Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Energetica e Nucleare

Application of Data Analytics techniques for the analysis of Building Energy Performance during operation: the case of Politecnico di Torino.

Relatore:
Prof. Alfonso Capozzoli
Correlatore:
Dott. Marco Savino Piscitelli

Candidata:
Maria Teresa Zitelli

Anno Accademico 2021/2022

Abstract

In 2019, building operations were responsible for about the 28% of the global CO_2 emissions[3], taking into account not only the share directly due to the daily activity, but also the indirect part produced by the generation of power that supplies the building. The building consumption is strongly affected by the *Energy Performance Gap*, which is the deviation of the actual energy performance of the building with respect to the expected and designed one. As a consequence, the scope of improvement is relevant for the buildings and, in this context, an effective energy management has a key role.

The aim of this thesis work is to provide a Data Analytics methodology whose results can be helpful to increase the knowledge about the system and that can be a tool to implement for the energy management. In particular, the methodology is applied to an educational building, the Polytechnic of Turin, focusing on a defined subsection of the system that includes energy-intensive loads and a photovoltaic production plant. The analysis follows two parallel paths, taking into account, first, the load-side and then the production-side of the domain. The load-level analysis identifies typical profiles of consumption - with correspondent external conditions - of a chiller unit and an independent building, by means of an hierarchical clustering technique and a classification tree. Then, the focus is on the baseload of each profile, intended as the minimum value of demand that is always present, in order to find reference power ranges that are used to define a Key Performance Indicator, that ranks the daily energy-related behaviour. At this point, the energy waste of the loads is detected with a comparison between the actual consumption and a simulated one, considering improved values of baseload power. The production-level analysis, instead, consists in the development of an Artificial Neural Network for the forecast of the power production of the photovoltaic plant; the results of the neural network are then used to develop an anomaly detection algorithm in order to automatically find faulty operating conditions of the system, providing a daily warning that distinguish between strong and possible anomalies. Finally, a predictive maintenance procedure is proposed with the aim to recommend extraordinary maintenance actions if a series of anomalous day are consequently reported.

Contents

List of Figures

List of Tables

List of Acronyms

PV	Photovoltaic
HVAC	Heating, Ventilation and Air Conditioning
KPI	Key Performance Indicator
ANN	Artificial Neural Network
ODM	One-Diode Model
SVM	Support Vector Machine
MPP	Maximum Power Point
EWMA	Exponentially Weighted Moving Average
PNN	Probabilistic Neural Network
DT	Decision Tree
CART	Classification and Regression Tree
RELU	Rectified Linear Unit
MV	Medium Voltage
LV	Low Voltage
KPI	Key Performance Indicator
MAPE	Mean Average Percentage Error
MAE	Mean Absolut Error
MSE	Mean Square Error
LL	Lower Limit
UL	Upper Limit
TRP	True Positive Rate
FPR	False Positive Rate
AUC	Area Under Curve

1 Introduction

On 13th November 2021, at the end of the XXVI United Nations Climate Change conference, the Parties signed the *Glasgow Climate Pact* with which they confirm the commitment, undertaken in 2015 with the *Paris Agreement* [2], to face the climate change, recognising the necessity to limit the increase in the global temperature to values below 1.5 °C with respect to pre-industrial level. In order to reach this goal, the document states the need for a *"rapid, deep and sustained reductions in global greenhouse gas emissions"*[1].

According to 2019 data, building operations cause the 28% of the global CO_2 emissions [3], taking into account not only the share directly due to the daily activity, but also the indirect part produced by the generation of the power that supplies the building. In particular, this 28% includes both residential and not-residential buildings that are responsible for the 11% of the global emissions [3]. These percentages are direct consequences of the amount of energy that serves this sector: residential and not-residential buildings represent the 30% of the global final energy use [3].

Shifting the attention to a limited context and focusing on a specific Country, in Italy the total electricity consumption in 2019 was about 292 TWh [4]: 66 was for residential buildings and about 90 TWh supplied only commercial and public services.

From these data, it is clear the strong impact that buildings have on the final energy consumption of a Country and this suggests a great opportunity of improvement; in fact, another aspect that has to be taken into account, and that strongly affects the consumption of the residential sector, is the *Energy Performance Gap* that is the deviation of the energy performance of the building from the expected one: it is the difference between the behaviour of the building and its actual energy use with respect to the project, with consequent worsening of the estimated consumption.

Despite the importance of this topic, the interest among the scientific community and the possible benefits due to its analysis, the magnitude of the gap is not estimated in a precise and consistent way; the main reason behind this poor evaluation, in addition to the choice of the sample of buildings, is the fact that the expected energy con-

sumption can be calculated by means of different methods and models, affecting the resulting value of the difference with the actual consumption and its comparability. However, it results that the Energy Performance Gap assumes lower values for households, with respect to the case of not-residential buildings [5]: schools, universities, public and commercial buildings show a bigger deviation of the actual consumption from the designed one. For this type of buildings, the Energy Performance Gap can reach values from 22% to 156% for universities, from 37% to 117% for schools and from 30% to 93% for offices [6].

The Energy Performance Gap can be considered connected to a variety of factors, different from each other for cause, predictability and solution. First of all, the origins of the gap can be technical: malfunctions and degradation of the devices, or an incorrect set-point, reduce the overall energy performance of the building, together with a wrong sizing of the system and errors in modelling and simulation [5]. However, the elements affecting the energy behaviour of the buildings are not only linked to the design of the system, but they can also be connected to human-related factors: the Energy Performance Gap can be due also to issues related to the behaviour of the occupants, their attitude and their comfort [5].

The significant role of the buildings in the total energy consumption of a Country - and consequent emissions - and the Energy Performance Gap, as a diffused and relevant issue, suggest the existence of a considerable scope for improvement in the energy performance of the buildings. In facts, the opportunities of energy savings can involve both technical features of the systems and aspects related to control, operation and occupant's habits.

In this context, the energy management of the whole building has a key role: the possible actions have to be identified, implemented and periodically updated. The basis of an advanced and effective management of the energy consumption is represented by a monitoring system: a set of meters and sensors that is able to provide measured information about the consumption of the building, at different levels of aggregation, from a single device to the overall system. The general idea of an effective process is the possibility to measure the actual consumption of a building and its subsystems,

identify the wastes, elaborate a valid strategy for energy saving and optimization and check the effectiveness of the actions by a continue monitoring of the energy-related parameters.

Nowadays, the monitoring system of the buildings, especially the non-residential ones, is able to provide sets of data that are extensive and detailed, in terms of both quantity in a time range and variety of measures. In fact, the consumption data that can be measured in a building are different and more or less aggregated: electricity supply of the overall buildings or of delimited areas, consumption of a single subsystem such as the HVAC or the lighting one, the production of on-site power generation (e.g. rooftop PV plant or heat pumps) and also characteristics of the indoor air such as temperature and humidity. Together with these building-related data, it is possible to easily access to meteorological information specifically referred to a certain location. Another point that has to be highlighted about the available data, is the possibility to reach a great volume of measurements in terms of time aggregation: quantities can be monitored periodically with time steps up to the order of minutes.

The result is the availability of extensive and heterogeneous sets of data, characterised by a large volume and a great variety of information and that can be obtained in real time. It follows that a data sets with such characteristics can be quite difficult to handle; in fact, quantities have to be collected, organised and correlated with each others to find useful information in order to take advantage from them for a correct management process: the goal is to extract knowledge from data and use them to model and analyse the energy behaviour of the building. This aim cannot be pursued by means of only conventional methods such as statistical and physics-based models: it is necessary an approach that is able to deal with a wide-ranging operational data [9]. The solution can be found in the application of Data Analytics that, as an alternative to direct models, allow a data-driven approach for constructing an effective energy modelling of the building and, in general, of a system. Data Analytics consists in a series of techniques and tools that are used to explore large sets of heterogeneous data in order to extract knowledge from them: it allows an effective description of the system by means, for example, of the identification of typical patterns and association

rules among quantities. Moreover, it can be used to identify correlations and construct prediction models.

In general, Data Analytics offers the opportunity to enhance the energy management of a building with a continuous monitoring and feedback, with consequent possible updating of the actions, with an investment that is very low because it does not require particular hardware, except for the metering system that is already installed in the majority of structures.

