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

Fabia Miorelli

E-Mail: fabia.miorelli@tecnico.ulisboa.pt

Instituto Superior Técnico, Universidade de Lisboa, Avenida Rovisco Pais 1, 1049-001 Lisbon, Portugal

ARTICLE INFO

Date:

July 2019

Keywords:

Energy Management, Demand-Response, Artificial Neural Networks, Machine Learning, Battery Energy Storage System.

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.

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. 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 depend 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.

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, the goal of machine learning is to build a computer system that automatically learns from the data, disregarding its internal working principles [3]. In machine learning present data can be used to predict future data, by learning the relationship between the features and the label.

Numerous works address electricity consumption forecasting in buildings, as it is of the utmost importance when it comes to reducing energy consumption and reaching predefined energy targets. 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 [4].

Wei [5] 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, either alone or more often together with other optimization techniques, or using deep recurrent neural networks [6], [7], [8], [9], [10]. 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 linear regression, 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 [11], [12], [13].

When it comes to decentralized generation, such as solar power production, its forecasting 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 all 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 meteorological 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.

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 [14], [15], [16], [17]. 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 [18]. Also in the case of solar power production, the prevalent machine learning methods seem to be ANNs and SVMs [19], [20] but in some cases, simpler models, such as k-Nearest Neighbours (kNN) or Gradient Boosted Regression Trees (GBRT), also proved to accurate [17], [18].

Being able to predict building electricity consumption and decentralized solar power generation allows to assess energy management strategies and flexibility options that could potentially reduce electricity consumption and lead to economic savings. Flexible buildings are indeed an example of prosumer that could help the transition to a low carbon energy system. Junker identifies several factors influence the building's ability to provide energy flexibility [20], 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 CO_2 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 [21], 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 [22]. 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 [23], which consists of a change in the consumption pattern of a customer which can either reduce or shift its peak consumption to off-peak hours. Following a drop in their prices, electro-chemical batteries, especially 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 demand-response to perform load shifting at customer level. 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 [24]. Multiple approaches to determine the optimal capacity of battery energy storage system (BESS) for peak shaving to reduce a building's annual energy costs have been proposed [25], [26], [27] and all show that a significant reduction can be achieved, although the compensation of the upfront investment is not always granted and is highly dependent on the battery size and the achievable profit.

Optimal energy consumption schemes based on Time-of-Use 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.

The aim of this paper is to compare different 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 the civil building on the campus is simulated and analysed by means of real data. An evaluation of possible demand-response strategies exploiting time-of-use energy tariffs and 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.

The present paper is divided as follows: Section 2 describes the basic concepts behind the tested machine learning forecasting algorithms whereas the methodology used to develop the prediction models and the demand-response model is described in Section 3. The results are presented in Section 4 with the consequent conclusions addressed in Section 5.

2. Background Concepts

In this work, four different low-complexity algorithms are tested to forecast building energy consumption and solar power production, namely multiple linear regression, k-Nearest Neighbours, decision trees and artificial neural networks, being these among the most common ones found in literature which combined a relatively low complexity and a good prediction accuracy. The models are developed in Python using the Scikit-Learn package, which 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 [28]. These models are applied to the civil building of Instituto Superior Técnico to forecast its electricity consumption and rooftop solar power production based on real data.

2.1 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, and others). 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:

$$y = mx + b$$

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:

$$y = b_0 + m_1 b_1 + m_2 b_2 + \cdots + m_n b_n \mathrm{y}$$

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):

$$\min \left(\sum_{i=1}^{n} (y_i - \bar{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2\right)$$

2.2 Decision Tree

A decision tree is a non-parametric model governed 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. 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 K-Nearest Neighbours

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 [28]. 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:

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

where y_i is the nearest neighbour. The parameter k is a very sensitive parameters that control the fit of the algorithm. 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.

2.4 Artificial Neural Networks

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, which takes several inputs, multiplies them by specific weights to produce an output.

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 [29]. When an ANN presents multiple layers it forms a multilayer perceptron (MLP).

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 numbers 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.

