decision Tree Regression
• k-NN Regression
• Artificial Neural Network
• MAE, RMSE, R² calculation
demand_response.py
• Battery Model
• Financial Calculations

Figure 3.3: Overview of the developed model.

Weather Data The weather data used was collected at the Instituto Superior Técnico Meteo Station, situated on top of the South Tower of the campus (38.736°N, 9.138°W, 90 m a.s.l. [1]) and is available with a 5 minute resolution.

[1] above sea level

Occupancy Data Building occupancy is of great importance for the implementation of energy efficient measures in buildings as it is strictly related to the energy consumption [91]. The analysed building is not equipped with presence sensors and therefore it is difficult to know the exact number of people in each area. Many works have investigated and analysed the performance of indirect indicators of people presence and, among others, WiFi connected users have proved to be a good estimate [92], even though such an indicator is obviously characterized by the uncertainty due to that fact that not all people in a building might be connected to the WiFi network with a device, or, on the other hand, they could be connected with multiple devices. For the scope of this work, the WiFi connected users are considered a valid indicator of the number of people in the building.

All information about the WiFi infrastructure network of the analysed building is stored using RRDTool (https://oss.oetiker.ch/rrdtool/index.en.html), an open-source industry standard data logging and graphing system for time series data, which acquires the data at regular time intervals through Simple Network Management Protocol (SNMP). RRDTool uses data consolidation features to store the data which is available to download through the Cacti software (https://www.cacti.net). The time series of the logged number of devices are split between all 53 available access points (APs) of the building. For the purpose of this work, the sum of the connected users of all the 53 APs has been taken into account as indicator of people presence in the building.

Electricity Consumption Data Energy consumption data was collected with hourly resolution from smart meters installed in the building provided by the IST project *Campus Sustentável* (http://sustentavel.unidades.tecnico.ulisboa.pt). These data are acquired periodically and correspond to the average current in Ah of the last hour, which was then converted into kWh using an average voltage of 230 V and a power factor of 0.90 for conversion.

3.2 Electricity Production Modelling

The first model developed aims to calculate the power output from the PV panels installed on the roof of the building, as no historical data longer than a week were available to develop the forecasting model. Whether the aim is to develop a univariate model, that only uses previous production data, or a multivariate model, that combines the use of past power data and meteorological data, such as solar irradiance, temperature, humidity and wind speed as presented by Rana [50], the need for historical data of PV power is plain. The following sections present the calculation and analysis process of the potential power production and gives a better understanding of the necessary inputs and outputs of the model.

3.2.1 Model Development Process and Input Data

The first step in setting up and running a model to assess PV power production is obtaining irradiance and, more in general, weather data. Such data is typically used to predict the output power of the proposed system, while power data, if available, is used to validate that the model, or, if in a real-time application, the PV system is functioning properly. The solar power generated from PV systems depends on solar irradiance and other weather variables such as temperature, humidity, wind speed and cloud cover. In particular, the model used as input the total solar irradiance and the ambient temperature from 01/01/2017 to 31/12/2018 with a 5 minute resolution. The development of such a model include three main tasks:

1. the implementation of the Perez model to to transform the total horizontal irradiance into the effective irradiance on the surface of the PV panel throughout the day;

2. the usage of a solar cells model (the one diode and three parameters model) to find the current and the voltage at the maximum power point;

3. the calculation of the AC power using the peak inverter efficiency.

3.2.2 Solar Radiation Model

Among the different existing models to compute the solar radiation presented in 2.2.1 and, in particular, the total irradiance on a tilted surface, the Perez model is the one offering the most accurate representation of the diffuse component of the solar radiation with respect to its three components, namely the isotropic diffuse, the horizon brightening and the circumsolar radiation [20]. In the model to assess the solar power output calculation, it was decided to include a model to transform the total horizontal irradiance into the effective tilted irradiance, so that the overall model to calculate the solar power output could be easily used both in the case where the panels are installed flat on a roof and in the case of tilted panels. After the preliminary calculations of solar time, astronomical parameters and solar angles, the total irradiance on the tilted surface has been calculated according to the Perez model (see section 2.2.1.1):

$$\frac{I_T}{I} = \left(1-\frac{I_d}{I}\right)R_b + \frac{I_d}{I}F_1\cdot\frac{a}{b} + \frac{I_d}{I}(1-F_1)\left(\frac{1+cos\beta}{2}\right) + \frac{I_d}{I}F_2 sin\beta + \rho_g\left(\frac{1-cos\beta}{2}\right) \qquad (3.1)$$

3.2.3 Solar Cell Model

There are innumerous commercial PV modules on the market today, each possessing its own strengths and weaknesses. The maximum power point and the AC power produced have been calculated thanks to the data sheets provided, the ones of the photovoltaic panel Amerisolar AS-6P30 (250W) and the inverter Involvar MAC250A. This particular panel and inverter were chosen because they are already installed in the solar energy laboratory and for which a week of historical data was available to validate the model. The key performance metrics for this type of module and inverter can be seen in the data sheet in Appendix A and B.

3.2.4 Model Validation

The developed model was validated using historical data from the solar energy laboratory (a photovoltaic panel) installed at the Department of Physics on the South Tower on the main campus of Instituto Superior Técnico. Given specific input parameters, such as the solar radiation, the temperature and the data sheets of the panel and the inverter, it was possible to calculate the parameters and expected outputs of the installation and compare them with the experimental measurements of power. This way, it was possible to conclude whether the developed model, which makes use of the Perez model to calculate the total radiation on a tilted surface, accurately models the energy produced. The model validation used previous power data from a week in December 2017 and showed an average error in the power output between 5% and 10%. Figure 3.4 shows an example of the daily calculated solar power produced versus the power measured for a 250W panel, where the deviation between the two graphs could be due to an underestimation of the cell temperature.

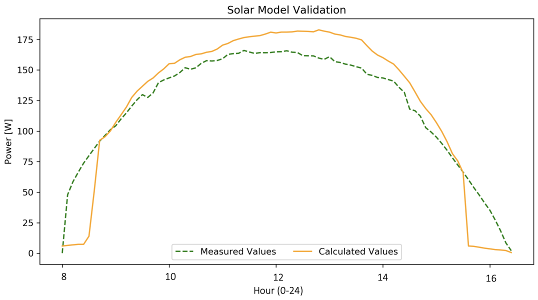


Figure 3.4: Measured power output versus calculated power output for one 250W panel.

3.3 Machine Learning Models

One of the most commonly used Python packages to implement machine learning is Scikit-Learn and it was therefore chosen to carry out the development of this work. The Scikit-Learn package offers a wide range of machine learning algorithms, both for supervised and unsupervised learning, as well as testing and validation features and it is characterized by a clean uniform interface. Once the basic use and syntax are clear, it allows to switch between algorithms or models in a reasonable amount of time [74].

Choosing the most suitable ML tool implies a trade-off between advantages and disadvantages of each one of the models. Therefore, among all supervised regression learning tools, four different methods (multiple linear regression, decision trees, kNN regressors, and artificial neural networks) were tested for the two tasks of predicting electricity consumption and production and their specific parameters optimized being these among the most common ones found in literature which combined a relatively low complexity and a good prediction accuracy (see sections 2.3.2 and 2.3.3).

3.3.1 Model Development Process

The methodology used to develop data-driven prediction models, whether to forecast electricity consumption or the PV panels power output followed the same main steps. In data-driven model development, the development process can be divided in the following main steps:

- Collecting, preparing and pre-processing the data (section 3.3.2);
- Choosing an evaluation metric and a testing procedure (section 3.3.3);
- Identifying the important features (section 4.1.1 and 4.1.6);
- Developing the model and tuning its hyper-parameters (section 3.3.4);
- Evaluating the model performance and proceed to a further optimization if needed.

This process was applied iteratively for all models, both varying the input data, the percentage of train-test data and the hyper-parameters fed to the different models to sense the effect of such parameters on the quality of the predictions. To this purpose, different script were developed in Python for each of the model tested.

3.3.2 Data Collection and Preprocessing

For all machine learning tools, it is good habit to pre-process the input data to see if there is any significant trend, periodicity and irregularity. In general the databases used were relatively clean and, as a result, the main task in data preprocessing consisted of scaling (up or down) all features in order for them to have the same timestamp to be merged in the same database. The resulting database was therefore containing 70080 rows for two years of data, equivalent to 96 points a day.

A few missing points have been replaced interpolating the previous and the following point in the dataframe. Considering the data has a 15 minute resolution, an interpolated value is considered a good estimation. At this point, outliers have been identified, as for example, a few positive values of solar power output at night, and removed. Another preprocessing procedure included denormalising categoric variables into a set of boolean variables, as for examples holiday flags in the electricity consumption dataframe, paying attention to continuous input variables, such as wind direction, which are difficult to handle by classifiers since there is continuity between 0°and 360°, or normalization of input variables.

After the preprocessing phase, the dataframes were sorted correctly according to a regular timestamp of 15 min. It should therefore be clarified that all values refer to the energy consumption or production in kWh per 15 minutes.

3.3.3 Performance Measures

To be able to compare different learning methods and evaluate their overall performance, it is useful to identify different performance measures against which the difference models can be compared. In particular, to measure regression performance during model testing of both electricity production and consumption the chosen metrics include the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE) and the coefficient of determination (R^2).

- **Mean Absolute Error (MAE):** The mean absolute error is the average of the difference between a real and its correspondent predicted value. It measures how far the predicted values are from the actual output but it disregards the direction of the error, i.e. whether the value is over or under the real value. Mathematically, it is represented as :

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\widehat{y_i} - y_i| \tag{3.2}$$

- **Root Mean Square Error (RMSE):** Specifically, the RMSE measures the square root of the average value of the squares of the errors, formulated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\widehat{y_i} - y_i)^2} \tag{3.3}$$

- **Coefficient of determination ($\mathbf{R^2}$):** The coefficient of determination accounts for the percentage of points that fall on the line formed by a regression equation. It varies from 0 to 1, being 1 a perfect representation of the modelled data. The higher the coefficient, the more points the regression line passes through and the higher the ability of the model to predict future events. The coefficient of determination can also be thought as a probability of a new point to fall on the regression line. More specifically, it indicates the proportion of the variance in the dependent variable that is predicted or explained by the predictive variables.

$$R^2 = 1 - \frac{\sum_n^i (y_i - \widehat{y_i})^2}{\sum_n^i (y_i - \widehat{y_i})^2} \tag{3.4}$$

3.3.4 Development and Testing Process

A time series forecast can be made for different steps in the future, such as one hour ahead or one day day ahead. The lengths of the steps in the future is the forecast horizon whereas the starting point from which it is made, is called the forecast origin. It can be easily sensed that longer forecast horizons are usually less accurate because of the variability of the features influencing the output and the aggregation of errors. The basic testing method for machine learning models consists of splitting the data into two sets, a training set and a testing set, training the model using the training set, make prediction using the testing set and assess the performance of the model with an appropriate metric. There are two obvious drawbacks to this methodology: the first is that the training set might be too small to be split into two subsets, and the second one is that depending on where the split is performed the model performance might change. To reduce this issue common practice is to perform cross-validation (CV) procedures, such as k-fold cross validation. K-fold cross-validation refers to randomly dividing the set into k folds, iteratively using k-1 folds for training and the k fold for testing. This process is repeated until all the folds are used and the value of the chosen performance metric is calculated as the average of the errors of each of the folds. It is clear that such approach is not appropriate for time series datasets as observations are normally strictly dependent on previously occurred observations and intrinsically carry with them a time attribute (the order in the dataset).

The chosen approach is therefore to extract feature from the timestamp of the dataset such as the day, the month, the day of the week, the hour and the minute. In this way the timestamp features can be passed to the algorithm as numerical values and the algorithm can learn the relationship between the timestamp and the output rather than learning the relationship between historical outputs and the current one.

To achieve a more rigorous procedure for validation, different train-test splits have been tested, incrementally increasing the percentage of the training set and always choosing a test set from the end of the data set, i.e. performing last block validation [93].

The method described above was the general development and testing method for the algorithms. Another factor influencing the implementation process of the different methods are the features. To test the most important features for the different algorithms a forward stepwise method was chosen. This procedure included starting the model with very few predictors and iteratively assessing its performance while increasing the number of predictors to see if the accuracy increases or not. The procedure is repeated until no further improvements can be achieved and only the four best combination of predictors are retained. At this point the hyperparameters of each of the four models need to be further optimized, such as the number of neighbours in kNNs, the number of nodes and depth of decision tree and the overall architecture of ANNs. The tuning of such hyperparameters directly affects model performance and it requires hard work to be found. Empirical approaches have to be used as no universal procedure exists [60] and care needs to be taken not to overfit or underfit the model (see section 3.3.4.2) while tuning the hyper-parameters.

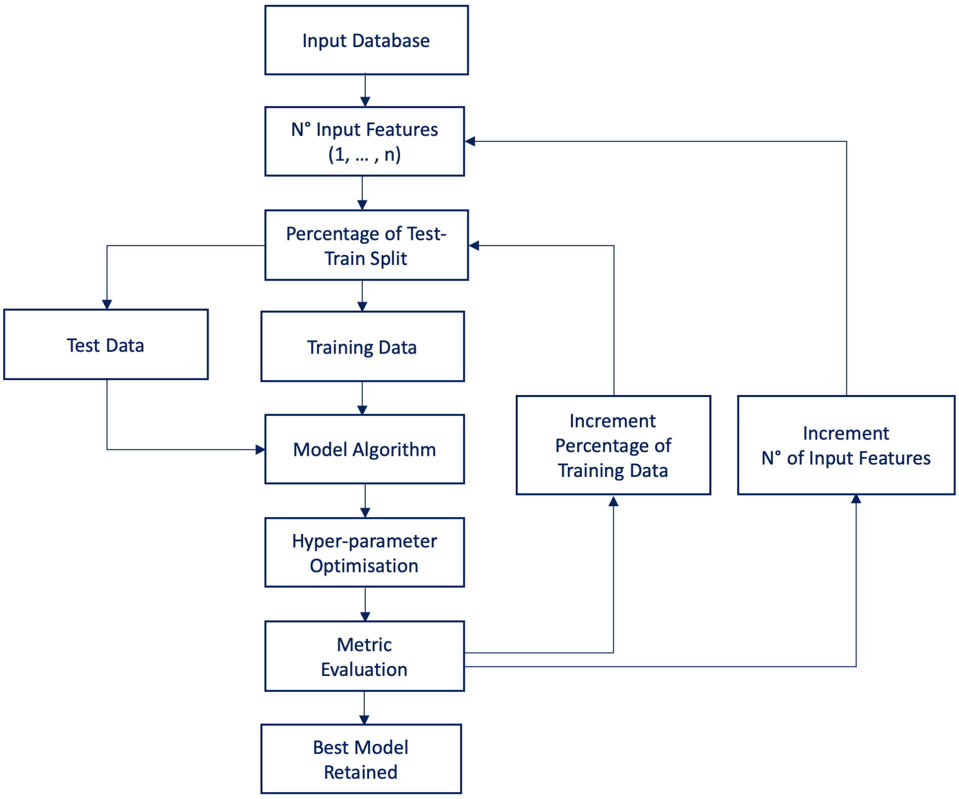

Input Database
N° Input Features (1, … , n)
Percentage of Test-Train Split
Test Data
Training Data
Model Algorithm
Increment Percentage of Training Data
Increment N° of Input Features
Hyper-parameter Optimisation
Metric Evaluation
Best Model Retained

Figure 3.5: Model development and evaluation procedure. At the end of each simulation, the MAE, RMSE and coefficient of determination are computed.

3.3.4.1 Hyper-parameters

- **K-Nearest Neighbours (kNN):** The scikit-learn package for supervised neighbours-based learning offers different methods and parameters that can be tuned. The overall principle behind this method is to find the best number of training samples closest in distance to the input and make a prediction from these samples. The number of these samples needs to be set as a constant (k-nearest neighbour learning), or can vary according to the local density of points (radius-based neighbour learning). The distance parameter can be any metric: standard Euclidean distance and Minkowski distance are the two common choices. Considering that kNN is sensitive to the domain of its input values, it is important to normalise to the same scale the values beforehand to make different distances comparable. Additionally two different weighting functions were tested, one that provides a uniform weighting to all of the nearest neighbours, and one that weights the contribution according to the distance from the input point. The tested parameters for the method included a varying number of neighbours from 5 to 60, two distance metrics (Euclidean and Minkowski) and two different weighting of the features (uniform and distance).

  Considering that increasing the training test size, the number of features to consider and the number of neighbours or the radius, increases the computational time and that kNN is not very good in the presence of uninformative variables, always less than 6 input features were tested every time.

- **Decision Tree:** It was decided to test decision trees since they are very adequate in dealing with categoric and boolean features, such as holiday flags. The parameters to be tuned include the size of the tree, its maximum depth and its minimum number of leaves. Not optimising these parameters proved to be very computationally demanding and lead to fully grown unpruned trees. It is possible to see that if the maximum depth of the tree is set too high, the decision tree learns too fine details from the training data resulting in an overfitting. At each branching split the algorithm performs internal optimization to decide which attributes to use using different criterion metrics. The tested criteria included a maximum depth ranging from 0 to 20 and two criteria, 'mse' for the mean squared error and 'mae' for the mean absolute error.