With specific regard to the energy sector and in particular to the energy modelling and analysis of buildings, Data Analytics finds a great variety of applications. It can be used to describe the demand in terms of typical load profiles and patterns that can be helpful in making considerations about the energy wastes and their causes, in order to develop and implement strategies of energy savings. Moreover, another application can be found in the field of building retrofit: Data Analytics techniques, such as clustering analysis combined with decision trees, can be used to identify potential buildings for retrofit actions by comparing them to a benchmark one [10]. Data Analytics techniques can be also adopted to develop forecast models that are able to predict the energy demand of a building or one of its subsystems (e.g. HVAC systems) [7] but also the production of a power generation system (e.g. on-site PV plant), as a function of external quantities such as meteorological information or indoor air characteristics. This kind of application can be helpful to predict the peaks of demand and analyse influencing factors, but it can also represent the first step to a deeper analysis that includes the detection and diagnosis of faults and anomalies in the production or consumption of a system. In fact, the predicted profile can be considered as a normal behaviour of the system and it can be used as term of comparison with actual data, in order to identify not acceptable values or patterns due to malfunctions of the system of interest. The anomaly detection related to the energy consumption can be adopted to identify not only anomalous behaviour or faulty appliances, but also non-technical loss (e.g. malfunction of the monitoring devices) and occupancy [8].

It is important to highlight the fact that even if Data Analytics is a strong tool that can really enhance the management and the decision-making about energy-related is-

sues of a building, it is necessary the knowledge of an expert about the system and the related problems, in order to be able to understand the results, evaluate their coherence and wisely use them.

In this context, the present thesis work has the aim to develop a methodology that can be considered as a contribution in a general approach of advanced energy management of a building and it can helps in the identification of possibilities for performance improvement. In particular, the analysis regards an academic building, the Polytechnic of Turin, whose management is not so trivial because it consists of spaces with different purposes, generation plants and energy-intensive loads. The proposed methodology is divided into two main and parallel parts: on the one hand, the focus is on the generation side consisting in a rooftop PV installation. An artificial neural network is used to develop a forecast model that is able to predict the production of the plant as a function of meteorological data and position of the sun; the result of the model is then used to carry out an anomaly detection and predictive maintenance procedure in order to identify and report, firstly, days of lower production as consequence of malfunctions, and then cases in which it might be necessary an extraordinary maintenance on the system. The other part of the methodology is about the load-side of the domain: typical load profiles are identified and the focus is on the baseload, considered of particular interest because it is the lower value of the consumption, that is always present and whose analysis and reduction can represent an effective strategy of energy saving. The baseload is analysed in terms of normal ranges of power, corresponding to specific external conditions (e.g. external air temperature, working day and month of the year); these ranges are used to construct a KPI that ranks each day in terms of its consumption with respect to values that are considered adequate. At this point, the methodology continues with a simulation of an improved scenario, in which the higher consumption is assumed equal to acceptable values, in order to show the possible energy savings that can be reached with only a different management of the loads.

This thesis work is subdivided in 5 chapters. Chapter 1 illustrates a review of the

works in literature that are related with the topics of the present analysis. Then, Chapter 2 deals with the methods of Data Analytics that are used to carry out the analysis. Chapter 3, instead, describes in detail the developed methodology; it is divided into three sections regarding the three main steps: the creation of the dataset, the load-level analysis and the production-level one. The case study, to which the methodology is applied, is illustrated in Chapter 4 that contains a description of the physical system of interest: the monitoring system, the loads and the production plant. At this point, Chapter 5 shows the results for each one of the steps of the methodology and finally, the Chapter 6 contains a discussion about the obtained results and the conclusion of the overall analysis.

1.1 Literature review

This section contains a summary of the collected scientific papers dealing with the main topics of this thesis work. In particular, in the first part there is an overview of the studies regarding the use of Data Analytics in energy-related applications for buildings; the second part, instead, deals with an outline of papers about different techniques for the anomaly detection applied to photovoltaic power plants.

Xiao et al [21] apply cluster analysis and association rule mining to the consumption of a building. They first process the data, transforming them from numerical to categorical, in order to apply the chosen methods: meteorological quantities (e.g. air temperature and humidity) are categorized into 6 levels, while power consumption data are described with 3 categories (i.e. low, medium and high) by means of the equal-frequency binning method. Then, different clustering methods are applied and all of them result in an optimal number of clusters equal to 3 corresponding to 3 typical operating patterns; at this point, association rule mining is applied to each cluster. The majority of the identified rules could be derived with only the knowledge of the system, but some of them have been helpful to discover potential hidden knowledge and improve the performance of the building; in fact, one of them regards the relation between the cooling load and the operation of the pumps for chilled water: it

represents a useful rule to detect abnormal operations of the devices.

Amasyalia et al. [22], instead, develop a methodology to predict the lighting energy consumption for office buildings. The prediction model is based on a SVM approach, that is chosen because of its ability to solve non-linear problem even with a small training dataset; it is used to predict the lighting energy as a function of two features: day type (e.g. working day or holiday) and daily average sky cover that varies from 0 to 1. The model shows acceptable results, reaching a Coefficient of Variation (CV) of 6.83%, quantifying the variation of the prediction error with respect to the mean of the target.

Another application of Data Analytics in building and energy sector is given by Sendra-Arranz et al. [23] who develop a prediction model of the power consumption of a HVAC system that supplies a self-sufficient small building; in particular, the model consists in a Long Short-Term Memory (LSTM) artificial neural network and the aim of the study is the prediction of the next-day power, given the one of the previous day, in order to implement a potential demand-side management system.

A similar algorithm is proposed by Lin et al. [24]: they carry out a forecast analysis of the power demand of a building in order to discover anomalous patterns. More in detail, power and temperature data are collected from electric meters and a rooftop weather station, they are pre-processed and then used for the power demand forecasting that is obtained by means of a LSTM artificial neural network. At this point, the time series is transformed by means of the SAX algorithm to obtain true and predicted patterns, in order to distinguish between motifs (i.e. regular typical patterns) and discords (i.e. unusual patterns).

Moreover, Capozzoli et el. [25], analyse electrical consumption of a mechanical room of the Polytechnic of Turin. After the collection and the pre-processing of the data, they identify typical electrical load profiles of the system, by means of the application of an hierarchical clustering with Ward's linkage method. Then, the identified four clusters are used as dependent variable in a Classification and Regression Tree whose explanatory variables are the month of the year and the day of the week.

Furthermore, Yang et al. [26] develop a methodology to identify consumption pat-

terns of 10 institutional buildings, by means of a k-shape algorithm. This approach, similar to the k-means clustering, is used to cluster time-series data, its performance is evaluated with the Mean Average Percentage Error and the optimal number of cluster is chosen according to the Elbow method. At this point, the resulting output of the clustering analysis is used in a forecasting step, consisting in a Support Vector Regression model, in order to increase the accuracy of the prediction.

Moreover, Zhenjun et al. [28] apply and show the effectiveness of a methodology to discover typical daily heating profiles of educational buildings; more specifically, after the collection and a proper pre-processing of data, they use a Partitioning Around Medoids (PAM) clustering technique with Pearson Correlation Coefficient as dissimilarity measure. This methodology results to be effective in the knowledge extraction about the building: it is possible to identify peaks, variations of demand and information about starting and ending time of the heating system.

Finally, Alam et al. [27] focus on an educational building with a variety of devices and end-uses; they apply a clustering procedure to analyse the electrical energy consumption patterns of different spaces. Once the data are collected and pre-processed, a k-means clustering algorithm is used to identify patterns for gas and electricity consumption, distinguish them according to the responsible load such as lighting system and plug loads in different spaces.

Regarding the anomaly detection on photovoltaic power plant, the main scientific papers are summarised in the following lines.

De Benedetti et al.[11] propose a methodology for anomaly detection and predictive maintenance on a PV plant: an algorithm able to identify anomalies in the system and potential trends of degradation, in order to early plan maintenance services. Firstly, a model is constructed to give a forecast of the production of the photovoltaic system: a multi-layer perceptron artificial neural network is used to predict the power production, depending on external air temperature and global irradiance. The structure of the neural network consists in 1 hidden layer with 10 neurons and 1 year of measurement is used as dataset, dividing it in learning and testing set. The resulting predicted data are used for the anomaly detection procedure, calculating the residuals

between measured data and predicted ones; the residuals, both hourly and daily, are then compared with limit values defined as multiples of the RMSE obtained during the testing phase of the ANN. Warnings are given if the residuals exceed the limits, distinguishing between possible and strong anomalies; then, the moving average of the residuals and its derivative are calculated: a positive derivative corresponds to an increase in the residuals and so to a degradation trend. Finally, if a warning is in correspondence to degradation area, a predictive maintenance alert is given.

Harrou et al.[12], instead, model the nominal behaviour of the PV plant with a more physical-based approach, the One-Diode model with which they simulate the system, calibrating the parameters of a modelling electric circuit. The anomaly detection is obtained by means of a One-Class Supporting Vector Machine that is an unsupervised techniques which learns a decision rules, identifying an optimal separation plane, on the basis of anomaly-free data and it consequently classifies the new measurements. The inputs of the fault detection procedure are the residuals between the simulation model and the measurements, referring to current and voltage at the MPP. The results of this methodology are obtained simulating scenarios with possible anomalies and evaluating the quality of the detection: in the case of intermittent faults, the efficiency is acceptable reaching a True Positive Rate of 0.997 and an Area Under Curve of 0.871.