3. Methodology

This section describes the case-study building, the available input data and the modelling process used for the forecasting and for the development of the demand-response assessment.

3.1 Analysed Building

The civil building is one of the main buildings on the Alameda campus, with a total area of about 25'152 m^2 and is composed by seven floors, three below ground and four upper floors. 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 semester 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.

3.2 Available Data

The aforementioned models and their development process are introduced in greater detail in this chapter and form the different steps to be able to assess the profitability potential of the suggested energy management strategy. 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 and graphically presented in Figure 1.

3.2.1 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\degree N, 9.138\degree W, 90 m above sea level) and is available with a 5-minute resolution.

3.2.2 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 [30]. 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 [31], 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 OpenSource 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 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.

3.2.3 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 [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.

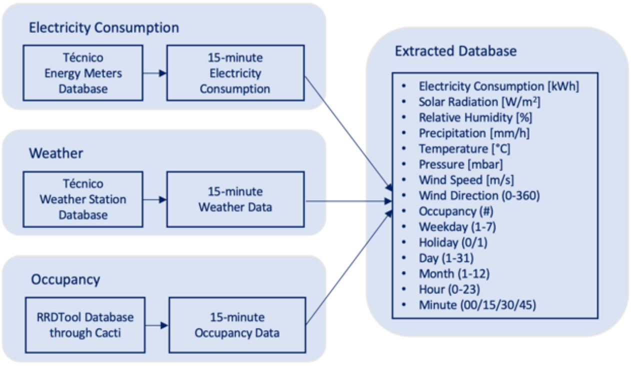


Figure 1: Data extraction and integration process to create the database for the development of energy consumption and production prediction models with a 15-minute sampling rate.

3.3 Forecasting 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, choosing an evaluation metric and a testing procedure, identifying the important features, developing the model and tuning its hyper-parameters, evaluating the model performance and proceeding 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. Figure 2 summarises the model development and evaluation process.

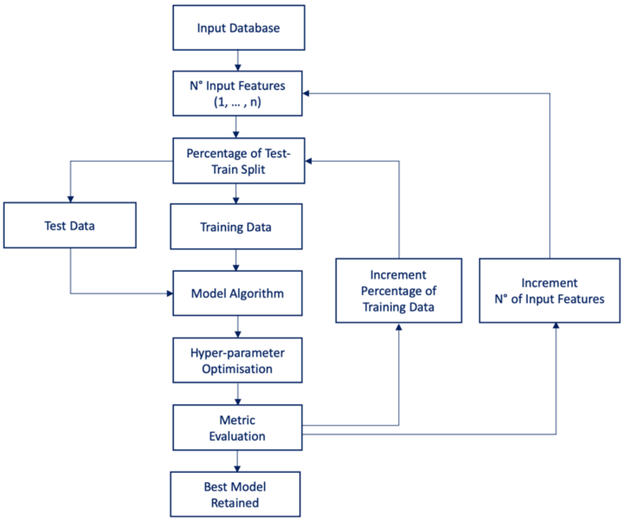

Input Database
N° Input Features (1, … , n)
Percentage of Test-Train Split
Test Data
Training Data
Model Algorithm
Hyper-parameter Optimisation
Metric Evaluation
Best Model Retained
Increment Percentage of Training Data
Increment N° of Input Features

Figure 2: Model development and evaluation procedure. At the end of each simulation, the MAE, RMSE and coefficient of determination are computed.

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 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 [32].

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 feature for the different algorithms a forward stepwise method was chosen. This procedure included starting the model with very few predictors and iteratively assess 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 affect model performance and it requires hard work to be found. Empirical approaches have to be used as no universal procedure exists [33], while tuning the hyper-parameters care needs to be taken not to overfit or underfit the model.

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).

3.4. Demand-Response

To carry out the assessment of the economic feasibility of the implementation of a battery energy storage system to perform demand-response, a behind-the-meter battery is optimally sized 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 the lifetime of the system considering current market-based battery prices. Lastly, the economic advantages of the chosen battery are assessed taking into consideration the consumption and PV production forecasts.