- **Artificial Neural Network:** The hyper-parameters that can be controlled for the artificial neural network are the number of hidden layer, the number of neurons per hidden layer, the activation function and the solver. The learning rates tested included adaptive, constant and inverse scaling whereas the activation function included the logistic sigmoid function, the hyperbolic tangent function and the rectified linear unit function. The solver used tries to optimise the squared-loss using stochastic gradient descent.

3.3.4.2 Bias versus Variance Tradeoff

The bias-variance tradeoff is the problem associated with finding the perfect balance between two sources of errors in machine learning models: errors due to variance and error due to bias. Bias refers to the error between the expected prediction of the model and its real values in the training data. This results in underfitting the model, i.e. the model is too simple to learn the underlying structure of data and predictions are therefore bound to be inaccurate. Underfitting and, therefore a high bias, can occur when fitting a simple model to complicated data. The opposite situation is called overfitting and occurs when fitting a complex model to simple data. Overfitting leads to a low bias because the model results to be flexible, but, at the same time, the variance is high because the model is too influenced by the training data. As a result, the model will perform well on the training data, but it is not capable of generalizing well to new unseen instances.

To avoid overfitting, good approaches are trying to simplify the model by choosing a model with fewer parameters, selecting fewer input features in the training data or reducing the noise in the training data so that even a complex algorithm does not learn the noise in the training data. Tuning the hyperparameters of the different methods is an important part of building a machine learning system. They are normally set before learning and an experimental iterative process needs to be carried out to find their optimum values as they greatly influence the bias-variance trade-off.

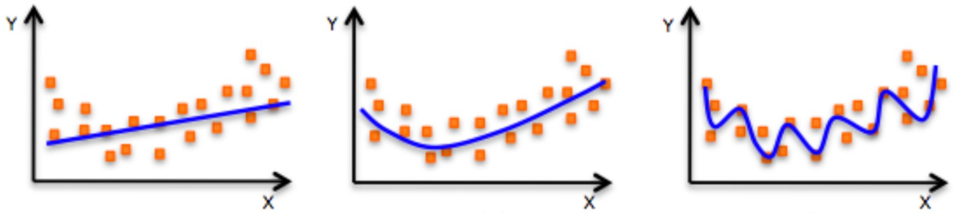


Figure 3.6: Graphical illustration of underfitting (left), overfitting (right) and an adequate fitting (center) of the model on the data [74].

3.4 Demand-Response Model

The main goal of this section is the assessment of the economic feasibility of the implementation of a battery energy storage system to perform demand-response. The first task consists of optimally sizing a behind-the-meter battery to minimise electricity costs exploiting cost arbitrage of time-of-use electricity tariffs. Subsequently, the solution is analysed from the cost perspective point of view to ensure the investment can at least be paid back within a fixed hypothetical lifetime of the system considering current market-based battery prices.

To optimise the schedule of the battery and choose a reasonable battery size, linear programming methods are employed. On the line of the previous models developed, it was chosen to keep employing open-source programming languages such as Python. Python scripts are indeed easy to build-on and can easily be translated in other languages to be integrated in other programs [94]. In particular, the open-source software package Pyomo was employed, which possesses broad optimization capabilities to formulate, solve, and analyse optimization models [95]. Pyomo allows fairly easy formulations to define the objective function of the model, the constraints, the decision variables and the parameters.

The resulting problem is a Mixed-Integer-Linear-Problem (MILP) whose solution can be found using the solver GLPK (GNU Linear Programming Kit). For MILPs, GLPK employs as default the branch-and-bound algorithm together with Gomory's mixed integer cuts, and is able to solve the problem in a matter of minutes.

This method was chosen in place of more complex ones, because it accommodates rapid cost-optimal battery sizing and has the advantage of being easily implemented in devices with low computational power.

3.4.1 Model Development Process

The formulated problem aims to cost-optimise the overall electricity bill for the building, taking into account real market electricity tariffs while considering technical battery constraints and real consumption and production data.

The considered load profile consisted of 12 months of data for the year 2018 for the analysed building, taking also into account the month of August, when the consumption and the activities in the building are negligible compared to the rest of the year. As in the analysed period, the PV production was always lower than the consumption of the building, the best strategy for the solar power production is direct consumption. As a result, the input load to the model for the analysed months is the net load resulting from the subtraction of the power production from the given load collected by the smart-meters. Similarly to the previous models and analyses, the input data used consisted of a 15 minute dataset.

The economic opportunity that the problem aims to model is the shift of the net consumption from peak hours, when the electricity price is higher, to off-peak hours, when the associated cost is much lower. The battery can indeed charge during off-peak hours and discharge during peak hours, resulting this way in a lower overall energy cost. In order to do so, it is of the utmost importance to understand the applied tariff structure applied by the utility company, in this case Energias de Portugal (EDP). The applied tariff is a typical Medium Voltage (MV) time-varying tariff that is divided into seasonal and daily periods. This tariff has been used in the simulations as its values are very close to the real ones but it is not the one actually applied to the Alameda campus. The overall tariff includes two major charges: an energy charge referred to the amount of kWh consumed and a power demand charge that is correlated with the maximum peak demand over a day or over a month. The EDP MV weekly electricity purchasing price applied is reported in Table 3.1 and it includes VAT at the present level in Portugal (23%) [96]. As it can be seen from Table 3.1, the tariff is divided in four trimesters per year, and it depends on the day of the week as well as on the time during the day, for a total of eight different prices per year. The solar power production already contributes to decreasing the electricity tariff as the production peaks always occur in the peak hours within the tariff scheme. The bigger share of the electricity consumption is though still concentrated in the peak hours of the weekdays.

Based on the active energy tariff, first the overall electricity costs are calculated and, once the battery is sized, the possible achievable savings in the presence of an optimally scheduled battery are calculated.

Table 3.1: EDP medium voltage quadri-hourly tariff schedule.

Period	Time of Use	Monday - Friday	Saturday	Sunday	Tariff [€/kWh]
Periods I, IV	Peak	9:30 - 12:00 18:30 - 21:00	/	/	0.1382
	Half-Peak	7:00 - 9:30 12:00 - 18:30 21:00 - 00:00	9:30 - 13:00 18:30 - 22:00	/	0.1111
	Normal Off-Peak	00:00 - 2:00 6:00 - 7:00	00:00 - 2:00 6:00 - 9:30 13:00 - 18:30 22:00 - 00:00	00:00 - 2:00 6:00 - 00:00	0.0777
	Super Off-Peak	2:00 - 6:00	2:00 - 6:00	2:00 - 6:00	0.0666
Periods II, III	Peak	9:30 - 12:30	/	/	0.1408
	Half-Peak	7:00 - 9:30 12:30 - 00:00	9:00 - 14:00 20:00 - 22:00	/	0.1124
	Normal Off-Peak	00:00 - 2:00 6:00 - 7:00	00:00 - 2:00 6:00 - 9:00 14:00 - 20:00 22:00 - 00:00	00:00 - 2:00 6:00 - 00:00	0.0791
	Super Off-Peak	2:00 - 6:00	2:00 - 6:00	2:00 - 6:00	0.0728

3.4.2 Optimization

To find the optimum battery size an optimization procedure is developed. The model takes as input a vector for the load and a vector of the respective electricity costs, with a 15 minute timestep and tries to find the optimum schedule, i.e. charged and discharged energy to and from the battery in order to minimise the costs of the electricity bill, while respecting the technical specifications for the battery. The starting point is the definition of the objection function (section 3.4.2.1), of the variables and their boundaries in an algebraical way so that they can be interpreted by Pyomo.

Pyomo allows to define two different types of models, a so called concrete model, where parameters are defined at the same time as the model definition and an abstract model, in which the parameters are specified at a later stage. A concrete model is used as all parameters are specified during model development and before the simulations are run. Initially multiple simulations are run, varying the charging and discharging power and the capacity of the battery to find the best combination for the given problem and in a second moment, once the specifications of the battery are known, a comparison using the forecast load and solar power is carried out.

3.4.2.1 Objective Function

The main goal of the optimisation is to reduce the total electricity bill by increasing or decreasing the electricity bought from the grid at different times. Considering the site has a PV installation, the model could possibly be more complex and include a selling price for the electricity sold to the grid in the case of negative consumption but, as this situation does not occur in the analysed months, the decision variable reduces to the positive net load consumed at each timestep (a negative load would correspond to an excess in production which could be sold to the grid). The net load to be bought from the grid is a function of the load of the building and of the battery action and can be written as:

$$Load_{net}(t) = Load_{building}(t) + Load_{battery}(t) \tag{3.5}$$

The objective function becomes then:

$$objectiveFunction = min\left(\sum_{t=0}^{n}(Load_{net}(t) \cdot Price_{electricity}(t))\right) \tag{3.6}$$

At this point, constraints and boundaries have to be defined so that the model represents the battery parameters. In particular, the constraints aim to guarantee the following characteristics:

- a minimum and a maximum state of charge (SoC) of the battery, SOC_{max} and SOC_{min}, have to be respected;
- a charging and a discharging power limits have to be set;
- a charging and discharging efficiency, η_{charge} and $\eta_{discharge}$, have to be taken into account (battery charging and discharging efficiencies of 90% were considered.).

On its turn, the SoC is defined splitting it into a positive and a negative component, namely $\Delta_{SoC,pos}$ and $\Delta_{SoC,neg}$, which represent the energy coming from the grid to the battery and energy exiting the battery ($SoC(t) = \Delta_{SoC,pos}(t) + \Delta_{SoC,neg}(t)$). When charging or discharging the battery, an efficiency needs to be taken into account, and $\Delta_{SoC,pos}$ and $\Delta_{SoC,neg}$ are therefore defined as:

$$E_{grid}(t) = \Delta_{SoC,pos}(t)/\eta_{charge} \tag{3.7}$$

$$E_{outBattery}(t) = \Delta_{SoC,neg}(t) \cdot \eta_{discharge} \tag{3.8}$$

3.4.2.2 Equality Constraints

The equality constraints aim to ensure that the balance of energy flows within the system is respected. The term referring to the battery in equation 3.5, can indeed be split in two other terms: the energy coming from the grid and the energy going out from the battery:

$$Load_{net}(t) = Load_{building}(t) + E_{grid}(t) + E_{outBattery}(t) \tag{3.9}$$

The mixed integer formulation of the problem aims on the other hand, to ensure that the battery can only either charge or discharge at a particular timestep. To this purpose, boolean variables for charging and discharging are set, $bool_{charge}(t)$ and $bool_{discharge}(t)$, and the following constraint is added:

$$bool_{charge}(t) + bool_{discharge}(t) = 1 \tag{3.10}$$

3.4.2.3 Inequality Constraints

For a BESS ,the charged and discharged energy is limited by the capacity of the battery and by the charging and discharging power limits. This consists of two governing rules: one rule representing the minimum and maximum states of charge and another rule setting how much the state of charge changes at each period. The maximum and minimum states of charge can be set in the bounds of the variable SoC as they are simple constants to be assigned when initializing it:

$$0 \leq SoC(t) \leq C_{battery} \tag{3.11}$$

For the charging and discharging power limits, there are also two constraints to be set and therefore the following equations were added:

$$E_{grid}(t) \leq P_{limit_{charge}} \cdot t \tag{3.12}$$

$$E_{outBattery}(t) \geq P_{limit_{discharge}} \cdot t \tag{3.13}$$

A last constraint to ensure physical sense within the model guarantees that the energy used locally coming from the battery cannot exceed the actual demand and can be written as follows:

$$E_{outBattery} \leq Load_{net}(t) \tag{3.14}$$

In addition to the previous constraints, upper and lower boundaries are attributed to all the decision variables. In particular, the SoC is limited through upper and lower bounds and charging and discharging rates are constrained by the technical limits of the batteries through time.

Table 3.2 summarises all parameters used in the optimisation.

Table 3.2: Battery optimisation parameters.

Symbol	Parameter	Unit
$Load_{net}(t)$	Net load	kWh
$Load_{battery}(t)$	Battery load	kWh
$Load_{building}(t)$	Building load	kWh
$SoC(t)$	Battery State of Charge	-
SoC_{max}	Battery maximum State of Charge	-
SoC_{min}	Battery minimum State of Charge	-
η_{charge}	Battery charging efficiency	%
$\eta_{discharge}$	Battery discharging efficiency	%
$E_{grid}(t)$	Energy coming from the grid	kWh
$E_{outBattery}(t)$	Energy exiting the battery	kWh
$P_{limit_{discharge}}$	Discharging power limit	kW
$P_{limit_{charge}}$	Charging power limit	kW
$bool_{charge}(t)$, $bool_{charge}(t)$	Boolean variables for charging and discharging	-
$\Delta_{SoC,neg}(t)$	Negative change in battery SoC	-
$\Delta_{SoC,pos}(t)$	Positive change in battery SoC	-

Chapter 4

Results and Discussion

The main goal of this thesis is to evaluate the feasibility of machine learning tools to predict the electricity consumption and production of a higher education building, to be able to investigate energy efficient strategies that could possibly be implemented, such as demand-response using a battery energy storage system. Pursuant to this main goal, a model to calculate the PV power output is developed and data-driven supervised learning models are built both to estimate consumption as well as to predict the electricity generation based on different sets of readily available parameters, such as day of the week, hour, occupancy and weather variables. All machine learning tools' specific settings, from the input variables to the specific hyper-parameters, such as the number of neighbours to consider in kNN regression to the number of neurons for a neural network, have to be optimized. This chapter presents and analysis of all different models tested with their respective parameters.

4.1 Electricity Consumption Model

The available data to develop the electricity forecasting model was already presented in section 3.1.2. Figure 4.1 shows the data extraction and integration process to create the necessary database for the development of energy consumption prediction models with 15-min sampling rate and recalls the available input features. Additionally to the available variables, a variable representing the academic calendar and the Portuguese national holiday was added (named holiday), as it is considered important to distinguish days with a lower activity in the building from regular working days. The variable taking into account the day of the week is not able indeed to carefully represent holidays or lower activity periods, such as the month of August.

4.1.1 Correlation between Input Variables

Before going into the actual development of the models and their results, it is useful to carry out an analysis, often called exploratory data analysis, to examine the available input variables and their correlation with the electricity consumption. Such correlations will indeed provide an insights to the relationships between the features and the target value to be forecast and might help explaining the output's sensitivity on the inputs in the prediction model. In machine learning, exploratory data analysis provides insights to the importance of the predictive variables based on the change in the model performance that occurs including or excluding predefined predictors from the inputs.

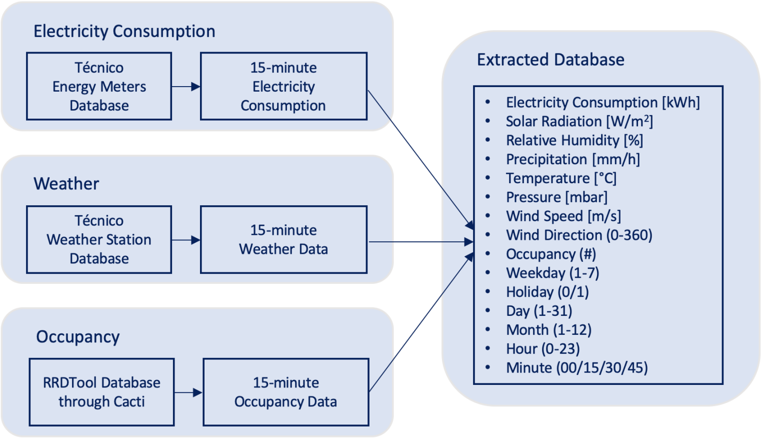


Figure 4.1: Data extraction and integration process to create the database for the development of energy consumption and production prediction models with 15-min sampling rate.

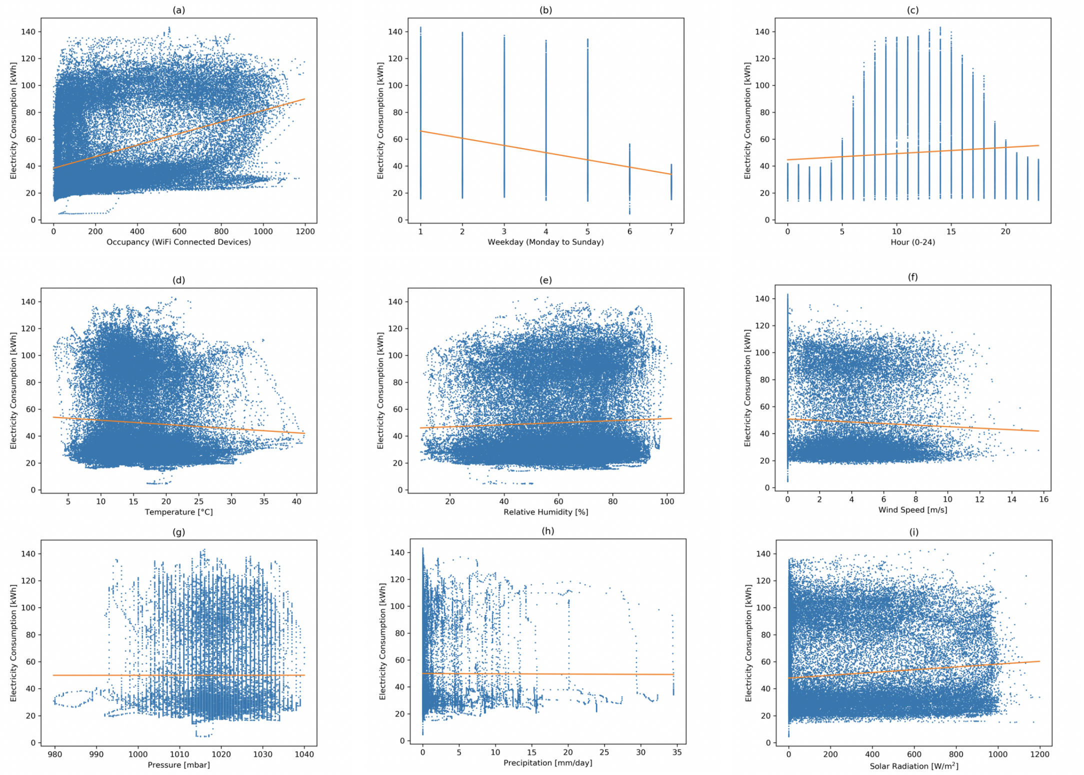


Figure 4.2: Scattered plots showing the correlation between energy consumption and (a) occupancy, (b) day of the week, (c) hour of the day, (d) temperature, (e) relative humidity, (f) wind speed, (g) pressure, (h) precipitation and (i) solar radiation.