Furthermore, Garoudja et al.[13] propose a different methodology to detect and classify the anomalies in a PV plant; as the previous paper, the nominal behaviour of the panels is modelled by means of an ODM and it is used for the calculation of the residuals, as difference with measured data of current, voltage and power at the MPP. Then, the moving average of the residuals is calculated; in particular an Exponentially Weighted Moving Average is chosen, because of its capability to accounts for a wide time range and not only recent information, that results in the possibility to detect even small anomalies. The output of the EWMA is then compared with two control limits that define the acceptable range of operation and that depend on the average of the fault-free data and the standard deviation of the output of the EWMA. Once the anomalies are detected, they are classified depending on the values of power, current

and voltage.

Taghezouit et al. [18], instead, simulate the expected behaviour of a PV system by means of a parametric model in order to obtain the expected power as a function of real climatic condition. Then, a Double EWMA (DEWMA) scheme is used to carry out the anomaly detection with a comparison between its output and a control limit.

A different approach is proposed by Le et el. [14] who develop a methodology for the anomaly detection of PV modules based on infrared radiation cameras and deep neural networks. In particular, an UAV (i.e. unmanned aerial vehicle) with IR cameras is used to provide the temperature distribution on the photovoltaic modules; this information is used by the neural network in order to detect and classify faults and anomalies, reaching an accuracy higher than 90% for the detection, both in training and validation phase.

Moreover, Gaoudja et al. [15] develop an algorithm for the detection and classification of faults on the DC side of the PV system, taking advantage of a Probabilistic Neural Network (PNN). As a first step, the operating condition of the photovoltaic plant is simulated by means of a ODM, that is used to build a dataset with healthy and faulty data. This database is the input of two PNNs, one for the detection and the other for the classification of the fault. The dataset is composed of four type of information (temperature, irradiance, current and voltage) referred to four different operating conditions: nominal operation, three modules short-circuited in a string, ten modules short-circuited in a string and a string completely disconnected from the array.

Another methodology is proposed by Dhoke at al. [16] who apply a simple procedure to detect and localise intra-string line-line faults in a solar PV system. The procedure is based on the calculation of the residuals and their comparison with a threshold value; the residuals are computed as the difference between the measured current in the *i-th* string and an average value (i.e. the current in the array over the number of strings), while the threshold depends on the norm bounded of the string currents in no-faults condition. If the residuals are higher than the threshold, an alarm is given and the faults is localised by means of a regression model as a function of solar radi-

ation and string current.

Moreover, Benkercha et al. [17] suggest an algorithm based on a Decision Tree to identify faults in a grid connected PV plant, depending on climatic condition. The algorithm is based on real data that are collected by means of an acquisition system which records temperature, irradiance and electric variables of the system in both healthy and faulty conditions; in particular the recorded faults include string fault, short circuit fault and line-line one. This algorithm allows to reach an accuracy higher than 99%.

Zhao et al. [19], instead, develop an algorithm based on a Decision Tree for the detection and classification of faults in a PV array. The data that are measured and collected are: air and operating temperature, operating and short circuit current, operating and open circuit voltage; depending on the values of the attributes, the DT is able to detect an occurring faults and to classify it (e.g. shading, line-line fault). Different sizes of the model have been tested in order to verify the accuracy of detection and classification: the detection accuracy goes from about 93% up to more than 99.9%, while the classification accuracy reaches values between 85% and 99.8%.

Finally, Madeti et al. [20] use a One-Diode Model to simulate the nominal and anomalous behaviour of the system, for different values of temperature and solar irradiance; the fault detection is obtained with the k-nearest neighbours algorithm that classifies an object on the basis of the characteristics of the k nearest objects. The authors use, as input of the model, current, voltage and power at MPP and the corresponding values of temperature and solar irradiance, while the output classes correspond to an anomalous or normal operation. With this methodology, it is possible to classify the occurring fault and, in particular, it is possible to distinguish open circuit fault, line-line fault, partial shading and inverted bypass diode fault.

2 Methods

This section contains a description of the methods of Data Analytics that have been used in the procedure followed in this thesis work. More specifically, the first section treats the Clustering analysis and in particular the Hierarchical algorithm; in the second section, instead, there is a description of the Classification and Regression Tree technique and, finally, the third section deals with Artificial Neural Networks.

2.1 Hierarchical clustering

The clustering analysis is an unsupervised technique that is used to identify, in a dataset, groups (i.e. clusters) containing objects characterised by a certain similarity. In particular, the aim of this technique is to label the data in order to obtain groups with an high intra-class similarity and a low inter-class one. The clustering analysis can be carried out by means of three types of algorithms:

- Partitive clustering;
- Density-based clustering;
- Hierarchical clustering.

In this thesis work, only the hierarchical clustering has been used and it will be further described in the following lines.
The Hierarchical Clustering algorithm is a technique by which it is possible to identify nested groups of objects. It can be divided in two types: agglomerative clustering and divisive one, that is the less used; with the last kind of algorithm, the whole dataset is firstly considered as a single cluster and then, with a certain number of iterations, the objects are separated from other ones that are not similar. The agglomerative technique, instead, works in the reverse way: at the beginning, each objects is considered as a single cluster and then, after each iteration, the data points are aggregated according to their similarity, up to the creation of a single cluster that includes all the objects. The similarity between objects is evaluated computing the distance between

them; the most used one is the Euclidean distance that, in a 2-dimensional space, can be calculated with Equation 1.

$$d(X, Y) = \sqrt{\sum_i (x_i - y_i)^2} \tag{1}$$

Once the distance calculation method is defined, the linkage criterion has to be chosen: the issue is the choice of the points of the clusters between which the distance is computed and, as a consequence, which clusters have to be agglomerated. Some of the possible solutions are:

- Single linkage: the clusters that are merged are those with the smallest distance between the closest data points;
- Complete linkage: the clusters that are joined are those with the smallest distance between the furthest objects;
- Average linkage: the agglomeration is done on the basis of the smallest average distance between objects;
- Ward's linkage: the merge is done minimizing the within-cluster sum of squares [29];

Once the clustering procedure ends, the result can be visualised by means of a dendrogram, a tree-like structure that clearly shows the hierarchical structure of all the clusters, starting from the unique one that contains all the data and proceeding with successive splits highlighting the groups contained in bigger ones. At this point, the tree has to be cut on the basis of the optimal number of clusters, that can be evaluated mainly in two ways: either using the Elbow Method or on the basis of the computation of pre-defined metrics. The Elbow Method consists in plotting the percentage of variance explained with the increasing number of clusters: the optimum is the elbow of the curve that corresponds to a situation in which adding groups does not provide much more information [30].

2.2 Classification and Regression Tree (CART)

Classification and Regression Trees are unsupervised techniques that recursively partition the datasets to predict the value of a certain dependent variable as a function of independent ones; the name comes from the possibility to describe the partitioning process with a decision tree [31]. Classification and regression trees are different only for the type of the dependent variables: the former regards categorical variables, while the latter predicts the value of continuous numerical variables. The tree starts from a single node, the *root node*, and it proceeds with successive splits originating from intermediate nodes, the *decision nodes*, dividing the dataset in even smaller parts; at each split the data are divided into two mutually exclusive groups [32]. The nodes at the end of the tree are called *leaf nodes* and represent the final classification of the variables. A tree that is too deep can cause an over-fitting of the model, so it has to be stopped on the basis of some criteria. Among the others, two solutions can be adopted: either to fix the minimum number of data in the leaf nodes or to directly fix the maximum depth of the tree. In order to evaluate the prediction efficiency of the model, a confusion matrix is constructed with the rows representing the predicted values and the columns containing the actual ones; it is filled with the responses of the models, and then some metrics can be computed to asses the performance:

- **Accuracy** is the fraction of correctly predicted values over the total amount of predictions;

$$accuracy = \frac{true\ positives + true\ negatives}{true\ pos. + true\ neg. + false\ pos. + false\ neg.} \tag{2}$$

  Using the confusion matrix, accuracy can be calculated as the sum of the diagonal values divided by the total cases (sum of all the elements of the matrix);

- **Precision** represents the fraction of true positively-predicted values over all the positively predicted ones. In other words, how many values are correct among all the ones that the model predicted as positive.