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. In particular, the open-source software package Pyomo was employed, which possesses broad optimization capabilities to formulate, solve, and analyse optimization models [34]. Pyomo allows fairly easy formulations to define the objective function of the model, the constraints, the decision variable 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.

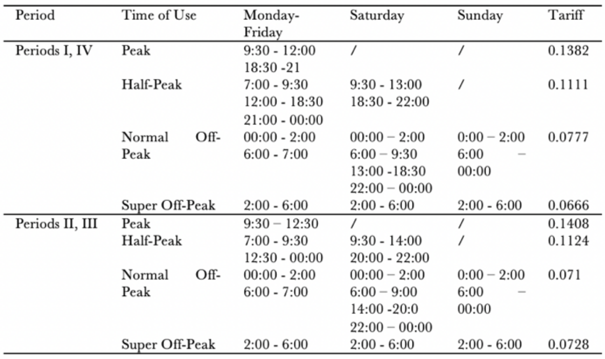
This method was chosen in place or more complex ones, because it accommodates rapid cost-optimal battery sizing and has the advantage of being easily implemented in devices with low computational power.

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 generation data.

The considered load profile considered the year 2018 for the analysed building. 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 12 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. 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 1 and it includes VAT at the present level in Portugal (23%) [35]. 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 its production peaks always occur in the peak hours within the tariff scheme. The major of the electricity consumption is though still concentrated in the peak hours of the weekdays.

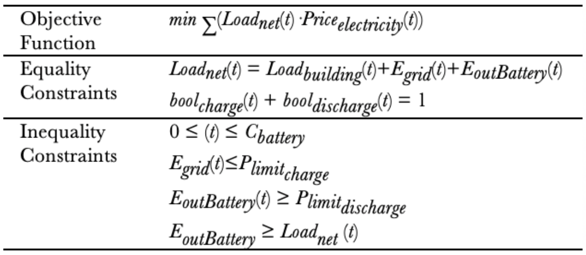
Based on the active energy tariff, first the overall electricity costs are calculated and then compared to the costs which could be achieved in the presence of an optimally scheduled battery. Table 2 shows the implemented formulas in the MILP optimisation procedure.

Table 1: EDP MV time-varying tariff.



Period	Time of Use	Monday-Friday	Saturday	Sunday	Tariff
Periods I, IV	Peak	9:30 - 12:00 18:30 -21	/	/	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	0:00 – 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	9:30 - 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 -20:0 22:00 – 00:00	0:00 – 2:00 6:00 – 00:00	0.071
	Super Off-Peak	2:00 - 6:00	2:00 - 6:00	2:00 - 6:00	0.0728

Table 2: MILP optimization formulation



Objective Function	$min \sum(Load_{net}(t) \cdot Price_{electricity}(t))$
Equality Constraints	$Load_{net}(t) = Load_{building}(t)+E_{grid}(t)+E_{outBattery}(t)$ $bool_{charge}(t) + bool_{discharge}(t) = 1$
Inequality Constraints	$0 \leq (t) \leq C_{battery}$ $E_{grid}(t)\leq P_{limit_{charge}}$ $E_{outBattery}(t) \geq P_{limit_{discharge}}$ $E_{outBattery} \geq Load_{net}\ (t)$

4 Results

This section presents the most accurate prediction models developed. The impact of different relevant input features to predict electricity consumption and generation is assessed, enhancing those which contribute to achieving accurate models. Finally, the results of the optimum sizing of a BESS system to perform demand-response is presented.

4.1 Electricity Consumption Forecasting

Additionally to the available data presented in Figure 1, a variable representing the academic calendar and the Portuguese national holiday was added (named holiday) to the electricity consumption forecasting database, 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. Another feature used as input to the model is an auto-regressive feature representing the consumption of the 15 minutes and it is labelled as 'Energy-1' throughout this paper.

4.1.1. Correlation Between Input Variables

Figure 3 shows scattered plots showing 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, it was noticed 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 trend of the solar radiation and the consumption a direct proportional relationship between these two variables becomes evident it 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 the consumption but, at the same time, should be independent from each other.