Table 4.1: Correlation coefficients between variables.

Correlation (%)	Variable 1	Variable 2
38.5%	Electricity Consumption [kWh]	Occupancy [#]
36.0%	Electricity Consumption [kWh]	Day of the week (Mon-Sun)
44.2 %	Electricity Consumption [kWh]	Holiday (T/F)
10.7%	Electricity Consumption [kWh]	Hour of the day (0-24)
10.8%	Electricity Consumption [kWh]	Solar Radiation [W/m^2]

Figure 4.2 shows scattered plots of the correlation between energy consumption and the possible input features of the model, such as occupancy, day of the week, hour of the day, temperature, relative humidity, wind speed, pressure, precipitation and solar radiation. From the first scatter plot (a) no evident correlation arises between consumption and occupancy, but at a closer look, plotting the daily change in occupancy and consumption, it is possible to notice that these two follow the same trend and increase and decrease in a comparable proportion during the day. From plot b, it is evident the building has a weekly consumption pattern, with a higher consumption along the weekdays (Monday to Friday) and considerably lower consumption over the weekend. The electricity consumption is also considerably lower during night and it follows a characteristic bell-shaped trend during the day, when most activities take place. The lower six scatter plots (d, e, f, g, h, i) show the relationship between weather variables and consumption and no significant trend is to be noticed, which suggests that a linear correlation between such variables and the electricity consumption is difficult to find. Similarly to the occupancy, when looking more in detail at the daily change in solar radiation and consumption a direct proportional relationship between these two variables becomes evident in their daily pattern.

To investigate further the correlation between the possible input features and the target value, the linear correlation is calculated for each combination of variables both with the consumption and with themselves. The best possible features to describe the electricity consumption patterns should indeed be correlated to the consumption but, at the same time, should be independent from each other. Table 4.1 presents a table with the most significant coefficients of correlation between the input variables and the consumption. The table reports only correlation coefficients higher than 10%.

It can therefore be inferred that the variables depicted in Table 4.1, such as occupancy, the day of the week, the hour of the day and whether it is a holiday or not, could possibly be the ones that best describe significant relations with the consumption. The highest correlation is between the energy consumption and whether it is a holiday (44.2%), followed by occupancy (38.5%) and the day of the week (36%). Another input feature that has been tested in combination with the most representative features is an auto-regressive feature representing the electricity consumption of the past 15 minutes and it is labelled as 'Energy-1' throughout the results.

The next subsections (4.1.2, 4.1.3, 4.1.4, 4.1.5) present the results of the application of the four models, multiple linear regression, decision tree, kNN and neural networks, to assess their feasibility in representing how electricity is consumed in the civil building. The testing and tuning procedure of the parameters followed the model development process explained in section 3.3.4 and the performance of the different models has been assessed according to the metrics presented in section 3.3.3.

4.1.2 Multiple Linear Regression

Different combinations of input variables and percentages of training and testing data have been tested for the linear regression model. The resulting most performing model according to different input parameters are presented in Table 4.2.

Table 4.2: MAE, RMSE and coefficient of determination of the tested consumption multiple linear regression models for different input features.

Simulation	Weekday	Day	Month	Hour	Minute	Occupancy	Holiday	Radiation	Energy-1	MAE [kWh]	RMSE [kWh]	R^2
1	•	•	•	•	•	•				22.6	29.3	0.19
2	•	•	•	•	•	•	•			19.2	25.8	0.25
3	•	•	•	•	•	•	•	•		19.0	25.6	0.26
4	•	•	•	•	•	•	•	•	•	4.6	6.7	0.95

The third best model for the linear regression was developed considering the variables relative to the day of the week, the ones representing the time of the year (day and month), the time of the day (hour and minute) and the occupancy, resulting in a MAE of 22.6 kWh, a RMSE of 29.3 kWh and a coefficient of correlation between predicted consumption and real consumption of 0.19. Adding more variables, like whether the analysed day is a holiday or not, decreases slightly the MAE to 19.2 kWh and the RMSE to 25.8 kWh, while the coefficient of correlation increases to 0.25. Adding the solar radiation increases the performance of the model even more, reducing the MAE to 19.0 kWh and the RMSE to 25.6 kWh, while enhancing a bit more R^2.

Adding further weather variables, such as wind speed, pressure, precipitation etc. does not show any further improvement of the model, and the corresponding performance indexes decrease slightly, highlighting that the most important features to describe electricity consumptions are those representing the time of the year and of the day, the occupation and the holiday. The solar radiation feature helps instead the linear regression model to better represent the electricity consumption, probably because of the similar daily trend of these two parameters.

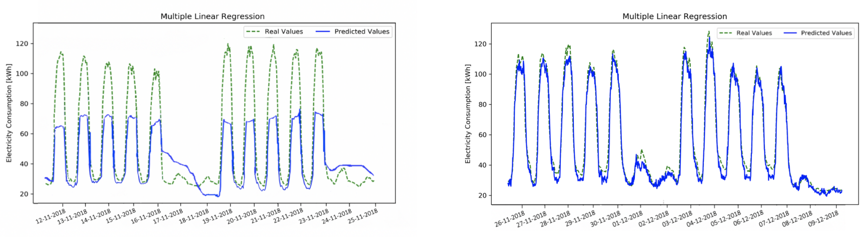


Figure 4.3: Multiple linear regression model versus real consumption (simulation 3 and 4).

Figure 4.3 shows two representative weeks in November: the dashed green line represents the real consumption measured by the smart-meters, whereas the blue line represents the predicted consumption of the best linear regression model. The left graph show the results of simulation 3, whereas the graph on the right uses the same input feature with the additional contribution of the auto-regressive feature. In the left graph, it is possible to see that the linear regression model does not capture well the load profile and its variations and its employability is therefore limited. The model does not track the daily variation of the consumption and has difficulties in recognising the lower consumption values that occur during weekends. Moreover, in the linear regression model, the training percentage of the data needed to accurately model the consumption was around 80% of the total available dataset, as with a lower percentage of training data the model was tracking the consumption even less accurately. This highlights how the needed period to train a linear regression model needs to be longer to account for possible variations in consumption, as otherwise the model would not be sensible to events like the lower consumption associated with the month of August.

4.1.3 Decision Tree

The second model tested was a decision tree regressor algorithm that extracts predictive information in the form of human-understandable rules. Its basic functioning rule is like an 'if-else' statement that explains the decisions that lead to a prediction. Being the decision tree algorithm based on conditional probabilities, it is expected that this algorithm performs better than the linear regression model, especially in differentiating between weekday and weekend consumption and holiday and working days.

At each conditional decision, i.e. at each branching split, the decision tree regressor performs an internal optimization to select the most important features at each step. This ensures that the decision tree is sensible to case specific values in its prediction. Table 4.3 outlines the performance of the four most performing decision tree models. Similarly to the linear regression model, the most representative model, is the one that, besides the time-related features, accounts for the holidays and occupancy with a MAE of 14.1 kWh, a RMSE of 15.3 kWh and a correlation coefficient of 0.52, metrics which all improve with the addition of the autoregressive feature.

Table 4.3: MAE, RMSE and coefficient of determination of the tested consumption decision tree regression models for different input features.

Simulation	Weekday	Day	Month	Hour	Minute	Occupancy	Holiday	Energy-1	MAE [kWh]	RMSE [kWh]	R^2
1	•	•	•	•	•				21.3	28.3	0.34
2	•	•	•	•	•	•			15.7	21.1	0.45
3	•	•	•	•	•	•	•		14.1	15.3	0.52
4	•	•	•	•	•	•	•	•	3.4	4.8	0.97

Figure 4.4 shows an example of the prediction of the decision tree regression model. In contrast to the multiple linear regression the decision tree is able to model the weekend considerably better than the linear regression, as it is able to make a clearer distinction between the days of the week. Nonetheless, the model is still not capable of capturing the daily peaks of consumption accurately, which might suggest that without the knowledge of the consumption of the past 15 minutes, the model does not include branches for higher values of the consumption.

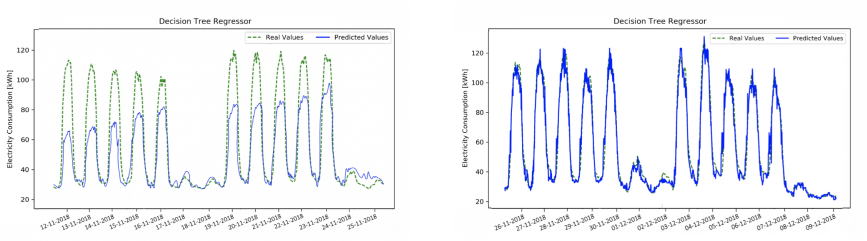


Figure 4.4: Decision tree regression model versus real consumption data (simulation 3 and 4).

4.1.4 K-Nearest Neighbours

The kNN algorithm takes a different learning approach from the previously tested algorithms as instead of using a global approach, if uses a local approach, which is usually easy to implement and interpret. Using a function to determine the similarity of points, it should perform well in estimating the consumption as it makes weaker assumptions about the data, and does not try to discover an underlying function to be able to model all data across the entire input domain. For this reason, it should perform better than the other algorithms with smaller set sizes. The algorithm was indeed reaching results very close to the ones depicted in Table 4.4 with just 50% of data used in the training set. The best results were though achieved with 80% training data and 20% testing data and are presented in Table 4.4.

As the set size increases and the algorithm needs to account for more data points, the time to fit the algorithm increases considerably as the number of examples it has to examine to determine the nearest neighbours becomes large. Similarly to the decision tree, the information related to the occupancy of the building and the holidays was fundamental to achieve a MAE of 15.0 kWh, a RMSE of 18.2 kWh, and a coefficient of determination of 0.53 which become 3.5 kWh, 4.9 kWh and 0.96 respectively when including the auto-regressive feature.

Table 4.4: MAE, RMSE and coefficient of determination of the tested consumption k-nearest neighbours regression models for different input features.

Simulation	Weekday	Day	Month	Hour	Minute	Occupancy	Holiday	Energy-1	MAE [kWh]	RMSE [kWh]	R^2
1	•	•	•	•	•				22.4	27.4	0.36
2	•	•	•	•	•	•			17.5	23.1	0.44
3	•	•	•	•	•	•	•		15.0	18.2	0.53
4	•	•	•	•	•	•	•	•	3.5	4.9	0.96

From the plot of Figure 4.5 it is possible to see that although the model is able to discern between weekdays and holidays, its forecast electricity consumption values (blue line) are always underestimated compared to the real values (dashed green line). The auto-regressive feature helps the model performance but create a lot of oscillations when forecasting the highest and the lowest values.

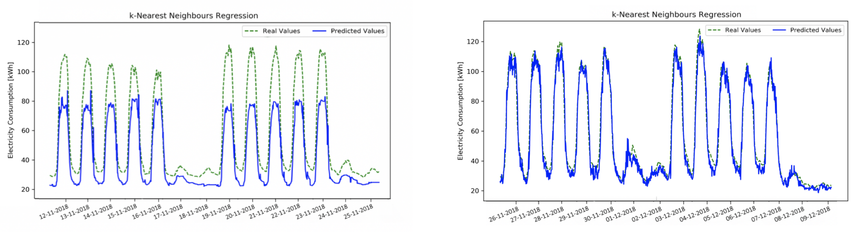


Figure 4.5: K-Nearest Neighbours regression model versus real consumption data (simulation 3 and 4).

4.1.5 Artificial Neural Network

The last tested model is a neural network model. In particular, the function used implements a multilayer perceptron (MLP) that trains using backpropagation with no activation function in the output layer. Therefore, it uses the square error as the loss function, and the output is a set of continuous values. After numerous iterations to find the best combination of layers and neurons, the most reliable model was implemented using all time-related variables, occupancy data and holiday and was composed of 7 neurons in the input layer, 7 neurons in the first hidden layer, 22 in the second hidden layer and 1 neuron in the output layer. According to Table 4.5 the model resulted in a MAE of 9.1 kWh, a RMSE of 10.4 kWh, and a R^2 of 0.63. This model followed a similar behaviour as the kNN model or the decision tree model, where the performance was decreasing when considering additional weather variables as input features, in opposition to the multiple linear regression model where the solar radiation variable was helping to achieve a better performance.

Table 4.5: MAE, RMSE and coefficient of determination of the tested consumption artificial neural network models for different input features.

Simulation	Weekday	Day	Month	Hour	Minute	Occupancy	Holiday	Energy-1	N°neurons	MAE [kWh]	RMSE [kWh]	R^2
1	●	●	●	●	●				33	13.4	21.1	0.57
2	●	●	●	●	●	●	●		33	10.2	15.6	0.61
3	●	●	●	●	●	●	●		7, 22	9.1	10.4	0.63
4	●	●	●	●	●	●	●	●	7, 22	1.4	2.9	0.97

As it can be seen in the left graph of Figure 4.6 the neural network model (simulation 3) follows better the peaks of electricity consumption during the day, but it is still not perfectly capable to track all the peaks, underestimating always the consumption. It can be though noticed from the right graph that the addition of the auto-regressive feature (simulation 4) almost completely solved this issue. Fine tuning its parameters is hard and it is difficult to find the best combination of hidden layers and neurons. As explained in the model development process (section 3.3.4), the best possible combination of model parameters were found increasing incrementally the number of neurons and hidden layers while checking the model performance metrics. When incrementing the number of layers to a third layer, the model performance kept decreasing and it was therefore chosen to stop the testing of additional hidden layers.

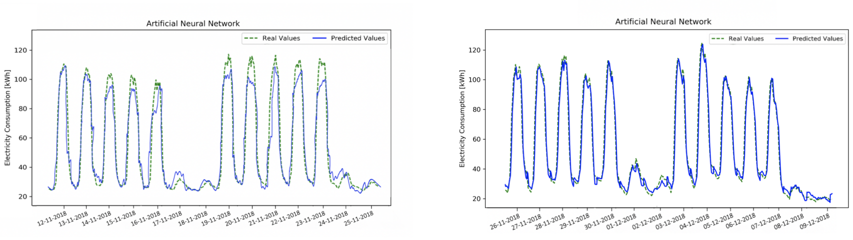


Figure 4.6: Artificial neural network model versus real consumption data (simulation 3 and 4).

4.1.6 Comparison of the Different Models

Comparing the four model developed and plotting the forecast values against actual values it can be seen that all algorithms tend to forecast values that in general are lower than the actual values, especially for the peaks of consumption during the working days and the values at night, when the consumption level is lower. A possible explanation could be that at night the consumption is not related to the activities in the building but it is caused by the base load which does not vary with time, and fitting a function over the consumption pattern may lead it to underestimate the base load at times due to the smoothness constraints of the functions. Overall all analysed tools clearly outperform the multiple linear regression, which, to be viable, requires at least 80% of the available data and still does not provide satisfactory results. Both kNN and decision tree perform better than the linear regression but are not able to capture well the daily consumption pattern. Artificial neural networks seem to be the best choice but more accurate results would require further tuning and testing of the algorithm.