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \tag{3}$$

Using the confusion matrix, precision can be calculated as the sum of the fractions of the diagonal values over the sum of the values on the rows, divided by the number of classes;

- **Recall** indicates the correctly predicted positive values over the actual positive values.

$$recall = \frac{true\ positives}{true\ positives + false\ negatives} \tag{4}$$

Using the confusion matrix, recall can be calculated as the sum of the fractions of the diagonal value over the sum of the values on the columns, divided by the number of classes;

2.3 Artificial Neural Network

An Artificial Neural Network is a supervised technique that can be used to find non-linear relations between inputs and outputs. It is a black-box mathematical model that can be described by a set of nested functions, represented by layers. In fact, the ANN is composed of elementary units, the *Neurons*, that are grouped in *Layers*, that can be of three types: input, output and hidden layers, that are the intermediate ones [33]. The mathematical function of a layer is described in the Equation 5

$$y = f_{ANN}(x) = g_{ANN}(W_{ANN}x + b_{ANN}) \tag{5}$$

W_{ANN} and b_{ANN} are parameters that are learned by the model at each layer, during the training phase, while g_{ANN} is the activation function that transforms the value before it is propagated to the other layers [33]. An example of activation function is the Rectified Linear Unit (Equation 6) with which negative inputs are turned into zero, while positive ones are left unchanged

$$relu(z) = \begin{cases} 0\ \ if\ z < 0 \\ z\ \ otherwise \end{cases} \tag{6}$$

The quality of the prediction can be assessed by calculating error metrics, such as:

- Mean Average Percentage Error:

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} |\frac{A_i - P_i}{A_i}| \tag{7}$$

- Mean Squared Error:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (A_i - P_i)^2 \tag{8}$$

- Mean Absolute Error:

$$MAE = \frac{\sum_{i=1}^{n} (A_i - P_i)}{n} \tag{9}$$

In Equations 7, 8 and 9, A_i is the actual value and P_i is the predicted value of the i-th point, while n is the number of predicted points.

3 Methodology

The main steps of the methodology followed in the thesis work are summarized in Figure 1 and they will be described in this section.

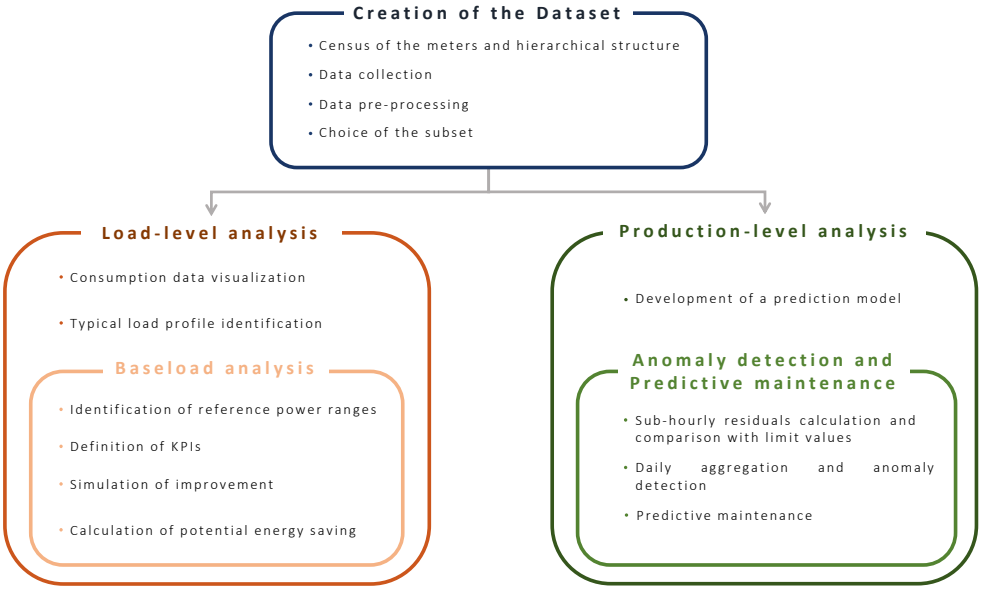

Figure 1: Steps of the methodology.

As a first step, there is the creation of a suitable data-set that has to be used in the analysis. It involves firstly the census of operating electric meters and their arrangement in an hierarchical order and then, the choice of a subset of the whole database in order to focus the analysis on a specific portion of the system and on a certain time range. At this point, electrical data can be collected from meters, but they have to be pre-processed in order to give them an usable structure. The data sources that are used are electrical meters, meteorological stations and online databases in order to provide a complete representation of the system, in terms of electrical consumption and production, climatic conditions and chronological information. Once the dataset is defined, the analysis follows two different and parallel paths: on the one hand it focuses on the load level of the system, on the other the focal point is the production side.

The load-level analysis consists in the visualization and exploration of the overall system and its sub-loads, identifying consumption features and typical profiles; then, a more detailed investigation is carried out on the baseload of the system in order to define average power ranges to be used in the definition of a KPI in order to detect energy wastes and possibilities of energy savings. At this point, an improvement of the consumption is simulated and the potential saving is calculated in terms of electric energy.
Finally, in the last part of the work, the production side of the system is taken into account: an ANN is used in order to develop a forecast model for the photovoltaic production, as a function of climatic information. The predicted production profile is later used to carry out an anomaly detection procedure on the PV system, by means of the calculation of residuals and their comparison with limit values. Moreover, a predictive maintenance approach is proposed in order to point out days in which extraordinary maintenance might be useful.
The process of data elaboration and visualization, as well as the construction of the neural network and the anomaly detection algorithm, is done with R[1], an open-source programming language and environment for statistical computing and graphics.
The steps of the methodology are described more in detail in the following paragraphs.

3.1 Creation of the dataset

This is the first step of the above-mentioned methodology, but it is fundamental in order to create a complete dataset that describes the system and that can be used to work on during the analysis. The data that have to be collected are both meteorological and electrical ones.

Census of the meters and hierarchical structure The latter category of data, the electrical ones, can be obtained from electrical meters that measure energy in a certain range of time and that can be installed at different levels: they can either refer to a specific sub-load, such as a chiller unit, or they can include the consumption of

[1] `https://www.r-project.org/`

an entire area with all its electrical devices. Because of the presence of different levels of measurement, it is necessary to have a structured database which gives information about the operating meters and their connection between the different levels of measurement. So, as a preliminary step in the creation of the data-set, it is necessary to identify meters that actually are in operation, and to define an hierarchical order among meters of different levels, in order to distinguish between *fathers* and *sons* and have a clear idea about loads that are included in each measure. A *father-meter* is a device installed at a lower level of detail on the system and it includes the data from one or more *son-meters* that are associated to loads at an higher level of detail. For this purpose, a census of all the meters is carried out. The information about the meters and their connection comes from various sources: electrical schemes, discussions with professional figures and official documents are used to identify useful devices and to define an hierarchical connection scheme.

Choice of the subset The next step in the creation of the final dataset is the selection of a portion of the measurement network, corresponding to a specific physical subsystem, and a certain time frame to consider during the thesis work. The choice of the area of interest is made on the basis of consideration about the level of detail of the measures, the type of monitored loads and the presence of a production system. In fact, it has been chosen a fraction of the measurement system that includes a photovoltaic production plant and that have a quite high share of monitored sub-loads of different types: chiller units, offices and laboratories.
Regarding the time frame, for the load-level analysis it has been chosen one year of operation to have an exhaustive representation of the variability of the consumption, while for the PV system it was necessary to consider more than one year to catch the seasonality of the production because of the incomplete available data.

Data collection Once the domain has been chosen, the data can be collected: they include energy measurements, information about the day and the type of hour of the measures and other characteristics of the device. The other type of data regards me-

teorological information: external temperature, solar irradiance and its components, and the position of the sun.

Data pre-processing After the collection of all the data, the next step is their pre-processing. It is a fundamental step in data-mining approaches which strongly depend on the quality of input data. In fact, in order to enhance the efficiency and the performance of the analysis, the collected raw data have to be manipulated to increase their quality.

A first cleaning-up of the data is done in terms of type of information of interest: only the columns, and so the quantities, of the raw dataset that correspond to useful information are selected and only the ranges of time at which both meteorological and electrical data are available are taken into account. Then, for load-related data the time granularity of the dataset is reduced: instead of 15 minutes samples, hourly averaged data are considered. This process allows to avoid the fluctuations of the measurement system without a relevant loss of information. For production-related data, instead, data are firstly considered both with 15-minutes time granularity and hourly averaged.

At this point, the pre-processing of the dataset continues with the identification and substitution of anomalous data and punctual outliers, that are observations that strongly deviate from the rest of the sample. Regarding the production side, a first type of anomalous measures is found in those values of power that exceed the size of the plant and in punctual observations with a positive power when the global radiation is equal to zero. These data are not removed, but they are replaced by means of linear interpolation. Then, the outliers identification on both load and production data is carried out by means of boxplots: a standardized representation that allows to graphically depict the distribution of a sample and to highlight the potential values that are out of bounds. The detected points are then evaluated using the knowledge about the system, in order to understand if they are relevant and coherent values or if they are punctual replaceable outliers; in this last case, the substitution is done by means of linear interpolation.