4.1.2 Forecasting Results

Comparing the four best model developed and plotting the forecast values against actual values (Figure 5), it can be seen that all algorithms tend to forecast values that in general are lower than the actual values, especially the peaks of consumption during the working days and 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 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 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 shows the results for the best preforming ANN without the autoregressive feature (on the left) and with the autoregressive feature (on the right) and Table 3 summarises the performance of all the tested algorithms.

Figure 5 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 slight better performance than the kNN model and it can be seen that the points are denser 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.

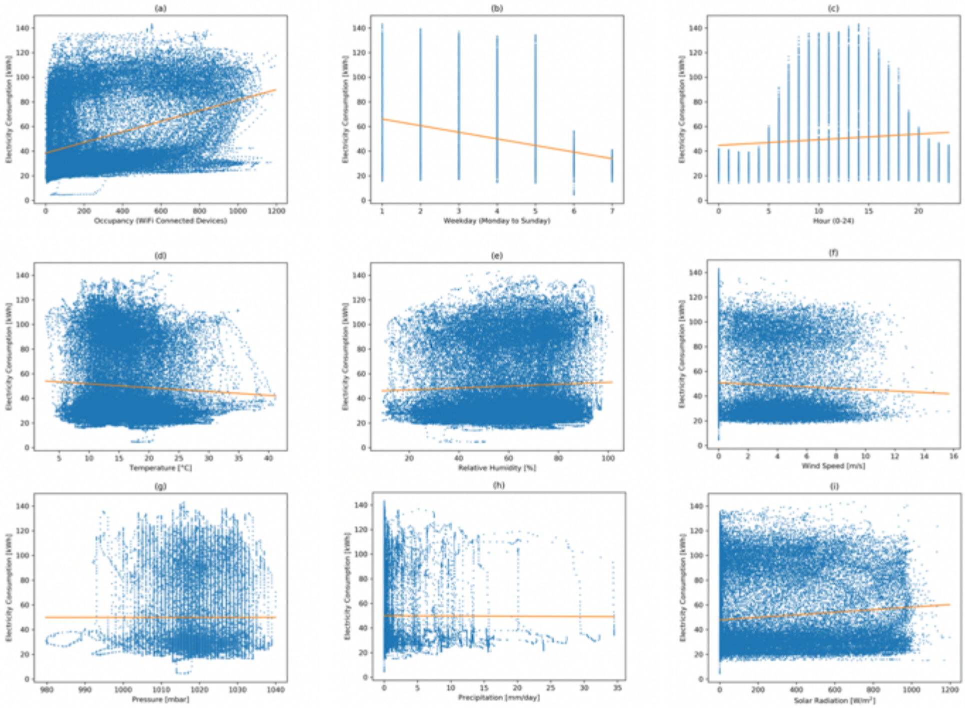


Figure 3: Scattered plots showing the correlation between energy consumption and (a) occupancy, (b) day of the week, (c) hour of the day, (d) temperature, (e) relative humidity, (f) wind speed, (g) pressure, (h) precipitation and (i) solar radiation.

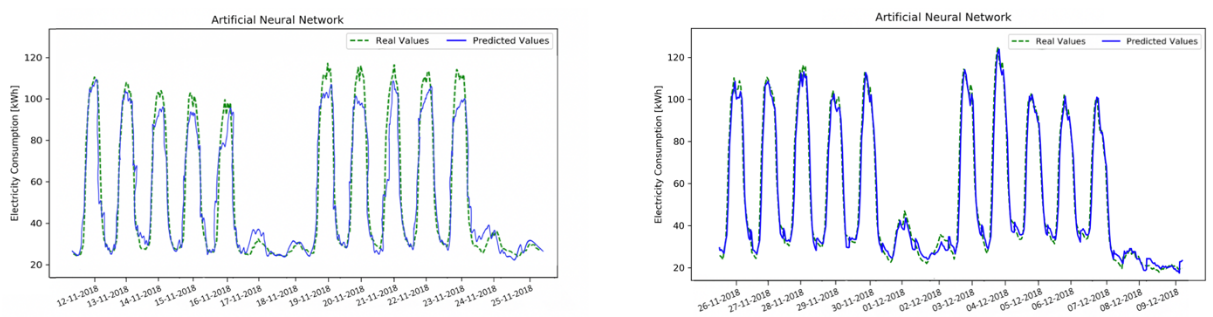


Figure 4: Artificial neural network model versus real consumption data.