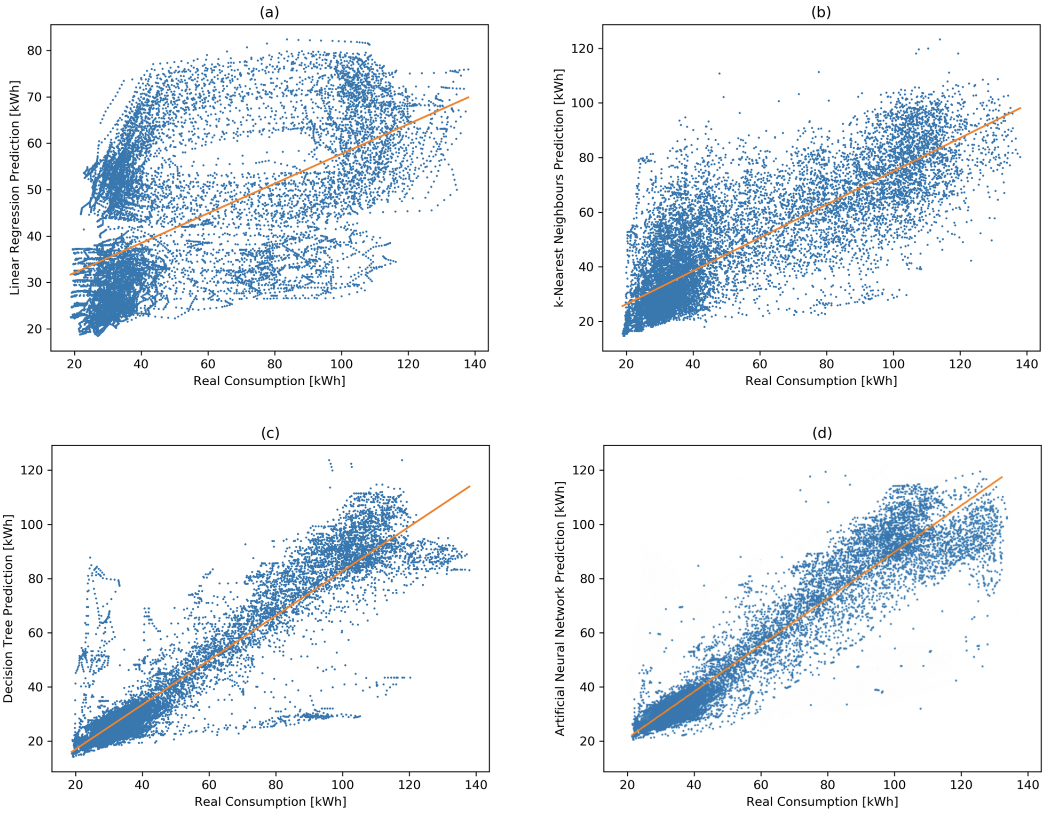


Figure 4.7: Comparison of predicted versus real consumption values for (a) multiple linear regression model, (b) k-nearest neighbours model, (c) decision tree model and (d) artificial neural network model.

Figure 4.7 shows scattered plots between modelled and real consumption values. It is possible to notice that the linear regression model present sparse points, whereas the kNN model captures more the relationship between the input features and the values to be predicted. The decision tree model shows a slightly better performance than the kNN model and it can be seen that the points are more dense the closer they are to the trend line. The artificial neural network shows a better linearity with the trend line, as expected by the higher coefficient of correlation between predicted and real values and proves to be the more precise tool to model the energy consumption in the civil building. Figure 4.8 shows the same trend in the results as Figure 4.7 but in this case, the input features included the auto-regressive feature relative to the consumption of the past 15 minutes (the plots represent the results for simulation 4 of all tested algorithms).

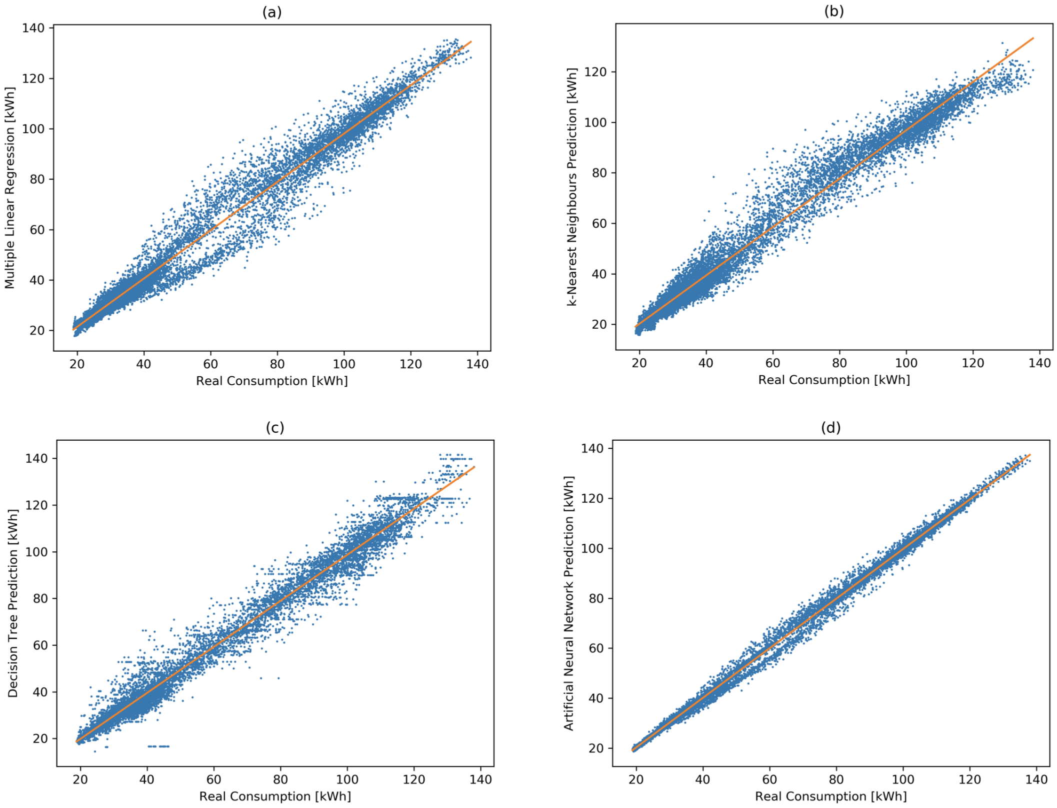


Figure 4.8: Comparison of predicted versus real consumption values for (a) multiple linear regression model, (b) k-nearest neighbours model, (c) decision tree model and (d) artificial neural network model including an autoregressive feature.

4.2 Electricity Production Model

As far as the solar power forecasting is concerned, the chosen forecasting approach involved a direct forecasting method (see section 3.2), i.e. the forecasting of the solar power output based on weather variables and the corresponding PV power data that has been previously calculated for the 120 kW installed on the civil building. Developing a machine learning approach for solar power forecasting results in a model which is specific of the location as the variation of the meteorological parameters and their related output is highly dependant on the specific PV plant layout and geographical location. This is due to the fact that the correlation of the meteorological parameters and the PV power output is not the same for different locations or for different technical specifications of the solar panel or inverter.

4.2.1 Correlation between Input Variables

Again, before starting the development of the actual predictive model, it is important to gain familiarity with the variables available. As it is possible to verify from Table 4.6, the PV power output has a high linear correlation with the solar radiation (70.9%) and with temperature (45.6%), which is expected since these two variables are the ones that influence the most the PV production and were used as inputs to the physical model (the Perez and the three-diode-one-parameter model) used to generate the power data for the machine learning model. The relative humidity is mildly correlated with the power output, which is reasonable as it probably accounts for more humid, rainy days. Wind speed, pressure and hour of the day are the last three variables that show a correlation higher than 10% but notably lower than the previously mentioned variables, 14.7%, 14.2% and 10.8% respectively.

Table 4.6: Correlation coefficients between variables for the PV forecasting model.

Correlation (%)	Variable 1	Variable 2
70.9%	Electricity Production [kWh]	Solar Radiation [W/m2]
45.6%	Electricity Production [kWh]	Temperature [°C]
31.5%	Electricity Production [kWh]	Relative Humidity [%]
14.7%	Electricity Production [kWh]	Wind Speed [m/s]
14.2%	Electricity Production [kWh]	Pressure [mbar]
10.8%	Electricity Production [kWh]	Hour of the day (0-24)

Figure 4.9 presents the scatter plots between the energy production and the weather variables. As expected, the relationship between the electricity production and the hour of the day shows a trend which reflects the average trend of the solar radiation during a sunny day. Plot (b) shows a strong linear correlation with the solar radiation and plot (a) a mild correlation with temperature. From the last scatter plot (plot (h)) it is immediate to identify that the power production increases with decreasing precipitation.

4.2.2 Multiple Linear Regression

In the case of multiple linear regression, the study of the correlation of the different meteorological inputs, such as solar radiation, humidity, wind speed and direction, and pressure, with PV power output, is of the utmost importance when building a multiple linear regression model. In a first attempt, as explained in

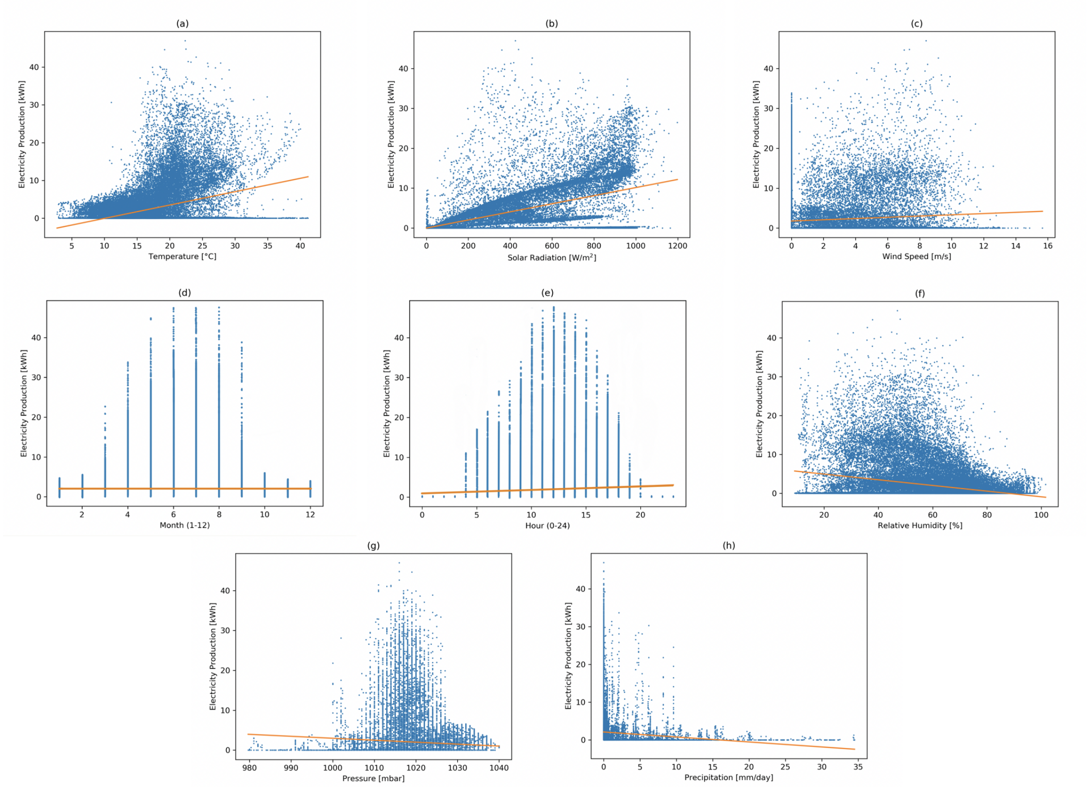


Figure 4.9: Scattered plots showing the correlation between energy production and (a) temperature, (b) solar radiation, (c) wind speed, (d) month, (e) hour, (f) relative humidity, (g) pressure and (h) precipitation.

section 3.3.1, only solar radiation and temperature where used as inputs, and the other variables were added iteratively to assess how the performance of the model was changing. By adding the hour of the day and the month, the model performance improved slightly, as it is possible to see from the MAE that decreases from 2.1 kWh to 1.7 kWh, the RMSE that went from 3.0 to 2.8 kWh and R^2 which increased from 0.36 to 0.37. Adding the additional feature of the pressure improved the model even further to an overall MAE of 1.5 kWh, a RMSE of 2.5 kWh, and a coefficient of determination of 0.39. The autoregressive feature relative to the PV power generation of the last 15 minutes improved the model notably (simulation 4).

Table 4.7: MAE, RMSE and coefficient of determination of the tested production multiple linear regression models for different input features.

Simulation	Radiation	Temperature	Relative Humidity	Wind Speed	Hour	Month	Pressure	Energy-1	MAE [kWh]	RMSE [kWh]	R^2
1	•	•	•	•					2.1	3.0	0.36
2	•	•	•	•	•	•			1.7	2.8	0.37
3	•	•	•	•	•	•	•		1.5	2.5	0.39
4	•	•	•	•	•	•	•	•	0.9	2.0	0.61

Figure 4.10 shows predicted value and real values of the power production for a week in November. It is to be noted that the term real in the ML models for the solar power production refers to the power data generated through the physical model, which has been validated with a week of real production measurements, as no historical data were available. From the graph, it can be seen that the model fails at capturing the trend of the production and is not even able to properly capture the absence of production at night (left graph) without the help of the autoregressive feature (right graph).

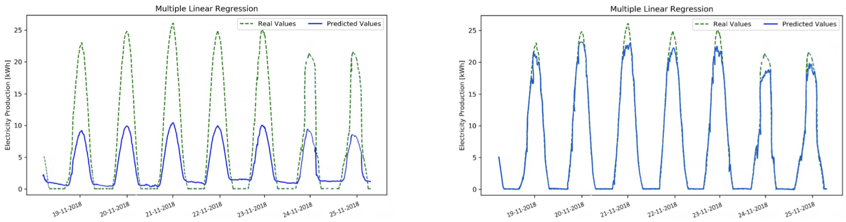


Figure 4.10: Multiple linear regression model versus real production data (simulation 3 and 4).

4.2.3 Decision Tree

The most performing decision tree implemented used solar radiation, temperature, relative humidity, wind speed, pressure and time related variables. The most performing tree that was not overfitting the data managed to reach a MAE of 1.1 kWh and a RMSE of 2.5 kWh accounting for a total R^2 of 0.37. Including the autoregressive feature improved the model even further to a MAE of 0.5 kWh, a RMSE of 1.9 kWh and a coefficient of determination of 0.66. Overall the decision tree performed better than the multiple linear regression as the algorithm was able to understand the trend of data.

Table 4.8: MAE, RMSE and coefficient of determination of the tested production decision tree regression models for different input features.

Simulation	Radiation	Temperature	Relative Humidity	Wind Speed	Hour	Month	Pressure	Energy-1	MAE [kWh]	RMSE [kWh]	R^2
1	•	•	•	•					1.9	2.9	0.33
2	•	•	•	•	•				1.6	2.6	0.35
3	•	•	•	•	•	•	•		1.1	2.5	0.37
4	•	•	•	•	•	•	•	•	0.5	1.9	0.66

Figure 4.11 confirmed that the decision tree model performs better than linear regression. The tree is able indeed to distinguish the production between day and night more precisely. The method fails though at modelling well the peaks of the power production during the day without overfitting (left graph) if no autoregressive feature is included (right graph).

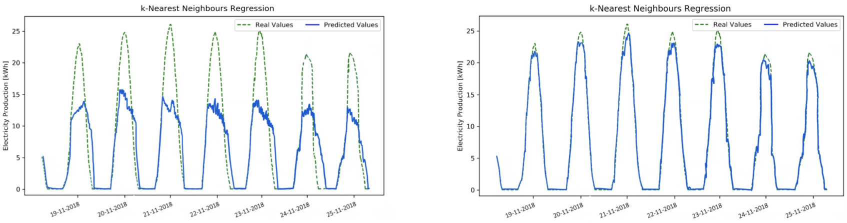


Figure 4.11: Decision tree regression model versus real production data (simulation 3 and 4).

4.2.4 K-Nearest Neighbours

The kNN regression model implemented the same features as the decision tree and the linear regression, but it was decided not to test it with a number of features above six, as the computational time would become too long to be considered for an application. The kNN algorithm performs slightly worse than the decision tree but overall its performance is quite close to it (see Table 4.9).

Table 4.9: MAE, RMSE and coefficient of determination of the tested production k-nearest neighbours regression models for different input features.

Simulation	Radiation	Temperature	Relative Humidity	Wind Speed	Hour	Month	Energy-1	MAE [kWh]	RMSE [kWh]	R^2
1	•	•	•	•				2.0	3.1	0.29
2	•	•	•	•	•			1.8	2.4	0.32
3	•	•	•	•	•	•		1.4	2.8	0.35
4	•	•	•	•	•	•	•	0.7	1.8	0.64

From Figure 4.12, it is clear how similarly the kNN algorithm performs to the decision tree algorithm. The biggest error in the kNN algorithm is the delayed decrease in the power production before night. At sunset, the algorithm overestimated production considerably (left graph). The same trend was present in the decision tree algorithm but it was less pronounced (see Figure 4.11). The autoregressive feature (right graph) though proved to be useful to help the algorithm follow a more realistic trend.

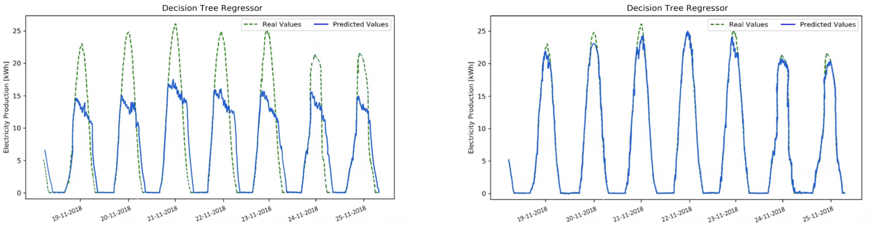


Figure 4.12: K-Nearest Neighbours regression model versus real production data (simulation 3 and 4).