3.2 Load-level analysis

After the definition of a suitable dataset, the second step of this methodology is an analysis on the load-side of the system.

Since the data are collected from a measurement network that has an hierarchical structure, it is necessary to begin with the calculation of the fraction of the measure of the principal meter that can be associated to specific sub-loads; in other words, the share of the total consumption that can be *labelled* with its sub-load has to be distinguished from the remaining *unlabelled* consumption that cannot be associated to a specific end-user. For this purpose, the energy in a whole year is calculated for the meter at the lowest level of detail and for each sub-load; as a consequence, it is possible to calculate the share of labelled and unlabelled consumption.

The load-level analysis starts with the visualization of data related to consumption to have a first overview of patterns and peculiarities; then, an unsupervised procedure for the identification of load patterns is carried out, in order to find typical load profiles, and a classification algorithm is used to define the conditions (meteorological and time-related) that let the load behave according to a specific profile. At this point, the analysis focuses on the baseload that is the lowest consumption of the system; its values are classified according to the previously defined clusters in order to obtain power ranges related to certain boundary conditions that can be used to make considerations about energy saving and anomalous operating conditions. These analysis have been applied on two sub-loads: a chiller unit and an office building, whose profiles and consumption are considered more relevant and complete from data point of view. The baseload of the two sub-loads is firstly analysed in terms of power ranges that can be associated to a normal operating condition; then, a solution for improvement is simulated to quantify the potential yearly energy saving.

These steps of the analysis are summarised in the Figure 2 and they will be further described in the following paragraphs.

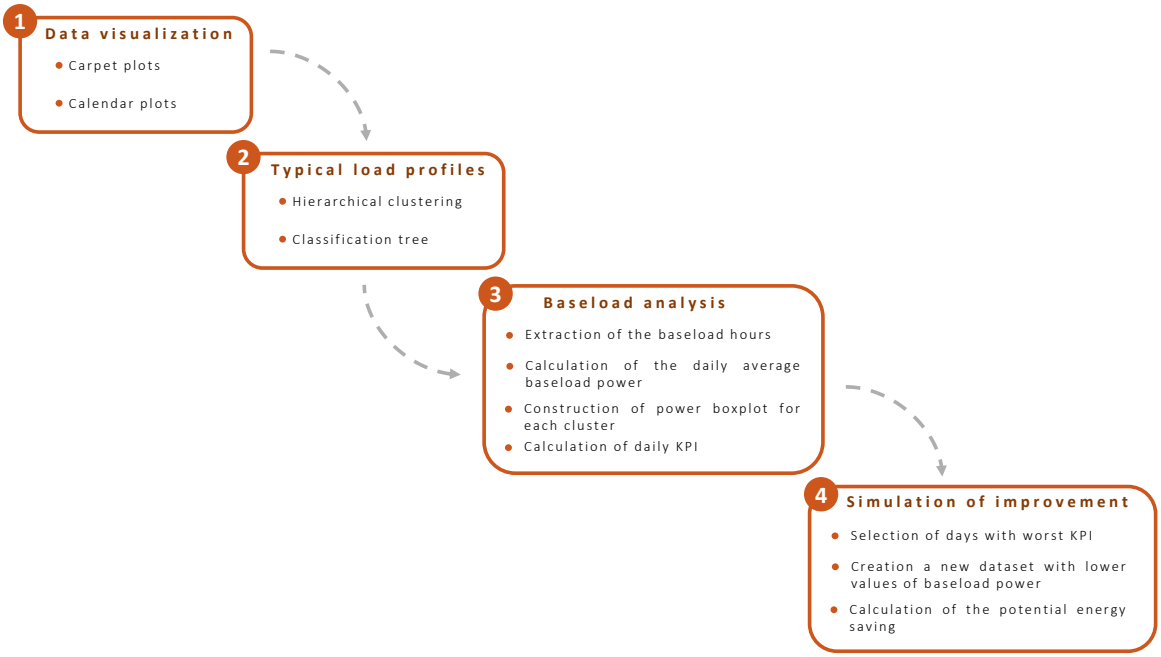

Figure 2: Steps of the load-level analysis.

Data visualization After the identification of the main sub-loads on which the thesis will be be focused, the first phase of the work is the data visualization. This step is very useful because of many reasons: the graphical representation of consumption data, in terms of time series or average value per time step, allows a better understanding and a general overview of the situation. If the visualization is well constructed, it is possible to catch important features about the system: patterns and periodicity can be immediately identified and possible unusual or anomalous situations can be highlighted. In this thesis work, some type of graphical representations are used: bar plots, carpet plots and calendar plots. *Carpet plots* are three-dimensional representations that are particularly effective in the visualization of time series, with the hours of the day on the x-axis, days on the y-axis and values of power or energy at each time step, filling the plot. This kind of graph allows to get an overview of daily or seasonal load patterns and to quickly detect potential missing values. *Calendar plots*, instead, allow to visualize the yearly variability of a relevant feature or quantity (such as the average power or the most consuming load) with daily information, grouping weeks

on the same column; this kind of representation is useful to identify peaks of demand, scheduling issues or to highlight specific days or periods that deserve a more detailed analysis. These visualization tools are widely used in this work to visualize data related to both the overall system and the single sub-load; all the above-mentioned type of plots have been constructed in R taking advantage of the package *ggplot2* [35].

Typical load profiles identification The procedure for the identification of typical load profiles starts with an hierarchical cluster analysis (described in Chapter 2) to identify clusters of daily load profiles that can be grouped on the basis of their similarity.

In particular, the technique has been applied by means of the programming language R, with the following procedure: firstly, the measured load profiles are arranged in an *mxn matrix*, with each of the m raws representing a day and the n columns representing the hours and containing the power value at that time; then, a *distance matrix* has been constructed, computing the distance between each pairs of values in the matrix. At this point, the hierarchical clustering with Ward's linkage method (see Chapter 2) is obtained with the function *hclust*. Once the clusters are formed, the next step is to identify a representative load profile for each cluster of objects: it is the centroid and it is obtained calculating for each time step the average power among the values of all the time series of the cluster. The identified clusters are then used as the dependent variable in a classification procedure (implemented in R with the function *rpart*) which has the final aim to predict the class of each object (i.e. daily load profile) on the basis of chosen independent variables, such as external air temperature, day of the week and month. The performance of the classification is evaluated by means of the construction of the confusion matrix and the calculation of the metrics described in Chapter 2 (Equations 3 and 4).

This procedure has been carried out testing different numbers of clusters in order to choose the best value: the optimal number of cluster is defined as a trade-off value that gives a good performance of the classifier and, at the same time, centroids that can be considered representative of the specific cluster.

3.2.1 Baseload analysis

The second part of the load-level analysis is focused on the baseload, intended as the consumption during *Off-peak hours*, that is the lowest value of consumption which is always present and it corresponds to conditions of non-operation of a device or of non-occupancy of a building. This topic has been considered of particular interest because the baseload is a fixed value under specific conditions that does not depend on occupancy or other non-predictable factors; as a consequence, its evaluation can be useful in order to identify sources of energy waste and saving opportunities.
The analysis is carried out separately for the two main sub-loads (e.g. chiller unit and office building) and it aims at the definition of normal ranges of power values, that can be considered as references and that can be used as terms of comparison with actual values. In fact, these ranges are used to define a KPI to evaluate the daily energy waste; then, an improved situation is simulated to verify the potential energy saving.

Identification of reference power ranges The procedure starts from clusters that are previously identified with the hierarchical cluster analysis for the definition of typical load shapes. For each sub-load and for each cluster, the baseload hours are extracted from the profiles that belong to that group: the power profiles are taken into account only during time ranges in which the load should have fixed consumption because it is not used. More specifically, for the chiller units the *Off-peak* hours are selected, and so, profiles are considered only during night hours and on holidays; for the office building, instead, the focus is on *non-occupancy hours* that correspond to non-working time ranges: the profiles are considered during holidays and in the periods from midnight to 8 a.m and from 19 p.m. to midnight, for the working days.
At this point, a single power value is calculated per day as the averaged quantity on the number of time observations (i.e. on the number of hours, since they are values of power for each hour): for chillers, 8 for working-nights and 24 for holidays, while for the office building, 13 hours for working-days and 24 for holidays.
The daily values of average baseload power are graphically represented in a boxplot that shows their distribution for each cluster and that allows the extraction of some

quantities: the minimum value and the maximum one, the first, second (i.e. the sample median) and third quartiles.