Table 3: MAE, RMSE and coefficient of determination of the best models tested for the electricity consumption model.

Algorithm	Weekday	Day	Month	Hour	Minute	Occupancy	Holiday	Radiation	Energy-1	MAE [kWh]	RMSE [kWh]	R²
Linear Regression	•	•	•	•	•	•	•	•	•	4.6	6.7	0.95
Decision Tree	•	•	•	•	•	•	•		•	3.4	4.8	0.97
kNN	•	•	•	•	•	•	•		•	3.5	4.9	0.96
ANN	•	•	•	•	•	•	•		•	1.4	2.9	0.97

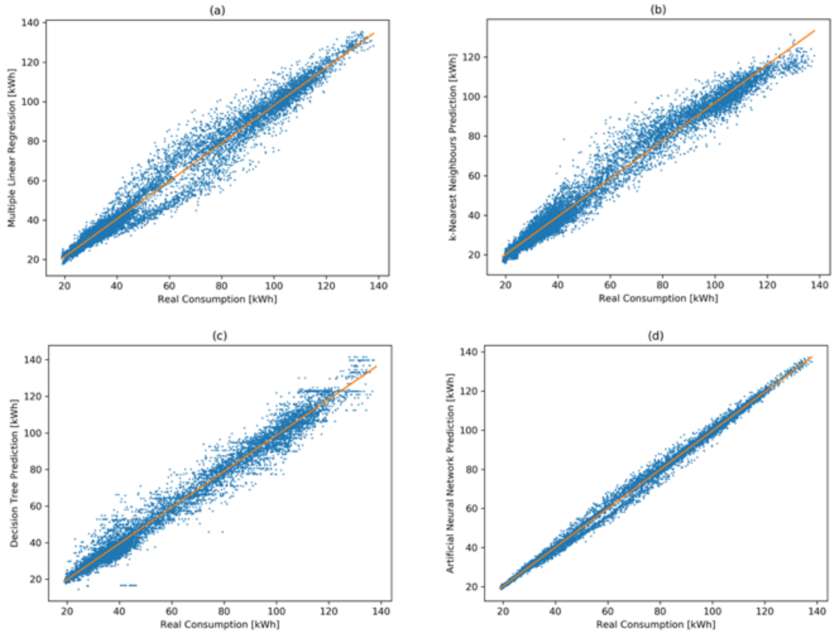


Figure 5: Comparison of predicted versus real consumption values for (a) multiple linear regression model, (b) k-nearest neighbours model, (c) decision tree model and (d) artificial neural network model including an autoregressive feature.

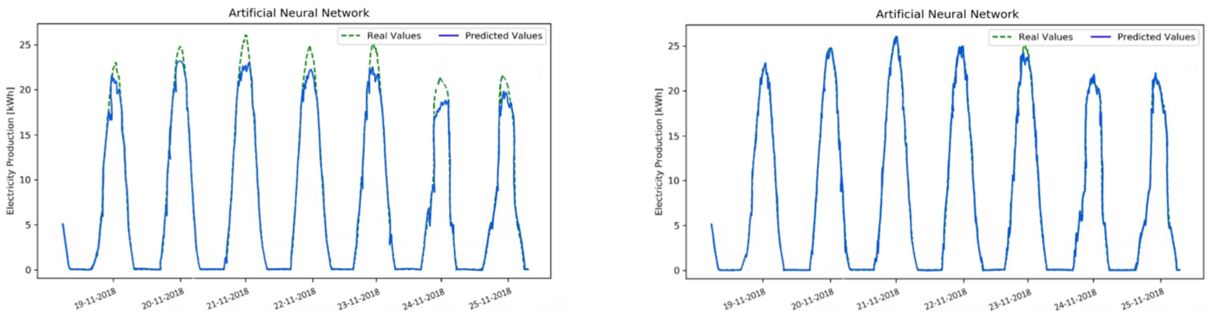


Figure 7: Artificial neural network model versus real consumption data.