4.2.5 Artificial Neural Network

The MLP is used to develop a suitable short-term forecasting model for 1 hour ahead solar power. After the tuning of its hyper-parameters, the best performing models consisted of two hidden layers with 10 and 20 neurons respectively. The input features used are the same as for the decision tree model and result in a MAE of 0.8 kWh, a RMSE of 2.4 kWh and a coefficient of correlation of 0.55.

Table 4.10: MAE, RMSE and coefficient of determination of the tested production artificial neural network models for different input features.

Simulation	Radiation	Temperature	Relative Humidity	Wind Speed	Hour	Month	Pressure	Energy-1	N neurons	MAE [kWh]	RMSE [kWh]	R^2
1	●	●	●	●					10, 20	1.5	2.8	0.45
2	●	●	●	●	●	●			10, 20	0.9	2.6	0.52
3	●	●	●	●	●	●	●		10, 20	0.8	2.4	0.55
4	●	●	●	●	●	●	●	●	10, 20	0.6	1.7	0.72

From Figure 4.13, it is clear how the MLP can reproduce well the pattern of the power production. Including the autoregressive feature, the model is able to capture both peaks and valleys considerably well. Its MAE is of 0.6 kWh in contrast to all the other model which presented MAEs above 1. The coefficient of correlation it accounts for is also much greater than in the other cases and its RMSE is the lowest of all tested models.

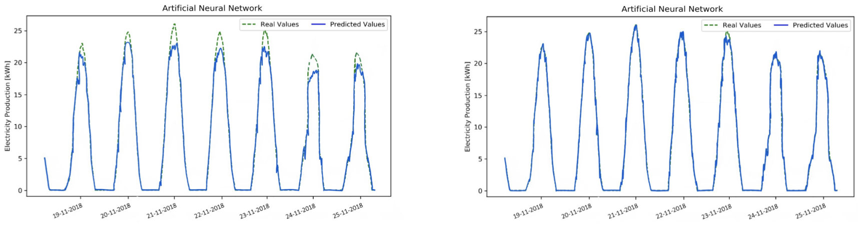


Figure 4.13: Artificial neural network model versus real production data (simulation 3 and 4).

4.2.6 Discussion of the Different Models

Analysing the four tested models for the solar power production the superior performance of the ANNs is clear. The linear regression, the decision tree and the kNN are not fully able to capture the relationship between all weather variables and the power output and present oscillatory trend during the day (kNN and decision tree) whereas the linear regression overestimates the performance at night. The ANN manages to understand the dynamics between the input features and the output, it models both the peaks of production during the day as well as no production at night.

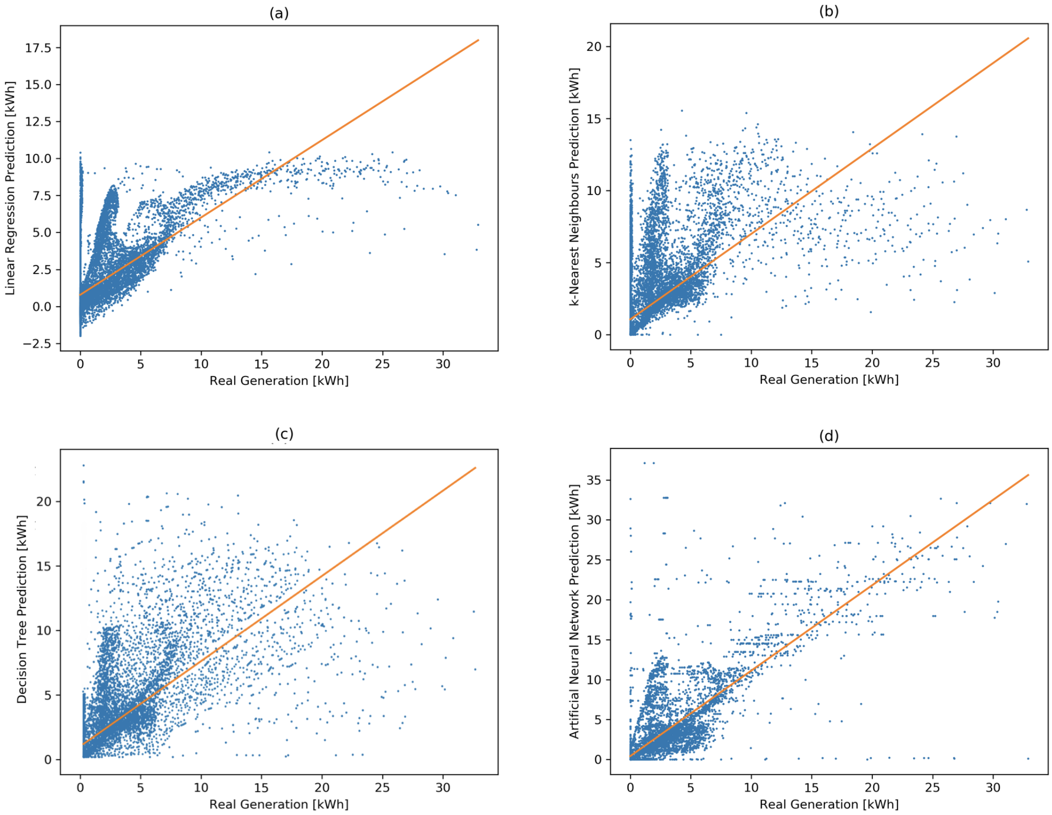


Figure 4.14: Comparison of predicted versus real production values for (a) multiple linear regression model, (b) k-nearest neighbours model, (c) decision tree model and (d) artificial neural network model.

Figure 4.14 represents scatter plots between the prediction and the real values for the four tested models. The main takeaways are that the linear regression model is not even close to accurately model the dynamics of the solar power production and its forecast values are much lower than the real ones. The kNN and the decision tree algorithm already perform better but it can be seen that for higher values of the real production they underestimate quite considerably. The ANN instead manages to concentrate the points along the trend line and shows a more symmetric graph than all other models. Figure 4.15 shows similar scatter plots for the simulation which included the autoregressive feature. The relative performance between the tested algorithms remains the same, as ANN prove to be superior in modelling power production even in this case. All models though present significant improvements in the prediction and in comparison to Figure 4.14, their prediction values lie much closer to the trend line than without the autoregressive feature.

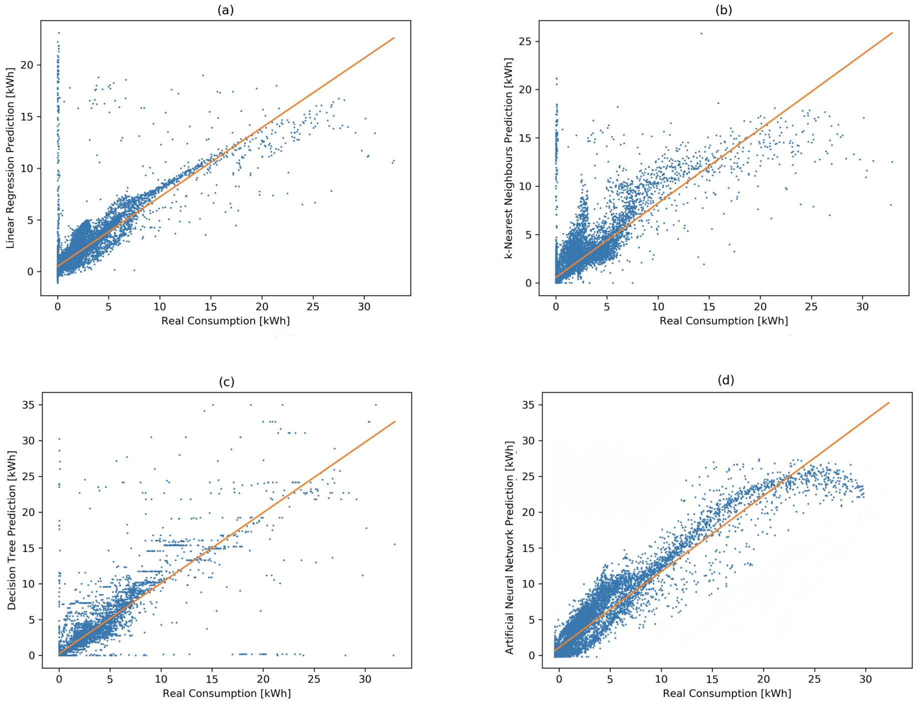


Figure 4.15: Comparison of predicted versus real production values for (a) multiple linear regression model, (b) k-nearest neighbours model, (c) decision tree model and (d) artificial neural network model including an autoregressive feature.

4.3 Time Horizon

The results presented in the previous section, both for the electricity consumption and solar power production considered a forecast horizon of 1 hour. With the aim of having an algorithm that could be further used to implement energy management strategies, simulations have been performed also for a longer forecast horizon, i.e. 24 hour. Table 4.11 reports the results for a longer forecasting horizon. The electricity consumption model show a slight improvement of its results, as the time related variables used as inputs to the model, help it understand the dynamics of the correlation between consumption and the input features used. In the case of the solar power production, the horizon of the forecast assumes a bigger importance in the model as the input features are more related to weather variables than to time dependent variable or occupancy trend. Considering the case when clouds suddenly appear in the sky, then there is no way for the model to know that and adapt itself based on previously seen observations. A possible solution to this could be feeding the model with forecasts of weather related variables and assess again its performance.

Table 4.11: Performance comparison based on forecast horizon.

Simulation	MAE [kWh]	MAE [kWh]	RMSE [kWh]	RMSE [kWh]	R^2	R^2
	1 hour	24 hours	1 hour	24 hours	1 hour	24 hours
Electricity Consumption	1.4	1.2	2.9	2.6	0.97	0.98
Solar Power Production	0.6	1.1	1.7	2.5	0.72	0.55

4.4 Limitations of Data-Driven Models

Data driven models have proven to be useful for prediction of both building energy consumption and solar power forecasting. However, to be considered for specific applications, it is important to consider the training set used as the model might not be suitable to explore scenarios that go beyond their training range. The models indeed might not be able to cope well with new instances that go beyond what it has learned, as they would not be able to generalise and make accurate predictions, especially if they were tested with limited data set. For example, a model that was trained with data collected from one building, might not perform well on unseen testing data representing another building, as it might have different physical properties, occupancy behaviours or operation strategies. The training set needs to have a sufficient variety to be representative for a specific application.

As mentioned in section 3.1, to further confirm the feasibility of the forecasting algorithm and input features proposed, simulations similar to the ones carried out for the civil building have been carried out for three other buildings, namely the central building, the north tower and south tower. The results are reported in Table 4.12, Table 4.14 and Table 4.13 respectively, and proved that the approach followed for the civil pavilion can be generalised to other buildings of the campus.

The second major problem of machine learning prediction models is that they function as black-box models. The internal procedures are not fully known and even though their prediction accuracy might be sufficient, they might be limited when it comes to providing an understanding of the causes behind them. To address this issue, physical models in combination with machine learning models (i.e. hybrid, grey-box models) help to minimise the disadvantages of both approaches.

Table 4.12: MAE, RMSE and coefficient of determination of the best models tested for the central pavilion.

Central Pavillion	Weekday	Day	Month	Hour	Minute	Occupancy	Holiday	Energy-1	Radiation	MAE [kWh]	RMSE [kWh]	R^2
Linear Regression	•	•	•	•	•	•	•	•	•	4.8	6.4	0.90
Decision Tree	•	•	•	•	•	•	•	•		3.4	4.2	0.95
K-Nearest Neighbours	•	•	•	•	•	•	•	•		3.6	4.4	0.94
Artificial Neural Network	•	•	•	•	•	•	•	•		2.8	3.0	0.97

Table 4.13: MAE, RMSE and coefficient of determination of the best models tested for the south tower.

South Tower	Weekday	Day	Month	Hour	Minute	Occupancy	Holiday	Energy-1	Radiation	MAE [kWh]	RMSE [kWh]	R^2
Linear Regression	•	•	•	•	•	•	•	•	•	5.0	7.5	0.91
Decision Tree	•	•	•	•	•	•	•	•		3.7	4.7	0.94
K-Nearest Neighbours	•	•	•	•	•	•	•	•		3.8	4.9	0.92
Artificial Neural Network	•	•	•	•	•	•	•	•		3.0	3.2	0.95

Table 4.14: MAE, RMSE and coefficient of determination of the best models tested for the north tower.

North Tower	Weekday	Day	Month	Hour	Minute	Occupancy	Holiday	Energy-1	Radiation	MAE [kWh]	RMSE [kWh]	R^2
Linear Regression	•	•	•	•	•	•	•	•	•	4.9	5.6	0.88
Decision Tree	•	•	•	•	•	•	•	•		3.2	4.0	0.91
K-Nearest Neighbours	•	•	•	•	•	•	•	•		3.8	4.1	0.92
Artificial Neural Network	•	•	•	•	•	•	•	•		2.9	3.2	0.98

4.5 Demand-Response Model

This section presents the results of the optimization of the battery schedule which had the goal to assess the profitability of the implementation of a BESS system to exploit load shifting from high-price intervals to low-price intervals, when tariffs favour off-peak consumption. Initially multiple simulations had to be carried out to find an optimal reasonable combination of capacity and maximum charging and discharging power, which would allow to exploit different time of use of electricity while taking into account battery life and the economics of the investment. The simulations were first run for a representative week in May, as it was clear that the biggest economic savings would arise from weeks of regular activities and classes in the building, rather than in summer or holiday periods. Once the optimal combination was found, yearly long simulations have been run and the profitability of the proposal was assessed taking into consideration the initial investment for the battery.

4.5.1 Sensitivity Analysis

Once the system is modelled, a sensitivity analysis was carried out, varying the size of the battery to find the one that would allow to achieve the most savings along its lifetime. To this purpose, following IRENA's 'Electricity storage and renewables' report [97] and Bloomberg technology report [98] a battery lifetime of 15 years and a corresponding industry price of 300 €/kWh Lithium-Ion batteries was assumed. Both reports highlight how Lithium-Ion battery system prices have fallen from 600 €/kWh in 2013 to around 275 €/kWh and are projected to fall even below 200 €/kWh in the coming few years, implying stand alone batteries are prone to become increasingly employed. The results of the possible annual savings are shown in Table 4.15 and are calculated comparing the electricity bill for the given tariff (see section 3.4.1) with and without the optimally scheduled battery system.

Table 4.15: Battery savings resulting from exploiting ToU electricity tariff.

Battery (Capacity, Power)	Post-Optimisation Cost (€)	Savings (€)	Savings (%)
300 kWh, 150 kW	188335.39	1583.36	0.84
500 kWh, 200 kW	187491.87	2426.88	1.28
700 kWh, 300 kW	186958.15	2960.60	1.56
800 kWh, 350 kW	186839.21	3079.54	1.62
900 kWh, 400 kW	186801.40	3117.35	1.64
1000 kWh, 400 kW	186820.18	3098.57	1.63

From Table 4.15 it can be seen that increasing the capacity reduces the minimal achievable cost for electricity and therefore increases slightly the savings. However, after a certain threshold, no further improvement can be noticed, as the battery price becomes too high compared with the possible savings (Figure 4.16). As the goal is to achieve the biggest possible savings while avoiding paying for an oversized battery, the cost-optimum solution for the civil building could be a 900 kWh battery.

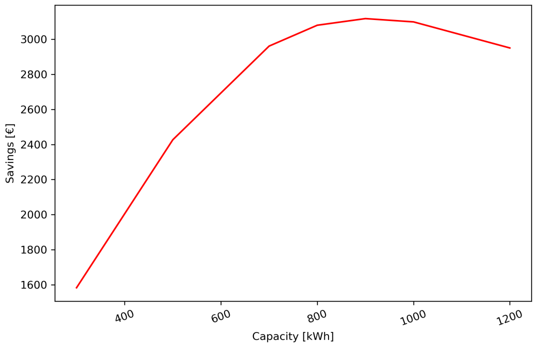


Figure 4.16: Possible savings with an optimally scheduled battery versus battery capacity.

4.6 Limitations of the Demand-Response Model

A significant limitation to the model is that a single pricing scheme (the EDP price scheme) is considered, whereas it could be interesting to investigate other tariffs with different pricing schemes which depends on the peak demand and load type and the possibility of stacking multiple revenue streams. Additionally, as the peak demand charge is calculated and charged at campus level and not at building level, the optimisation only takes into account the price of the energy and not the one related to the daily, monthly or yearly peak power depending on the type of contract with the utility. If that was the case, there is potential for additional savings should batteries be deployed throughout campus. In this case, the battery scheduling would try to exploit the period with a low tariff as well as reducing power demand peaks. Another limitation is that the battery degradation, cycling and maintenance mechanisms are not considered, and as such it does not accurately reflect energy losses occurring within the BESS. Considering the initial investment, other revenue streams or arbitrage opportunities for the battery could be investigated and implemented in the model. It would be wise as well to assess market conditions not only as they are nowadays but also how they are going to shift and evolve in the years to come, especially in relation to the price of batteries.

Finally, the model developed was built considering historical data, but being able to accurately forecast the load and the solar power production, the battery scheduling could be implemented online on a rolling basis to eliminate the assumption of perfect forecast connected to the offline model developed.