Definition of a KPI Once the boxplots are constructed, the extracted quantities are used to define a KPI, an index that allows to detect energy wastes, ranking a certain day and its baseload according to the average power value and its comparison with normal operating conditions.
More precisely, the daily KPI is calculated as the difference between the average baseload power of that day and the median of the corresponding cluster (Equation 10).

$$KPI = Average\ baseload\ power - Median\ of\ the\ cluster \tag{10}$$

The KPI assumes values between -3 and 3 depending on the results of the calculation and, consequently, on its position in the distribution of the power of the cluster, as it is shown in Table 1. Moreover, a different color is associated to each value of the KPI (as shown in Figure 3) in order to provide a clear and effective visualization of the results.

+3	if KPI > Maximum
+2	if Maximum > KPI > 2nd Quartile
+1	if 2nd Quartile > KPI > Median
0	if KPI = Median
-1	if Median > KPI > 1st Quartile
-2	if 1st Quartile > KPI > Minimum
-3	if KPI < Minimum

Table 1: Values of the KPI according to its position in the distribution of cluster's baseload power.

Figure 3: Color code of the values of the KPI.

The KPI is calculated for the baseload of each day of the year, in order to evaluate them in terms of energy waste. A KPI equal to +3 represents the worst case, in which the average baseload power of that day is higher than the maximum of the distribution of the cluster, revealing a situation of strong energy waste. The best case, instead, is identified with those days with a KPI equal to -3, during which the energy consumption is much lower than the median.

Simulation of improvement The last step of the baseload analysis, is the simulation of an improved situation, in which energy wastes are reduced, in order to make a comparison with the actual consumption and evaluate the potential savings.
The days with higher KPI (i.e. the worst ones) are selected and their baseload powers is simulated equal to the average value of the cluster. The choice of the average value, instead of the median of the distribution, has been done in order to provide a more conservative evaluation of the energy savings. In fact, the average value results to be higher than the median and it means that, generally, the simulated day will correspond to a KPI equal to 1 rather than 0: a more realistic situation is simulated, avoiding too ideal assumptions but testing a relevant decrease in consumption.
A new dataset is created with the modified power values, and it is used to carry out the calculation of the baseload energy. Firstly, the daily baseload energy is computed, considering the measured one for the days with acceptable KPI and calculating the simulated one for the worst days: the hourly average power is multiplied by the number of hours considered for the baseload (24 for holidays, 13 for non-occupancy hours of the office building and 8 for the nights for chiller units). Then, the daily energies

are aggregated firstly on a monthly basis and then on a yearly basis. At this point, as final calculation, the potential energy saving is calculated as the difference between the baseload energy in the real case and the one in the simulated one.
These steps have been followed in a separate and parallel way for the two sub-loads.

3.3 Production-level analysis

In parallel with the load-level analysis, described in Section 3.2, the thesis work focuses on the production side of the system, consisting in a rooftop PV production plant. In particular, after the collection and pre-processing of the related data (see Section 3.1), an artificial neural network is used to develop a forecast model that is able to predict the electricity production, as a function of meteorological and sun-related information. The resulting prediction is then used to carry out a procedure for the anomaly detection and predictive maintenance on the system, in order to detect abnormal behaviour of the plant and warning about the necessity of an extraordinary maintenance on the panels.

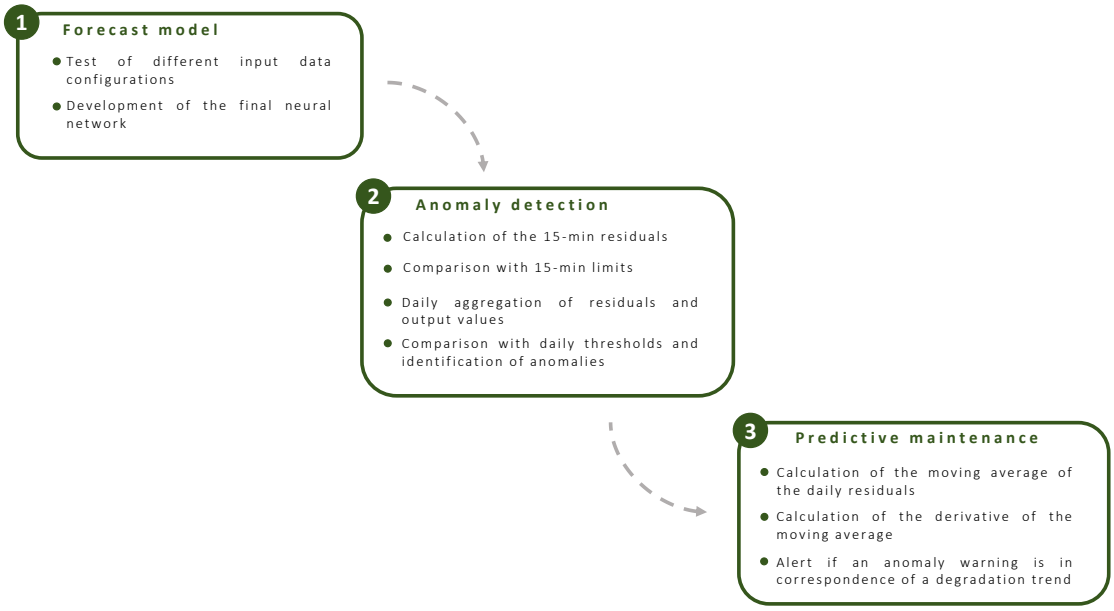

Figure 4: Steps of the production-level analysis.

The steps of this part of the analysis are summarised in the Figure 4 and they will be further described in the following paragraphs.

3.3.1 Forecast model

The first step of the production-level analysis is the development of a forecast model to predict the electric power production of the PV plant as a function of meteorological condition.

This aim is reached with a model based on an Artificial Neural Network, composed of 2 hidden layers with 32 neurons each. The model that has been constructed is not auto-regressive because the production at a certain time does not depend on the previous ones and the inputs of the model are only meteorological information related to sun's position, air temperature and components of irradiance. The parameters set for the learning phase are: 100 epochs, a batch size equal to 32 and the Mean Square Error (Equation 8) as error metric. Moreover, the activation function that has been chosen is the Rectified Linear Units (Equation 6), described in the Section 2.3.

The artificial neural network has been constructed in R by means of the package *Keras*: the function *keras_model_sequential* is used, adding 2 *layer_dense* as hidden layers and another one as output layer, and specifying all the above-mentioned parameters.

The available data from the on-site monitoring system allow the construction of three different models referred to three different outputs: the power production of the east pitch, of the west pitch and of the total plant. In other words, three different neural networks, with the same structure, are developed to predict separately the production of the pitches and the one of the whole plant, in order to verify the possibility to obtain good results with all of them.

Furthermore, regarding the inputs of the model, different data configurations have been tested in order to find the best solution. In fact, various attempts have been carried out considering the combination of different types of data and different data sources, to find the ones that allow the best performance of all the three models. In particular, data are taken firstly from databases available on payment, then from on-

site local monitoring network and finally from a combination of the two sources. The performance of the models is checked both in terms of error during the testing phase and with a visual comparison of the power curves of the production.

3.3.2 Anomaly detection

The predicted power production is used to carry out the detection of anomalous behaviour of PV plant: the procedure is applied at both the separate pitches and the total plant.

The first step is the calculation of the 15-minutes residuals (Equation 11): the difference is calculated between the real power, measured by the monitoring system, and the predicted one resulting from the ANN. The residuals represent a quantitative evaluation of how much the real production of the plant is different from the value it should have. The calculation is done considering the lower time aggregation that is the one with which the measurements are collected.

$$Residuals = P_{real} - P_{predicted} \tag{11}$$

The second step is the definition of 15-minutes limits in order to identify those residuals that are bigger than a certain threshold and, consequently, those operating situations in which the real production deviates from the predicted one, with a value that is not considered acceptable. Two thresholds are defined, both as a multiples of the MAE computed during the testing phase of the artificial neural network; in particular, the considered error is the mean absolute error calculated on the positive powers, so the one calculated taking into account only the situations of effective operation of the system. The defined limits are the Upper Limit (UL) and the Lower Limmit (LL) and they are expressed by the Equation 12.

$$\begin{aligned} LL &= 2.5 * MAE \\ UL &= 5 * MAE \end{aligned} \tag{12}$$

The third step of the procedure is the comparison of the residuals with the defined limits to detect anomalous behaviours. According to the position of the residual with

respect to the thresholds, two abnormal cases are considered: a *strong anomaly* occurs when the residuals are higher than the upper limit, while a *possible anomaly* takes place when the residuals are included between the lower and upper thresholds. The labelling criterion is expressed by the Equation 13.

$$\begin{aligned} \textbf{\textit{Possible anomaly}}\ if\ \ LL < Residual < UL \\ \textbf{\textit{Strong anomaly}}\ if\ \ Residual \geq UL \end{aligned} \tag{13}$$

This comparison and labelling of the residuals results in an output of the procedure that can assume three values, according to the evaluation of the specific residual value which can correspond to a situation of either anomalous behaviour or normal operating condition. The three different cases, corresponding to the three different output values, are summarised in the Equation 14.

$$output = \begin{cases} 1 \ \ if\ Strong\ anomaly \\ 0.5 \ \ if\ Possible\ anomaly \\ 0 \ \ otherwise \end{cases} \tag{14}$$

The next step of the anomaly detection is the aggregation on a daily basis of the residual, summing up for each day of the dataset both the 15-minute residuals and the output values obtained with the previous step. This aggregation is done to carry out the anomaly detection for each day, considering a daily information more practical for a management approach.