Table 4: MAE, RMSE and coefficient of determination of the best models tested for the electricity production model.

Algorithm	Radiation	Temperature	Relative Humidity	Wind Speed	Hour	Month	Pressure	Energy-1	MAE [kWh]	RMSE [kWh]	R²
Linear Regression	•	•	•	•	•	•	•	•	0.9	2.0	0.61
Decision Tree	•	•	•	•	•	•	•	•	0.5	1.9	0.66
kNN	•	•	•	•	•	•	•	•	0.7	1.8	0.64
ANN	•	•	•	•	•	•	•	•	0.6	1.7	0.72

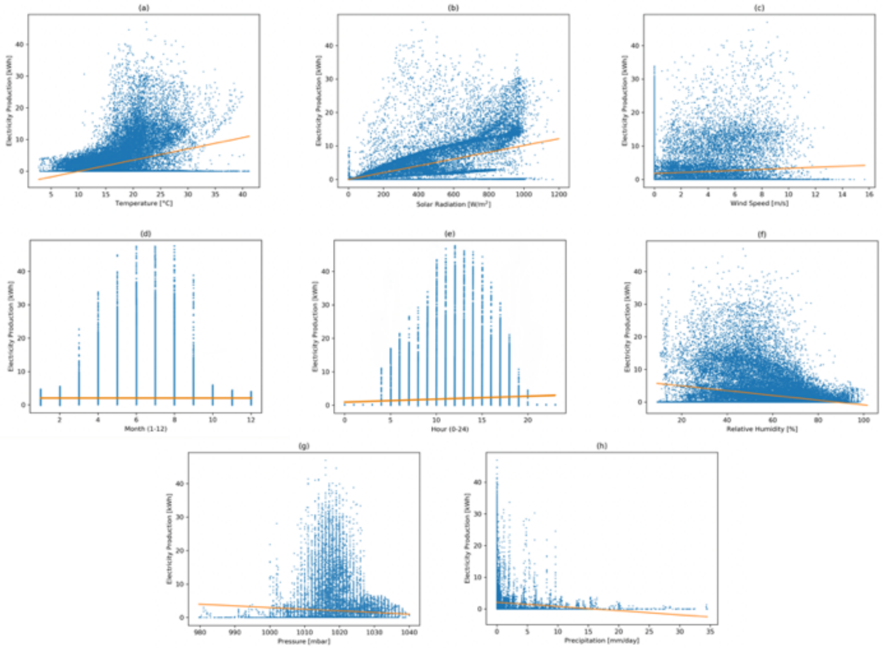


Figure 6: Scattered plots showing the correlation between energy production and (a) temperature, (b) solar radiation, (c) wind speed, (d) month, (e) hour, (f) relative humidity, (g) pressure and (h) precipitation.

4.2 Generation

The chosen forecasting approach for solar power production involved a direct forecasting method, 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 dependent 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

Figure 6 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 a mild correlation with temperature.

From the last scatter plot (plot h) it is immediate to identify that the power production increased with decreasing precipitation.