Chapter 5

Conclusion and Future Work

This final chapter aims to conclude with a brief summary of the work, the accomplishments made, and gives suggestion as to how the research could potentially be extended further. The main objective was to develop two models which would forecast the electricity consumption and production for some buildings of Instituto Superior Técnico at a predefined forecast horizon and assess which input variables to each of these models were paramount to obtain accurate results. To achieve this, an additional physical model to assess the electricity production of some rooftop solar panels had to be developed. Eventually, the results were then used to suggest possible energy management strategies. In particular, a MILP model was developed to schedule a battery system and quantify possible economic savings resulting from exploiting off-peak hours to charge the battery. The achievement of this goal involved multiple steps, which are explained in the following section together with an overview of possible further works related to this thesis.

5.1 Discussion and Conclusion

The initial task carried out was a through analysis of literature, which contributed to provide support in the development of this work while at the same time allowing it to differentiate itself from the others. The literature analysed the current state of the art involving both solar radiation models, data-driven models of both building electricity consumption and solar power generation, and possible demand-response options. In a second moment, Python, the chosen tool to carry out the thesis was studied and its packages thoroughly comprehended so that all models could be built using the same programming language. Next came the creation of different models.

The first model allowed for the simulation of the PV power output which was needed to carry out the predictions as no historical data was available. The model used the Perez model to transpose the total irradiance measured on the horizontal plane to the tilted plane of the panel and the three diode one parameter model to model the solar cell. Based on the simulations performed to validate the model using a week of historical data, the error of the model was proved to be between 5% and 10% compared to the recorded power data. As already noticed from literature, the Perez model probably overestimated the solar irradiance received by the panel which results in slightly higher values for the produced solar power.

Following this, the comparison of different data-driven models for both electricity consumption and power generation was performed, and included learning algorithms based on different methods such as distance-based algorithms, decision tree algorithms and artificial neural networks. The development of these models allowed the identification of the variables that best describe the consumption patterns in the analysed buildings, that proved to be highly correlated with the occupancy and the time of the day and of the year. The inclusion of such variables in the forecasting of electricity consumption allowed to consistently increase the performance index of the predictions. The occupancy data follows indeed a clear trend which can be noticed in the consumption as well and is strongly related to the activities occurring in the building. The inclusion of the knowledge of the academic calendar helped the model performance as well, by allowing to distinguish between actual working days and holiday periods. The model that proved to be the most suitable to predict the electricity consumption was the artificial neural network model which showed a MAE of 1.4 kWh, a RMSE of 2.9 kWh, and a coefficient of correlation of 0.97 when including input features related to the occupancy, the academic calendar and the time. Weather variables did not contribute to the improvement of the model performance indexes, indicating that electricity consumption is more correlated to occupancy than to weather conditions. The same procedure was carried out for the solar power forecasting. In this case, the artificial neural network model was again the one able to represent better how power is generated, showing a MAE of 0.6 kWh, a RMSE of 1.7 and a coefficient of correlation of 0.72. For this last model, the key input features were, as expected, the solar radiation, the temperature, the humidity, the wind speed and features related to time, such as hour of the day and month. These features were expected as the initial physical solar model developed uses the inputs related to the time of the year and of the day to calculate the angles between the sun rays and the panel surface and radiation and temperature to quantify the power output. The comparison of the above-mentioned data-driven models for both electricity consumption and power generation led to two different artificial neural network models able to carry out predictions with a relatively good accuracy and required countless hours of coding and iterations to find hyper-parameters useful for the scope.

With these models in place, a simulation was performed to assess possible economic savings resulting from the implementation of demand-response strategies exploiting the flexibility of a BESS system. A MILP optimisation was set up to find the optimum battery size and schedule which would allow to shave the daily peaks of energy consumption which characterise the building and 'shift' them to off-peak hours when the electricity tariff is lower. The results showed that using a behind-the-meter BESS could potentially lead to economic savings. Such savings are though highly dependent on the size of the battery and its consequent upfront investment. Assuming an optimistic battery lifetime, the net savings that could be achieved with an optimised schedule represent at most 1.64% of the total energy costs.

The optimisation developed did not take into account battery degradation mechanisms which lead to the reasonable assumption that the possible savings in reality would not be sufficient to reach break-even for the investment. On the other hand, the forecast possible decrease in the coming years in battery economics could alter the results of the proposed business case and prove it to be lucrative.

5.2 Recommendation for Future Work

Some areas that could be further explored and possibly extended related to this work include:

- It would be wise to validate the solar production model with a longer dataset of real historical data to confirm the model calculates an accurate amount of power even outside the range against which it was tested.
- It could be interesting to consider using typical meteorological years (TMY) for Lisbon as input to the solar model instead of real-data coming for the Técnico meteo station. Some of the radiation data provided were indeed strangely high for the location, which could be due to an inaccurate calibration of the radiation measuring devices. A simulation with a TMY could provide interesting and possibly more insightful results for a data-driven approach. The data coming from a TMY database is indeed the result of long-term averages of meteorological conditions for a year given a specific location. The TMY values are generated from years of data much longer than a year and specifically selected to present the range of weather of the location with a P50 case, which means these values will be exceeded in 50% of all years.
- The power forecasting model prediction accuracy could be boosted in a real-time implementation by eliminating the calculation of night time data (between sunset and sunrise). As the power production is zero, the model would not have to make a trade-off between fitting day time data and night time data, reducing this way the computational effort.
- The integration of an automatic API for the recognition of national holidays without a fixed day and automatic inclusion of the academic calendar would be useful for a real-time implementation of the forecasting.
- As far as the real-time forecasting is concerned, the software which was used to extract the data related to occupancy (RRDTool) offers the possibility to communicate with the wireless controller to grab statistical data from it which would definitely increase the accuracy of the predictions in the building electricity consumption model, as the results presented involved averages of historical data, as no data with a higher resolution was available.
- Concerning the actual implementation of the prediction models, the investigation of the model real-time performance and computational effort should be verified.
- As mentioned in the previous chapter, a penalty factor for battery cycling and degradation mechanics could be implemented to achieve more realistic battery operation modelling.
- An opportunity cost that the implement model does not account for involves the scheduling of the battery considering not only the energy tariff but the maximum power demand over the day, over the month or over the year depending on the specific contract with the utility company. Introducing the different charges related to peak power demand could result in an additional savings, or even a possible reduction of the contracted power. To this purpose, other variables in the MILP optimisation such as the peak power and its related charge could be introduced and an optimisation at campus level employing multiple batteries could be carried out.

- Similarly, instead of iteratively optimise for different battery sizes to find the optimum one, the optimisation could concurrently cost-optimise the battery sizing and operation, by defining as additional variables the battery capacity and maximum charging and discharging rate, instead of having them as fixed parameters, especially when considering multiple batteries.

- A limitation to the battery model is that it optimises the schedule considering the current EDP MV tariff. The same methodology and model could be adapted to investigate other electricity tariffs related to energy and/or peak demand charges available.

- In the eventuality that the PV power production is not directly consumed and there is excess power produced, the MILP optimisation model could be modified to account for an eventual storage of surplus solar power and/or a selling price of electricity to the grid, if it proves to be more profitable.

- Finally, the biggest drawback of batteries is their high initial capital cost. To compensate it, the investigation of the stacking of additional revenue streams such as ancillary services and not only tariff arbitrage could be investigated.

Bibliography

[1] International Energy Agency. *Cities, Towns and Renewable Energy: Yes in My Front Yard.* OECD, Dec. 11, 2009. 186 pp. ISBN: 9264076875. URL: https://www.ebook.de/de/product/10702691/organization_for_economic_cooperation_an_cities_towns_and_renewable_energy_yes_in_my_front_yard.html.

[2] International Energy Agency. *Transition to Sustainable Buildings: Strategies and Opportunities to 2050.* Tech. rep. International Energy Agency, 2013.

[3] Patrik Thollander, Patrik Rohdin, and Bahram Moshfegh. “On the formation of energy policies towards 2020: Challenges in the Swedish industrial and building sectors”. In: *Energy Policy* 42 (Mar. 2012), pp. 461–467. DOI: 10.1016/j.enpol.2011.12.012.

[4] Andrea Trianni et al. “Energy management: A practice-based assessment model”. In: *Applied Energy* 235 (Feb. 2019), pp. 1614–1636. DOI: 10.1016/j.apenergy.2018.11.032.

[5] Catherine Cooremans and Alain Schönenberger. “Energy management: A key driver of energy-efficiency investment?” In: *Journal of Cleaner Production* 230 (Sept. 2019), pp. 264–275. DOI: 10.1016/j.jclepro.2019.04.333.

[6] E.A. Abdelaziz, R. Saidur, and S. Mekhilef. “A review on energy saving strategies in industrial sector”. In: *Renewable and Sustainable Energy Reviews* 15.1 (Jan. 2011), pp. 150–168. DOI: 10.1016/j.rser.2010.09.003.

[7] B. Swords, E. Coyle, and B. Norton. “An enterprise energy-information system”. In: *Applied Energy* 85.1 (Jan. 2008), pp. 61–69. DOI: 10.1016/j.apenergy.2007.06.009.

[8] Marvin T. Howell. *Effective Implementation of an ISO 50001 Energy Management System (EnMS).* ASQ Quality Press, 2014. ISBN: 9780873898720.

[9] International Energy Agency. *Harnessing Variable Renewables. A Guide to the Balancing Challenge.* Tech. rep. 2011.

[10] Smart Grid Task Force. *2015 Regulatory Recommendations for the Deployment of Flexibility - EG3 REPORT.* Tech. rep. Smart Grid Task Force, 2015.

[11] E. Ela et al. “Wholesale electricity market design with increasing levels of renewable generation: Incentivizing flexibility in system operations”. In: *The Electricity Journal* 29.4 (May 2016), pp. 51–60. DOI: 10.1016/j.tej.2016.05.001.

[12] Pol Olivella-Rosell et al. “Local Flexibility Market Design for Aggregators Providing Multiple Flexibility Services at Distribution Network Level”. In: *Energies* 11.4 (Apr. 2018), p. 822. DOI: 10.3390/en11040822.

[13] Eric Hsieh and Robert Anderson. “Grid flexibility: The quiet revolution”. In: *The Electricity Journal* 30.2 (Mar. 2017), pp. 1–8. DOI: 10.1016/j.tej.2017.01.009.

[14] P. Lloret et al. *Smart system of renewable energy storage based on INtegrated EVs, bAtteries to empower mobile, Distributed, and centralised Energy storage in the distribution grid*. Tech. rep. 2017.

[15] Søren Østergaard Jensen et al. “IEA EBC Annex 67 Energy Flexible Buildings”. In: *Energy and Buildings* 155 (Nov. 2017), pp. 25–34. DOI: 10.1016/j.enbuild.2017.08.044.

[16] Rune Grønborg Junker et al. “Characterizing the energy flexibility of buildings and districts”. In: *Applied Energy* 225 (Sept. 2018), pp. 175–182. DOI: 10.1016/j.apenergy.2018.05.037.

[17] Jimeno A. Fonseca and Arno Schlueter. “Integrated model for characterization of spatiotemporal building energy consumption patterns in neighborhoods and city districts”. In: *Applied Energy* 142 (Mar. 2015), pp. 247–265. DOI: 10.1016/j.apenergy.2014.12.068.

[18] Sebastian Stinner, Kristian Huchtemann, and Dirk Müller. “Quantifying the operational flexibility of building energy systems with thermal energy storages”. In: *Applied Energy* 181 (Nov. 2016), pp. 140–154. DOI: 10.1016/j.apenergy.2016.08.055.

[19] Antonio Luque. *Handbook of Photovoltaic Science and Engineering*. Wiley-Blackwell, Dec. 21, 2010. 1164 pp. ISBN: 0470721693. URL: https://www.ebook.de/de/product/9604647/antonio_luque_handbook_of_photovoltaic_science_and_engineering.html.

[20] P.G. Loutzenhiser et al. “Empirical validation of models to compute solar irradiance on inclined surfaces for building energy simulation”. In: *Solar Energy* 81.2 (Feb. 2007), pp. 254–267. DOI: 10.1016/j.solener.2006.03.009.

[21] John A. Duffie and William A. Beckman. *Solar Engineering of Thermal Processes*. John Wiley & Sons, Inc., Apr. 2013. DOI: 10.1002/9781118671603.

[22] Richard Perez et al. “Modeling daylight availability and irradiance components from direct and global irradiance”. In: *Solar Energy* 44.5 (1990), pp. 271–289.

[23] R. Perez et al. “A new simplified version of the Perez diffuse irradiance model for tilted surfaces”. In: *Solar Energy* 39 (1987), pp. 221–232.

[24] Karthik Ramasubramanian and Abhishek Singh. *Machine Learning Using R*. Apress, 2017. DOI: 10.1007/978-1-4842-2334-5.

[25] Janusz Kacprzyk and Witold Pedrycz, eds. *Springer Handbook of Computational Intelligence*. Springer Berlin Heidelberg, 2015. DOI: 10.1007/978-3-662-43505-2.

[26] Yixuan Wei et al. “A review of data-driven approaches for prediction and classification of building energy consumption”. In: *Renewable and Sustainable Energy Reviews* 82 (Feb. 2018), pp. 1027–1047. DOI: 10.1016/j.rser.2017.09.108.

[27] Anibal de Almeida et al. “Characterization of the household electricity consumption in the EU, potential energy savings and specific policy recommendations”. In: *Energy and Buildings* 43.8 (Aug. 2011), pp. 1884–1894. DOI: 10.1016/j.enbuild.2011.03.027.

[28] Richard E. Edwards, Joshua New, and Lynne E. Parker. “Predicting future hourly residential electrical consumption: A machine learning case study”. In: *Energy and Buildings* 49 (June 2012), pp. 591–603. DOI: 10.1016/j.enbuild.2012.03.010.

[29] Hai Zhong et al. “Vector field-based support vector regression for building energy consumption prediction”. In: *Applied Energy* 242 (May 2019), pp. 403–414. DOI: 10.1016/j.apenergy.2019.03.078.

[30] Henrique Pombeiro et al. “Comparative assessment of low-complexity models to predict electricity consumption in an institutional building: Linear regression vs. fuzzy modeling vs. neural networks”. In: *Energy and Buildings* 146 (July 2017), pp. 141–151.

[31] Geoffrey K.F. Tso and Kelvin K.W. Yau. “Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks”. In: *Energy* 32.9 (Sept. 2007), pp. 1761–1768. DOI: 10.1016/j.energy.2006.11.010.

[32] Priyanka Singh and Pragya Dwivedi. “Integration of new evolutionary approach with artificial neural network for solving short term load forecast problem”. In: *Applied Energy* 217 (May 2018), pp. 537–549. DOI: 10.1016/j.apenergy.2018.02.131.

[33] L.G.B. Ruiz et al. “Energy consumption forecasting based on Elman neural networks with evolutive optimization”. In: *Expert Systems with Applications* 92 (Feb. 2018), pp. 380–389. DOI: 10.1016/j.eswa.2017.09.059.

[34] Saima Hassan et al. “A systematic design of interval type-2 fuzzy logic system using extreme learning machine for electricity load demand forecasting”. In: *International Journal of Electrical Power & Energy Systems* 82 (Nov. 2016), pp. 1–10. DOI: 10.1016/j.ijepes.2016.03.001.

[35] Aowabin Rahman, Vivek Srikumar, and Amanda D. Smith. “Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks”. In: *Applied Energy* 212 (Feb. 2018), pp. 372–385. DOI: 10.1016/j.apenergy.2017.12.051.

[36] Guillermo Escrivá-Escrivá et al. “New artificial neural network prediction method for electrical consumption forecasting based on building end-uses”. In: *Energy and Buildings* 43.11 (Nov. 2011), pp. 3112–3119. DOI: 10.1016/j.enbuild.2011.08.008.

[37] Saima Hassan, Abbas Khosravi, and Jafreezal Jaafar. “Examining performance of aggregation algorithms for neural network-based electricity demand forecasting”. In: *International Journal of Electrical Power & Energy Systems* 64 (Jan. 2015), pp. 1098–1105. DOI: 10.1016/j.ijepes.2014.08.025.

[38] Khuram Pervez Amber et al. “Energy Consumption Forecasting for University Sector Buildings”. In: *Energies* 10.10 (Oct. 2017), p. 1579. DOI: 10.3390/en10101579.

[39] M.A. Rafe Biswas, Melvin D. Robinson, and Nelson Fumo. “Prediction of residential building energy consumption: A neural network approach”. In: *Energy* 117 (Dec. 2016), pp. 84–92. DOI: 10.1016/j.energy.2016.10.066.

[40] Chirag Deb et al. “Forecasting Energy Consumption of Institutional Buildings in Singapore”. In: *Procedia Engineering* 121 (2015), pp. 1734–1740. DOI: 10.1016/j.proeng.2015.09.144.

[41] Radu Platon, Vahid Raissi Dehkordi, and Jacques Martel. “Hourly prediction of a building’s electricity consumption using case-based reasoning, artificial neural networks and principal component analysis”. In: *Energy and Buildings* 92 (Apr. 2015), pp. 10–18. DOI: 10.1016/j.enbuild.2015.01.047.

[42] K.P. Amber et al. “Intelligent techniques for forecasting electricity consumption of buildings”. In: *Energy* 157 (Aug. 2018), pp. 886–893. DOI: 10.1016/j.energy.2018.05.155.