The last step of the procedure is the identification of daily anomalous operations by means of a comparison of the summation of the 15-minutes outputs with two daily limits: lower threshold τ_L and upper one τ_U. As for the sub-daily case, also for the daily detection, two types of anomalies, strong and possible, are distinguished; the criterion with which strong and possible anomalies are labelled is expressed by Equation 15.

$$\begin{aligned} \textbf{\textit{Possible anomaly}}\ if\ \ \tau_L < \sum output < \tau_U \\ \textbf{\textit{Strong anomaly}}\ if\ \ \sum output \geq \tau_U \end{aligned} \tag{15}$$

At this point, a daily output value is assigned, similarly to the previous case, according to the Equation 16.

$$daily\ output = \begin{cases} 1 \ \ if\ Strong\ anomaly \\ 0.5 \ \ if\ Possible\ anomaly \\ 0 \ \ otherwise \end{cases} \tag{16}$$

The values of τ_L and τ_U have been chosen in order to guarantee the best performance of the detection, as described in the following paragraph.

Assessment of the detection and choice of the daily thresholds. In order to find the best values of the daily thresholds that result in a good performance of the detection, three cases, corresponding to three combinations of thresholds, have been tested for both the single pitch and for the total plant.
As a preliminary step, each one of the days has been labelled as *possible anomaly*, *strong anomaly* or *normal operation*, in order to be able to compare the results of the detection procedure with the actual behaviour of the production.
At this point, for each case, the performance of the detection has been estimated by means of the calculation of some metrics: True Positive Rate (Equation 17), False Positive Rate (Equation 18), Accuracy (Equation 2), Precision (Equation 3) and Area Under Curve (Equation 19).

$$TPR = \frac{true\ positives}{true\ positives + false\ negatives} = \frac{true\ positives}{positives} \tag{17}$$

$$FPR = \frac{false\ positives}{false\ positives + true\ negatives} = \frac{false\ positives}{negatives} \tag{18}$$

$$AUC = \frac{TPR - FPR + 1}{2} \tag{19}$$

In particular, thresholds are chosen according to the value of the Area Under Curve; in fact, if AUC > 0.9 the detection has a good performance [12]. Consequently, for each pitch and for the total plant, thresholds are defined in order to guarantee an high value of this metric.

3.3.3 Predictive maintenance

Once the anomalous days are identified, the next aim of the analysis is to define a procedure that allow to give an alert if it is recommended an extraordinary maintenance on the plant.

As a first step of this procedure, the Moving Average is calculated on the daily residuals in order to have an overview of their general behaviour; in particular Exponential Moving Average is chosen.

The next step is the calculation of the derivative of the moving average in order to highlight trends in the behaviour: if the derivative is positive, it means that the residuals increase and it is the case of a degradation trend of the system.

Finally, a predictive maintenance alert is given if a warning (i.e. possible or strong anomaly), from the anomaly detection procedure, is in correspondence of a degradation trend. In other word, a maintenance action is suggested if an anomalous day is detected in a period of time in which there are others anomalies and the plant shows more than one consecutive situations of malfunction.

4 Case Study

The methodology described in Chapter 3 has been applied to an educational building, the Polytechnic of Turin, that is a technical university located in the north of Italy. In particular, the analysis begins from its whole measurement system in order to find a suitable sub-system to which apply the described data-mining processes.
This specific case study has been chosen because of three main reasons: the first is the availability of the needed data and the possibility to easily collect them, the second one is the different purpose of its areas and the presence of a generation plant and a thermal-sensitive chiller unit, that gives the opportunity of exploring different systems; the third reason is the fact that it is a well-known system in terms of both physical space and its end-use. This last factor is not trivial because it allows to take advantage of the knowledge about the system to manage the data and understand the results.

4.1 Monitoring system

The Polytechnic of Turin is equipped with a large and nested measurement system, composed of electrical meters at different levels of detail and aggregation. The most aggregated measure, and so the one with the lower level of detail, comes from the principal electrical cabin at which the electrical energy is withdrawn from the national distribution grid at medium voltage (MV). This measure, together with the self-produced energy, gives the information about the total energy demand of the campus. From the principal cabin, the local distribution network takes origin: a medium voltage ring connects the principal cabin with the MV/LV transformation cabins, that are distributed in different locations all over the campus; each cabin is monitored in terms of received energy and it is responsible for the supply of a specific area, by means of a low voltage (LV) distribution network. From here on, the higher levels of the measurement system represent disaggregated measures of loads and sub-loads that are supplied by a transformation cabin. These deeper levels are not complete and exhaustive, in fact not all the devices and buildings are monitored separately, so for

each cabin there is a fraction of the consumption that can be referred to specific loads, while the remaining part cannot be disaggregated and it is considered as *unlabelled* consumption. In Figure 5 it is shown a schematic representation of the first two levels of the electrical monitoring network.

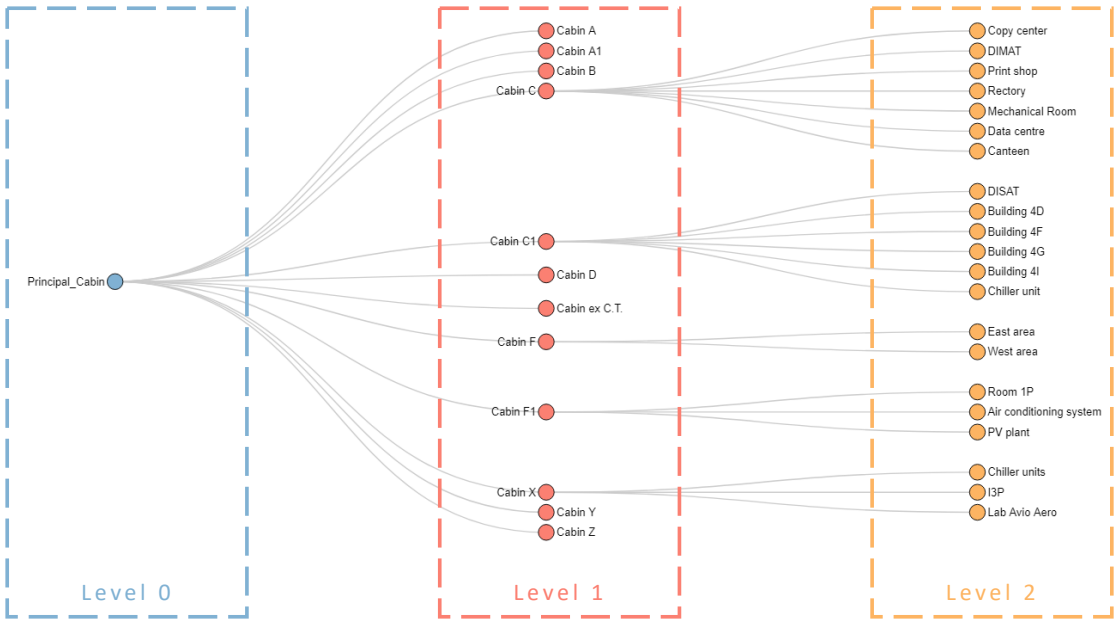

Figure 5: Hierarchical structure of the monitoring system.

The *Level 0* corresponds to the above-mentioned principal electrical cabin and it represents the most aggregated measure, while the *Level 1* consists in the MV/LV substations installed on the medium voltage ring. The transformation cabins are 12: the first 8 (from the Cabin A to the Cabin F) are located in the main headquarters, while the last 4 supply the more recent area, the '*Cittadella Politecnica*'. Finally, the *Level 2* represents all those sub-loads that are individually monitored. As it can be easily observed, most of the electrical consumption cannot be disaggregated because of the lack of installed meters; however, some of the substations give the opportunity to further explore the distribution of the electrical consumption, thanks to the presence of meters at higher level of detail.

Among this last type of cabins, the *Cabin X* has been selected to be the subject of the analysis of the thesis work and it will be further explored in the next paragraphs.