4.2.2 Forecasting Results

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

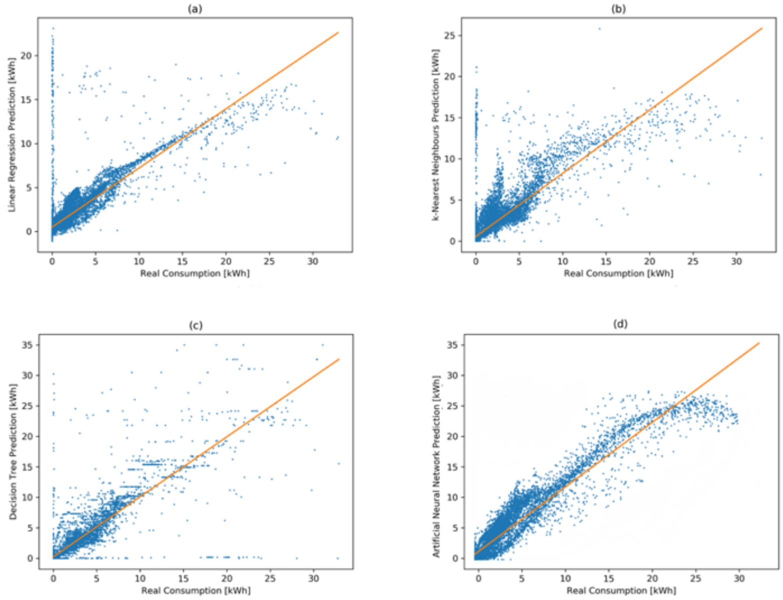


Figure 8: Comparison of predicted versus real production values for (a) multiple linear regression model, (b) k-nearest neighbours model, (c) decision tree model and (d) artificial neural network model including an autoregressive feature.

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 7 shows the results for the best performing ANN without the autoregressive feature (on the left) and with the autoregressive feature (on the right).

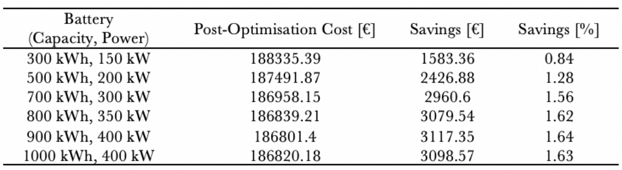
Figure 8, represents scatterplots between the prediction and the real values for the four tested models. The main takeaways are that the linear regression model is not even close to accurately model the dynamics of the solar power production and its forecasted values are much lower than the real ones. The kNN and the decision tree algorithm already perform better but it can be seen that for higher values in 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.

4.3 Demand-Response Results

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.

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 [36] and Bloomberg technology report [37] 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 5 and are calculated comparing the electricity bill for the given tariff with and without the battery.

Table 5: 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	2960.6	1.56
800 kWh, 350 kW	186839.21	3079.54	1.62
900 kWh, 400 kW	186801.4	3117.35	1.64
1000 kWh, 400 kW	186820.18	3098.57	1.63

From Table 5 it can be seen that increasing the capacity reduces the minimal achievable cost and therefore increases slightly the savings. However, after a certain threshold, no further improvement can be noticed, as the battery prices becomes too high compared with the possible savings. 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.

5 Conclusion

The goal of this article was to assess the performance of different forecasting data-driven models for both electricity consumption and power generation, including 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 variable that best describe the consumption patterns in the analysed buildings, that proved to be highly correlated with the occupancy data 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, and a RMSE of 2.9 kWh when including input features related to the occupancy, the academic calendar, the time and an auto-regressive feature. Weather variables did not contribute in 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, and a RMSE of 1.7. For this last model, the key input features were as expected the solar radiation, the temperature, the humidity, the wind speed, an autoregressive feature and features related to time, such as hour of the day and month. These features were expected as the most important parameters influencing solar power production are related to the time of the year and of the day, which influence the angles between the sun rays and the panel surface, and radiation and temperature.

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. 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 without a battery. 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.

References

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

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Federica Davo` 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Sebastian Stinner, Kristian Huchtemann, and Dirk Mu¨ller. "Quantifying the operational flexibility of building en- ergy systems with thermal energy storages". In: *Applied Energy* 181 (Nov. 2016), pp. 140–154. DOI: 10.1016/j. apenergy.2016.08.055.

[23] 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.

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

[25] Kyeong-Hee, Cho Seul-Ki Kim, and Eung-Sang Kim. "Optimal Sizing of BESS for Customer Demand Manage- ment". In: *Energy and Buildings, 2014* (2014).

[26] Unchittha Prasatsap, Suwit Kiravittaya, and Jirawadee Polprasert. "Determination of Optimal Energy Storage Sys- tem 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.

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

[32] Christoph Bergmeir and Jose´ 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.

[33] 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.

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

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

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

[37] 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.