[43] Kangji Li, Hongye Su, and Jian Chu. “Forecasting building energy consumption using neural networks and hybrid neuro-fuzzy system: A comparative study”. In: *Energy and Buildings* 43.10 (Oct. 2011), pp. 2893–2899. DOI: 10.1016/j.enbuild.2011.07.010.

[44] Pedro A. González and Jesús M. Zamarreño. “Prediction of hourly energy consumption in buildings based on a feedback artificial neural network”. In: *Energy and Buildings* 37.6 (June 2005), pp. 595–601. DOI: 10.1016/j.enbuild.2004.09.006.

[45] Alberto Hernandez Neto and Flávio Augusto Sanzovo Fiorelli. “Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption”. In: *Energy and Buildings* 40.12 (Jan. 2008), pp. 2169–2176. DOI: 10.1016/j.enbuild.2008.06.013.

[46] P.M.R. Bento et al. “Optimization of neural network with wavelet transform and improved data selection using bat algorithm for short-term load forecasting”. In: *Neurocomputing* (May 2019). DOI: 10.1016/j.neucom.2019.05.030.

[47] Gabriel Trierweiler Ribeiro, Viviana Cocco Mariani, and Leandro dos Santos Coelho. “Enhanced ensemble structures using wavelet neural networks applied to short-term load forecasting”. In: *Engineering Applications of Artificial Intelligence* 82 (June 2019), pp. 272–281. DOI: 10.1016/j.engappai.2019.03.012.

[48] Rui Zhang et al. “A composite k-nearest neighbor model for day-ahead load forecasting with limited temperature forecasts”. In: *2016 IEEE Power and Energy Society General Meeting (PESGM)*. IEEE, July 2016. DOI: 10.1109/pesgm.2016.7741097.

[49] Guo-Feng Fan et al. “Application of the Weighted K-Nearest Neighbor Algorithm for Short-Term Load Forecasting”. In: *Energies* 12.5 (Mar. 2019), p. 916. DOI: 10.3390/en12050916.

[50] Mashud Rana, Irena Koprinska, and Vassilios G. Agelidis. “Univariate and multivariate methods for very short-term solar photovoltaic power forecasting”. In: *Energy Conversion and Management* 121 (Aug. 2016), pp. 380–390. DOI: 10.1016/j.enconman.2016.05.025.

[51] Jie Shi et al. “Forecasting power output of photovoltaic system based on weather classification and support vector machine”. In: *2011 IEEE Industry Applications Society Annual Meeting*. IEEE, Oct. 2011. DOI: 10.1109/ias.2011.6074294.

[52] Jun Liu et al. “An Improved Photovoltaic Power Forecasting Model With the Assistance of Aerosol Index Data”. In: *IEEE Transactions on Sustainable Energy* 6.2 (Apr. 2015), pp. 434–442. DOI: 10.1109/tste.2014.2381224.

[53] Hugo T.C. Pedro and Carlos F.M. Coimbra. “Assessment of forecasting techniques for solar power production with no exogenous inputs”. In: *Solar Energy* 86.7 (July 2012), pp. 2017–2028. DOI: 10.1016/j.solener.2012.04.004.

[54] Changsong Chen et al. “Online 24-h solar power forecasting based on weather type classification using artificial neural network”. In: *Solar Energy* 85.11 (Nov. 2011), pp. 2856–2870. DOI: `10.1016/j.solener.2011.08.027`.

[55] Federica Davò et al. “Post-processing techniques and principal component analysis for regional wind power and solar irradiance forecasting”. In: *Solar Energy* 134 (Sept. 2016), pp. 327–338. DOI: `10.1016/j.solener.2016.04.049`.

[56] Caroline Persson et al. “Multi-site solar power forecasting using gradient boosted regression trees”. In: *Solar Energy* 150 (July 2017), pp. 423–436. DOI: `10.1016/j.solener.2017.04.066`.

[57] Jie Shi et al. “Forecasting Power Output of Photovoltaic Systems Based on Weather Classification and Support Vector Machines”. In: *IEEE Transactions on Industry Applications* 48.3 (May 2012), pp. 1064–1069. DOI: `10.1109/tia.2012.2190816`.

[58] Min-Cheol Kang et al. “Development of algorithm for day ahead PV generation forecasting using data mining method”. In: *2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE, Aug. 2011. DOI: `10.1109/mwscas.2011.6026333`.

[59] Mashud Rana, Irena Koprinska, and Vassilios G Agelidis. “Forecasting solar power generated by grid connected PV systems using ensembles of neural networks”. In: *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2015. DOI: `10.1109/ijcnn.2015.7280574`.

[60] Gokhan Mert Yagli, Dazhi Yang, and Dipti Srinivasan. “Automatic hourly solar forecasting using machine learning models”. In: *Renewable and Sustainable Energy Reviews* 105 (May 2019), pp. 487–498. DOI: `10.1016/j.rser.2019.02.006`.

[61] Adel Mellit and Alessandro Massi Pavan. “A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy”. In: *Solar Energy* 84.5 (May 2010), pp. 807–821. DOI: `10.1016/j.solener.2010.02.006`.

[62] Paras Mandal et al. “Forecasting Power Output of Solar Photovoltaic System Using Wavelet Transform and Artificial Intelligence Techniques”. In: *Procedia Computer Science* 12 (2012), pp. 332–337. DOI: `10.1016/j.procs.2012.09.080`.

[63] Maria Grazia De Giorgi, Maria Malvoni, and Paolo Maria Congedo. “Photovoltaic power forecasting using statistical methods: impact of weather data”. In: *IET Science, Measurement & Technology* 8.3 (May 2014), pp. 90–97. DOI: `10.1049/iet-smt.2013.0135`.

[64] Alberto Dolara, Sonia Leva, and Giampaolo Manzolini. “Comparison of different physical models for PV power output prediction”. In: *Solar Energy* 119 (Sept. 2015), pp. 83–99. DOI: `10.1016/j.solener.2015.06.017`.

[65] H. Mori and M. Takahashi. “Development of GRBFN with global structure for PV generation output forecasting”. In: *2012 IEEE Power and Energy Society General Meeting*. IEEE, July 2012. DOI: `10.1109/pesgm.2012.6345673`.

[66] Ruidong Xu, Hao Chen, and Xiaoyan Sun. “Short-term photovoltaic power forecasting with weighted support vector machine”. In: *2012 IEEE International Conference on Automation and Logistics*. IEEE, Aug. 2012. DOI: `10.1109/ical.2012.6308206`.

[67] Joao Gari da Silva Fonseca Junior et al. “Forecasting Regional Photovoltaic Power Generation - A Comparison of Strategies to Obtain One-Day-Ahead Data”. In: *Energy Procedia* 57 (2014), pp. 1337–1345. DOI: `10.1016/j.egypro.2014.10.124`.

[68] S. Leva et al. “Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power”. In: *Mathematics and Computers in Simulation* 131 (Jan. 2017), pp. 88–100. DOI: `10.1016/j.matcom.2015.05.010`.

[69] Marco Cococcioni, Eleonora D’Andrea, and Beatrice Lazzerini. “24-hour-ahead forecasting of energy production in solar PV systems”. In: *2011 11th International Conference on Intelligent Systems Design and Applications*. IEEE, Nov. 2011. DOI: `10.1109/isda.2011.6121835`.

[70] M. Lipperheide, J.L. Bosch, and J. Kleissl. “Embedded nowcasting method using cloud speed persistence for a photovoltaic power plant”. In: *Solar Energy* 112 (Feb. 2015), pp. 232–238. DOI: `10.1016/j.solener.2014.11.013`.

[71] Vincent P.A. Lonij et al. “Intra-hour forecasts of solar power production using measurements from a network of irradiance sensors”. In: *Solar Energy* 97 (Nov. 2013), pp. 58–66. DOI: `10.1016/j.solener.2013.08.002`.

[72] Hai-xiang Zhao and Frederic Magoules. “A review on the prediction of building energy consumption”. In: *Renewable and Sustainable Energy Reviews* 16.6 (Aug. 2012), pp. 3586–3592. DOI: `10.1016/j.rser.2012.02.049`.

[73] Sobrina Sobri, Sam Koohi-Kamali, and Nasrudin Abd. Rahim. “Solar photovoltaic generation forecasting methods: A review”. In: *Energy Conversion and Management* 156 (Jan. 2018), pp. 459–497. DOI: `10.1016/j.enconman.2017.11.019`.

[74] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research 12* (2012), pp. 2825–2830.

[75] Simon O. Haykin. *Neural Networks and Learning Machines (3rd Edition)*. Pearson, 2008. ISBN: 0-13-147139-2. URL: `https://www.amazon.com/Neural-Networks-Learning-Machines-3rd/dp/0131471392?SubscriptionId=AKIAIOBINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0131471392`.

[76] Peter Palensky and Dietmar Dietrich. “Demand Side Management: Demand Response, Intelligent Energy Systems, and Smart Loads”. In: *IEEE Transactions on Industrial Informatics* 7.3 (Aug. 2011), pp. 381–388. DOI: `10.1109/tii.2011.2158841`.

[77] Olivier Corradi et al. “Controlling Electricity Consumption by Forecasting its Response to Varying Prices”. In: *IEEE Transactions on Power Systems* 28.1 (Feb. 2013), pp. 421–429. DOI: `10.1109/tpwrs.2012.2197027`.

[78] Gianluca Dorini, Pierre Pinson, and Henrik Madsen. “Chance-Constrained Optimization of Demand Response to Price Signals”. In: *IEEE Transactions on Smart Grid* 4.4 (Dec. 2013), pp. 2072–2080. DOI: `10.1109/tsg.2013.2258412`.

[79] *Winter package.* Tech. rep. European Union, 2017 [accessed: 2019-02-28]. URL: `http://eur-lex.europa.eu`.

[80] Y. Ozturk et al. “An Intelligent Home Energy Management System to Improve Demand Response”. In: *IEEE Transactions on Smart Grid* 4.2 (June 2013), pp. 694–701. DOI: 10.1109/tsg.2012.2235088.

[81] Nandkishor Kinhekar, Narayana Prasad Padhy, and Hari Om Gupta. “Multiobjective demand side management solutions for utilities with peak demand deficit”. In: *International Journal of Electrical Power & Energy Systems* 55 (Feb. 2014), pp. 612–619. DOI: 10.1016/j.ijepes.2013.10.011.

[82] Ditiro Setlhaolo, Xiaohua Xia, and Jiangfeng Zhang. “Optimal scheduling of household appliances for demand response”. In: *Electric Power Systems Research* 116 (Nov. 2014), pp. 24–28. DOI: 10.1016/j.epsr.2014.04.012.

[83] Hans Christian Gils. “Assessment of the theoretical demand response potential in Europe”. In: *Energy* 67 (Apr. 2014), pp. 1–18. DOI: 10.1016/j.energy.2014.02.019.

[84] Pei-Hao Li and Steve Pye. “Assessing the benefits of demand-side flexibility in residential and transport sectors from an integrated energy systems perspective”. In: *Applied Energy* 228 (Oct. 2018), pp. 965–979. DOI: 10.1016/j.apenergy.2018.06.153.

[85] Francisco Diaz-González, Andreas Sumper, and Oriol Gomis-Bellmunt, eds. *Energy Storage in Power Systems*. John Wiley & Sons, Ltd, May 2016. DOI: 10.1002/9781118971291.

[86] Garrett Fitzgerald et al. *The economics of battery energy storage: how multi-use, customer-sited batteries deliver the most services and value to the customer and the grid*. Tech. rep. Rocky Mountain Institute, 2015.

[87] Shalinee Kishore and Lawrence V. Snyder. “Control Mechanisms for Residential Electricity Demand in SmartGrids”. In: *2010 First IEEE International Conference on Smart Grid Communications*. IEEE, Oct. 2010. DOI: 10.1109/smartgrid.2010.5622084.

[88] Unchittha Prasatsap, Suwit Kiravittaya, and Jirawadee Polprasert. “Determination of Optimal Energy Storage System for Peak Shaving to Reduce Electricity Cost in a University”. In: *Energy Procedia* 138 (Oct. 2017), pp. 967–972. DOI: 10.1016/j.egypro.2017.10.091.

[89] Guido Lorenzi and Carlos Augusto Santos Silva. “Comparing demand response and battery storage to optimize self-consumption in PV systems”. In: *Applied Energy* 180 (Oct. 2016), pp. 524–535. DOI: 10.1016/j.apenergy.2016.07.103.

[90] Kyeong-Hee, Cho Seul-Ki Kim, and Eung-Sang Kim. “Optimal Sizing of BESS for Customer Demand Management”. In: *Energy and Buildings, 2014* (2014).

[91] Zhenghua Chen, Chaoyang Jiang, and Lihua Xie. “Building occupancy estimation and detection: A review”. In: *Energy and Buildings* 169 (June 2018), pp. 260–270. DOI: 10.1016/j.enbuild.2018.03.084.

[92] Claudio Martani et al. “ENERNET: Studying the dynamic relationship between building occupancy and energy consumption”. In: *Energy and Buildings* 47 (Apr. 2012), pp. 584–591. DOI: 10.1016/j.enbuild.2011.12.037.

[93] Christoph Bergmeir and José M. Benitez. “On the use of cross-validation for time series predictor evaluation”. In: *Information Sciences* 191 (May 2012), pp. 192–213. DOI: 10.1016/j.ins.2011.12.028.

[94] Allen B. Downey. *Python for Software Design: How to Think Like a Computer Scientist*. Cambridge University Press, Mar. 16, 2009. 251 pp. ISBN: 0521725968. URL: `https://www.ebook.de/de/product/8217121/allen_b_downey_python_for_software_design_how_to_think_like_a_computer_scientist.html`.

[95] William E. Hart et al. *Pyomo — Optimization Modeling in Python*. Springer International Publishing, 2017. DOI: `10.1007/978-3-319-58821-6`.

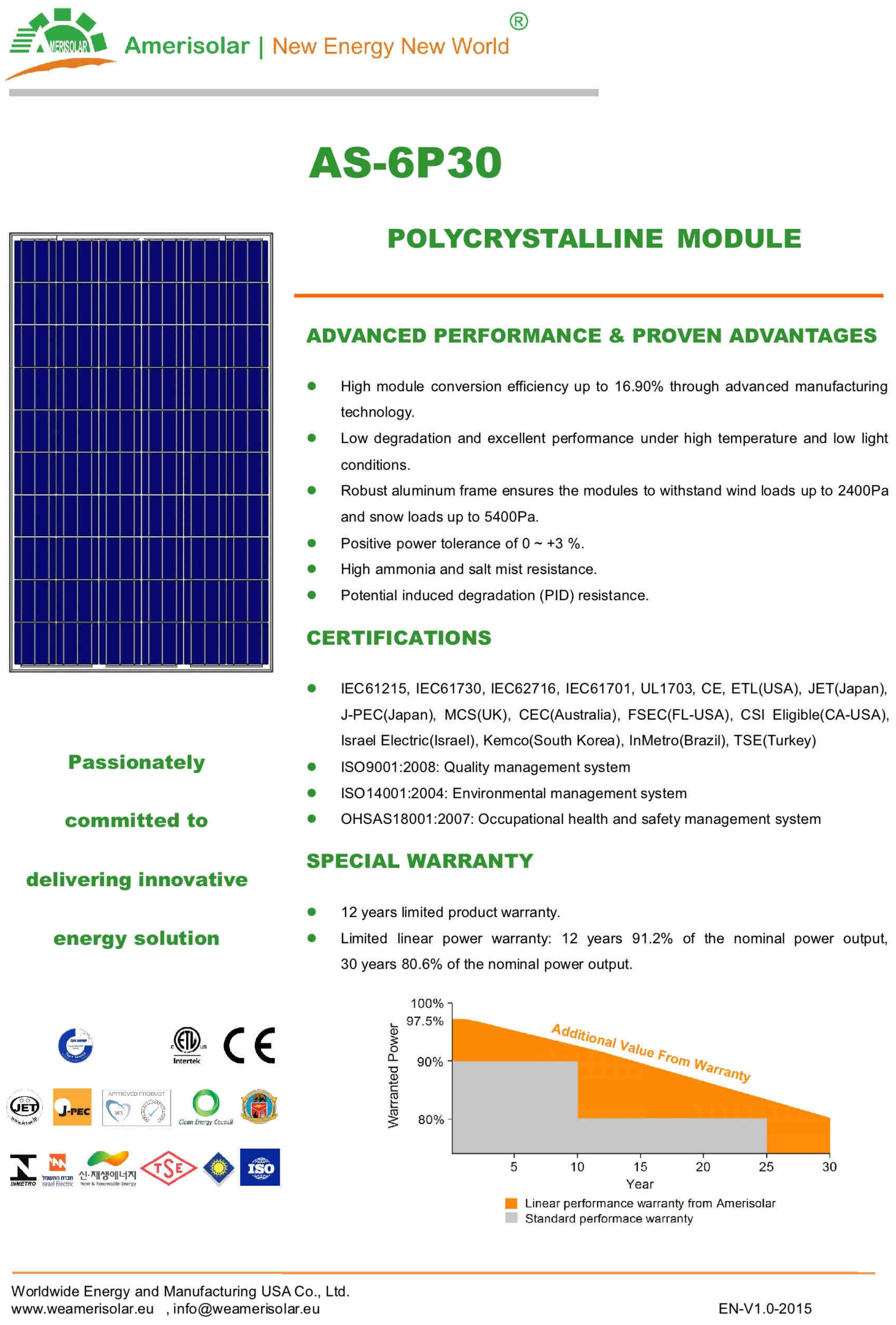
[96] Entidade Reguladora dos Servicos Energeticos ESRE. *Estrutura Tarifaria do Setor Eletrico em 2019.* Tech. rep. ESRE, 2019.

[97] International Renewable Energy Agency IRENA. *Electricity Storage and Renewables: Costs and Market to 2030*. Tech. rep. IRENA, 2017.

[98] T Randall. *Tesla's Battery Revolution Just Reached Critical Mass*. Tech. rep. available at https://www.bloomberg.com/ news/articles/2017-01-30/tesla-s-battery-revolution-just-reached-critical-mass, accessed on 28/2/2019, 2017.

Appendix

Appendix A: PV Panel Datasheet



Amerisolar | New Energy New World ®

AS-6P30

POLYCRYSTALLINE MODULE

ADVANCED PERFORMANCE & PROVEN ADVANTAGES

- High module conversion efficiency up to 16.90% through advanced manufacturing technology.
- Low degradation and excellent performance under high temperature and low light conditions.
- Robust aluminum frame ensures the modules to withstand wind loads up to 2400Pa and snow loads up to 5400Pa.
- Positive power tolerance of 0 ~ +3 %.
- High ammonia and salt mist resistance.
- Potential induced degradation (PID) resistance.

CERTIFICATIONS

- IEC61215, IEC61730, IEC62716, IEC61701, UL1703, CE, ETL(USA), JET(Japan), J-PEC(Japan), MCS(UK), CEC(Australia), FSEC(FL-USA), CSI Eligible(CA-USA), Israel Electric(Israel), Kemco(South Korea), InMetro(Brazil), TSE(Turkey)
- ISO9001:2008: Quality management system
- ISO14001:2004: Environmental management system
- OHSAS18001:2007: Occupational health and safety management system

Passionately

committed to

delivering innovative

energy solution

SPECIAL WARRANTY

- 12 years limited product warranty.
- Limited linear power warranty: 12 years 91.2% of the nominal power output, 30 years 80.6% of the nominal power output.

100%
97.5%
90%
80%
Warranted Power
Additional Value From Warranty
5
10
15
20
25
30
Year
Linear performance warranty from Amerisolar
Standard performace warranty

Worldwide Energy and Manufacturing USA Co., Ltd.
www.weamerisolar.eu , info@weamerisolar.eu

EN-V1.0-2015

ELECTRICAL CHARACTERISTICS AT STC

Nominal Power (P_{max})	240W	245W	250W	255W	260W	265W	270W	275W
Open Circuit Voltage (V_{OC})	37.7V	37.9V	38.0V	38.1V	38.2V	38.3V	38.4V	38.5V
Short Circuit Current (I_{SC})	8.57A	8.66A	8.75A	8.83A	8.90A	8.98A	9.06A	9.15A
Voltage at Nominal Power (V_{mp})	29.9V	30.1V	30.3V	30.5V	30.7V	30.9V	31.1V	31.3V
Current at Nominal Power (I_{mp})	8.03A	8.14A	8.26A	8.37A	8.47A	8.58A	8.69A	8.79A
Module Efficiency (%)	14.75	15.06	15.37	15.67	15.98	16.29	16.60	16.90
Operating Temperature	-40°C to +85°C							
Maximum System Voltage	1000V DC							
Fire Resistance Rating	Type 1(UL1703)/Class C(IEC61730)							
Maximum Series Fuse Rating	15A							

STC: Irradiance 1000W/m^2, Cell temperature 25°C, AM1.5

ELECTRICAL CHARACTERISTICS AT NOCT

Nominal Power (P_{max})	177W	180W	184W	188W	191W	195W	199W	202W
Open Circuit Voltage (V_{OC})	34.7V	34.9V	35.0V	35.1V	35.2V	35.3V	35.4V	35.5V
Short Circuit Current (I_{SC})	6.94A	7.01A	7.09A	7.15A	7.21A	7.27A	7.34A	7.41A
Voltage at Nominal Power (V_{mp})	27.2V	27.4V	27.6V	27.8V	27.9V	28.1V	28.3V	28.5V
Current at Nominal Power (I_{mp})	6.51A	6.57A	6.67A	6.77A	6.85A	6.94A	7.04A	7.09A

NOCT: Irradiance 800W/m^2, Ambient temperature 20°C, Wind Speed 1 m/s

MECHANICAL CHARACTERISTICS

Cell type	Polycrystalline 156x156mm (6x6inches)
Number of cells	60 (6x10)
Module dimensions	1640x992x40mm (64.57x39.06x1.57inches)
Weight	18.5kg (40.8lbs)
Front cover	3.2mm (0.13inches) low-iron tempered glass
Frame	Anodized aluminum alloy
Junction box	IP67, 3 diodes
Cable	4mm^2 (0.006inches^2), 900mm (35.43inches)
Connector	MC4 or MC4 compatible

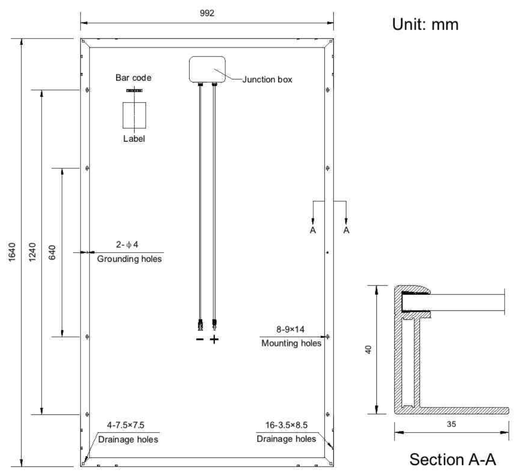
TEMPERATURE CHARACTERISTICS

Nominal Operating Cell Temperature (NOCT)	45°C±2°C
Temperature Coefficients of P_{max}	-0.43%/°C
Temperature Coefficients of V_{OC}	-0.33%/°C
Temperature Coefficients of I_{SC}	0.056%/°C

PACKAGING

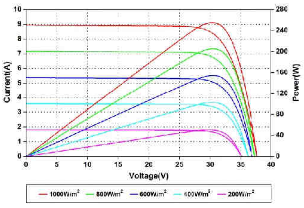
Standard packaging	26pcs/pallet
Module quantity per 20' container	312 pcs
Module quantity per 40' container	728 pcs

ENGINEERING DRAWINGS

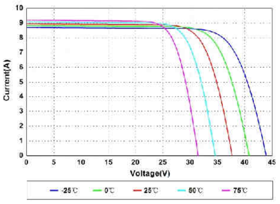


Specifications in this datasheet are subject to change without prior notice.

IV CURVES

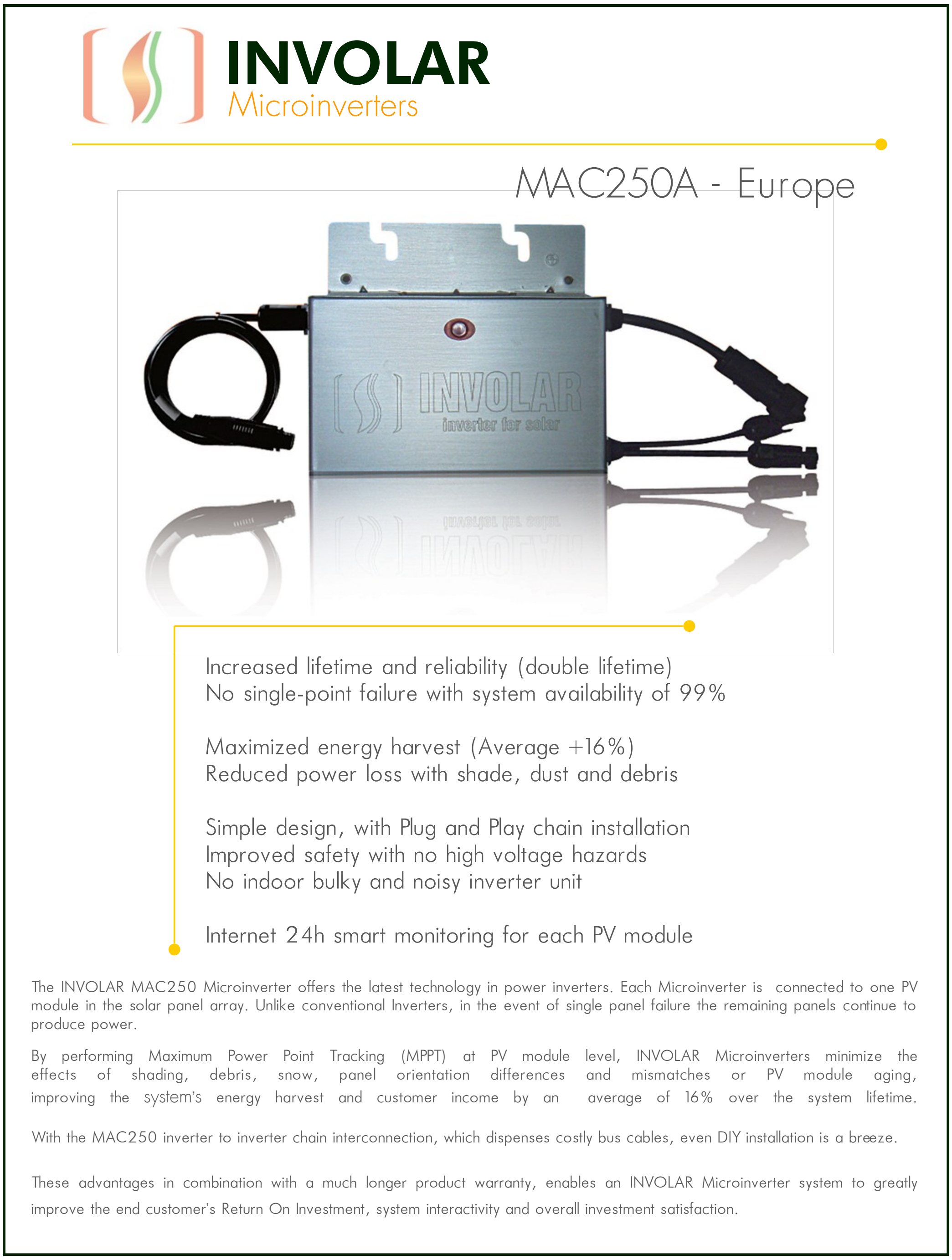


Current-Voltage and Power-Voltage Curves at Different Irradiances



Current-Voltage Curves at Different Temperatures

Appendix B: Inverter Datasheet



INVOLAR
Microinverters

MAC250A - Europe

Increased lifetime and reliability (double lifetime)
No single-point failure with system availability of 99%

Maximized energy harvest (Average +16%)
Reduced power loss with shade, dust and debris

Simple design, with Plug and Play chain installation
Improved safety with no high voltage hazards
No indoor bulky and noisy inverter unit

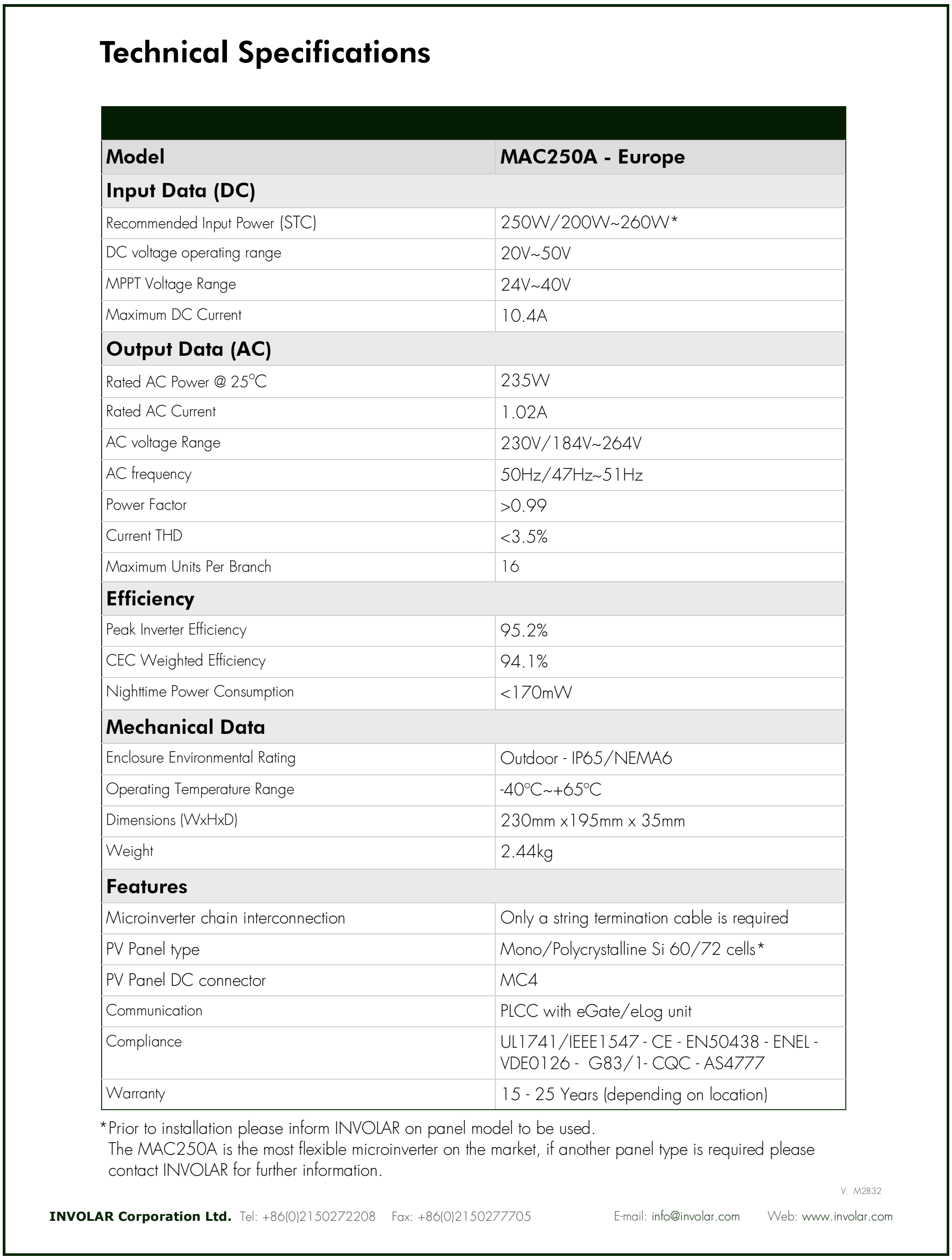
Internet 24h smart monitoring for each PV module

The INVOLAR MAC250 Microinverter offers the latest technology in power inverters. Each Microinverter is connected to one PV module in the solar panel array. Unlike conventional Inverters, in the event of single panel failure the remaining panels continue to produce power.

By performing Maximum Power Point Tracking (MPPT) at PV module level, INVOLAR Microinverters minimize the effects of shading, debris, snow, panel orientation differences and mismatches or PV module aging, improving the system's energy harvest and customer income by an average of 16% over the system lifetime.

With the MAC250 inverter to inverter chain interconnection, which dispenses costly bus cables, even DIY installation is a breeze.

These advantages in combination with a much longer product warranty, enables an INVOLAR Microinverter system to greatly improve the end customer's Return On Investment, system interactivity and overall investment satisfaction.



Technical Specifications

Model	**MAC250A - Europe**
Input Data (DC)	
Recommended Input Power (STC)	250W/200W~260W*
DC voltage operating range	20V~50V
MPPT Voltage Range	24V~40V
Maximum DC Current	10.4A
Output Data (AC)	
Rated AC Power @ 25°C	235W
Rated AC Current	1.02A
AC voltage Range	230V/184V~264V
AC frequency	50Hz/47Hz~51Hz
Power Factor	>0.99
Current THD	<3.5%
Maximum Units Per Branch	16
Efficiency	
Peak Inverter Efficiency	95.2%
CEC Weighted Efficiency	94.1%
Nighttime Power Consumption	<170mW
Mechanical Data	
Enclosure Environmental Rating	Outdoor - IP65/NEMA6
Operating Temperature Range	-40°C~+65°C
Dimensions (WxHxD)	230mm x195mm x 35mm
Weight	2.44kg
Features	
Microinverter chain interconnection	Only a string termination cable is required
PV Panel type	Mono/Polycrystalline Si 60/72 cells*
PV Panel DC connector	MC4
Communication	PLCC with eGate/eLog unit
Compliance	UL1741/IEEE1547 - CE - EN50438 - ENEL - VDE0126 - G83/1- CQC - AS4777
Warranty	15 - 25 Years (depending on location)

*Prior to installation please inform INVOLAR on panel model to be used.
The MAC250A is the most flexible microinverter on the market, if another panel type is required please contact INVOLAR for further information.

V. M2832

INVOLAR Corporation Ltd. Tel: +86(0)2150272208 Fax: +86(0)2150277705 E-mail: info@involar.com Web: www.involar.com