In parallel with the electrical loads, it is also monitored the self-production of the Polytechnic: it consists for the most part of photovoltaic systems, but includes also other devices such as endothermic engines. The PV generators can be connected either at the MV ring or in parallel with LV distribution lines.
In addition to the electrical (both production- and load- level) measurements, the monitoring systems is completed by an on-site meteorological station that is installed on the roof of the main building and it is equipped with all the devices needed to measure solar irradiance, atmospheric pressure, external temperature and wind speed.

4.1.1 Collected data

The meters composing the on-site measurement system allow the collection of an exhaustive variety of data: electrical and meteorological measures and chronological information. The time granularity for all the meters is 15 minutes.
Regarding the electrical meters, the main collected measure is the energy requested or produced in a quarter of an hour, in $[\frac{kWh}{\frac{1}{4}h}]$. This data can be easily manipulated in order to get the power in [kW]. Moreover, each sample is associated to the information about the hour and the day of the measure, such as the day of the week, whether the day is a work day or a holiday and the type of hour to which the data is referred. This last information is related to the tariff-related typology of hour, defined by the Italian authority for the electrical energy and gas (ARERA), with Resolution 181/2006; in this document, the following categorization is stated:

- *Peak hours (F1)*: from Monday to Friday from 8 a.m. to 7 p.m.;
- *Mid-level hours (F2)*: from Monday to Friday from 7 a.m. to 8 a.m. and from 7 p.m. to 11 p.m., on Saturday from 7 a.m. to 11 p.m.;
- *Off-peak hours (F3)*: from Monday to Saturday from 11 p.m. to 7 a.m., on holidays.

Regarding the meteorological station, the following data are collected:

- Global Irradiance in $[\frac{W}{m^2}]$: it is the total power per unit area received by the device;
- Direct Horizontal Irradiance in $[\frac{W}{m^2}]$: it is the power per unit area received by a horizontal surface on Earth;
- External air temprature in [°C];

External data sources Another source that is used to collect meteorological data for the analysis of the photovoltaic system is Solcast[2]. It is an Australian company which provides historical meteorological data with different time granularity (from 5 minutes to an hour) in a specific location. This data are not measures, but they are estimated at the selected geographical coordinates. From this source, the following data are collected each 15 minutes:

- Global Horizontal Irradiance (GHI) $[\frac{W}{m^2}]$: it is the total irradiance received on a horizontal surface;
- Diffuse Horizontal Irradiance (DHI) $[\frac{W}{m^2}]$: it is the horizontal component of the horizontal irradiance;
- Direct Normal Irradiance (DNI) $[\frac{W}{m^2}]$: it is the direct irradiance from the sun, measured on a surface that is perpendicular to the sun;
- Direct Horizontal Irradiance (EBH) $[\frac{W}{m^2}]$: it is the horizontal component of the direct normal irradiance;
- Solar Zenith [°] : it is the angle between the sun and a line perpendicular to the earth's surface;
- Solar Azimuth [°] : it is the angle between a line pointing to north and the sun position;
- Cloud Opacity [%] : it indicates how opaque the clouds are to solar radiation.

[2] `https://solcast.com/`

- Temperature [°C] : it is the temperature measured 10 meters above ground;

4.2 Cabin X

All the data that are mentioned in the previous section can be collected for each point of measurement in the described monitoring network. It has been chosen a defined portion of the whole system, in order to carry out the data-mining analysis on loads and production. The selected domain is the Cabin X with its sub-loads. The area that is supplied by this substation is the north-west side of the campus; this area includes a photovoltaic production system, chiller units and spaces with different intended use: rooms for lectures, offices and laboratories. Regarding the time frame that has been considered, for the load-side it has been considered one year of data, in particular it has been selected the 2019 because it is the last entire year of normal operating conditions. For the production-side, instead, more than one year have been considered in order to make possible the construction and validation of the forecast model.

Aggregated system As a first step, the Cabin X and its energy demand is analysed in its aggregated form. As it can be noticed in figure 6, the total amount of energy request for the sum of the loads, depends on the day of the week and more strongly on the season of the year. The presence of the chiller unit, in fact, causes an increase in the energy demand during the cooling season and in particular in June and July in which it can reach about 16 MWh in some days; it is interesting to notice that even if August is a summer month, with high external temperature, the consumption is lower and it does not have strong peaks of demand: the reason is the summer break during which most of the campus activities are suspended.

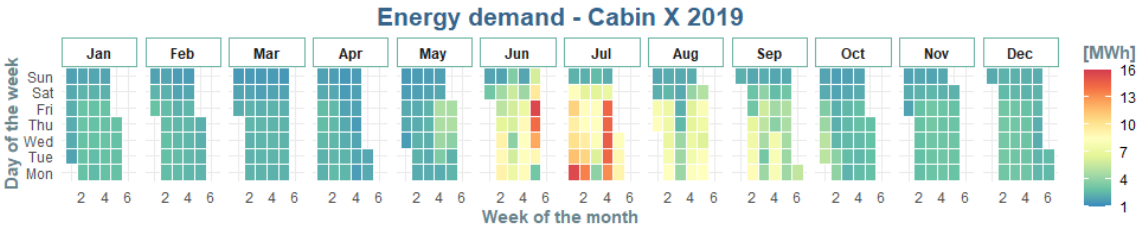

Figure 6: Daily energy demand of the Cabin X in 2019.

During the other months of the year, until May and from October onward, the total daily energy demand is quite regular, with values around 4-5 MWh during the working-days.

Moreover, other consumption patterns can be highlighted if another visualization is chosen: in Figure 7 it is shown the power profile during the day, for each day of the year. The power request follows the typical occupancy pattern of an educational building: it is higher during working-hours and lower when the buildings are not used. Then, as expected, there is an increased power demand during summer months, with peaks in the middle of the day, during the hours in which the external temperature is high and the spaces are used.

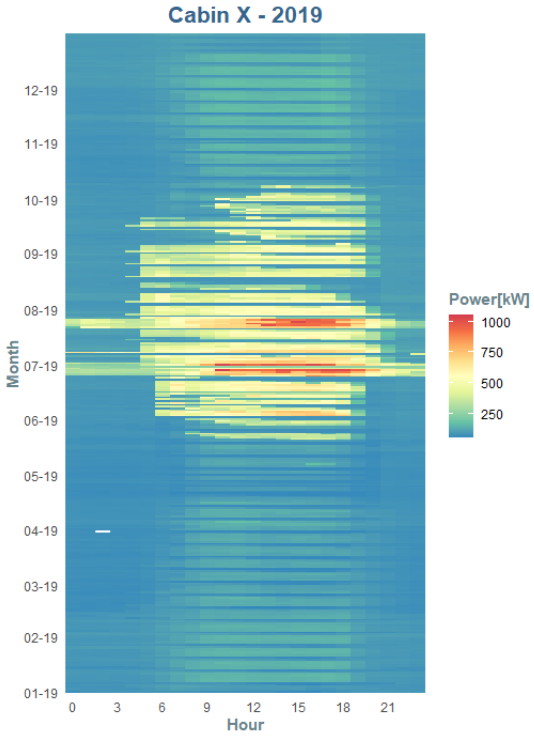

Figure 7: Power demand of the Cabin X in 2019.

The total energy demand of the whole cabin is satisfied by both the withdrawn of electricity from the national grid and the self-production system: a PV production plant that is installed in the geographical area of the cabin and that releases electricity in a LV line downstream the substation.

The following sub-sections will go into detail of the Cabin X, describing firstly the PV production system feeding part of the area of interest, and then the monitored loads.

4.3 Production-side

The above-mentioned production system is a PV plant identified with the name *'ex fucine'* and installed on the roof of the building that hosts the Innovative Companies Incubator[3] (I3P); it injects the produced electricity in parallel with one of the LV distribution branches downward the cabin. The plant has a size of 35[kWp][34], assumed being equally divided into its two pitches: in fact, the system is composed of a *East*

[3] `https://www.i3p.it/`

Pitch and a *West Pitch*, placed on the two side of the sloping roof.

Available data The two pitches of the photovoltaic system are individually monitored and this offers a good level of detail, but the issue concerning this portion of the measurement system is the lack of data during large periods of time because of the malfunctioning of the monitoring devices. In Table 2 it is reported a summary of the complete months in which data are available and it can be noticed that the missing months are often different for the two pitches.

Year	**East Pitch**	**West Pitch**
2018	From January to October	All the months
2019	From July to December	All the months except for August
2020	From January to October	From January to May, October and November
2021	From January to September	From January to June

Table 2: Months with complete PV data per pitch.

4.4 Sub-loads

The Cabin X supplies a large area of the campus and a great variety of loads, but not all of them are monitored, so not all the consumption can be disaggregated and associated to a specific load. In fact, there are only three loads whose consumption is individually monitored: the I3P headquarters, that is an office building, Avio Aero laboratory, that deals with research in aerospace field, and a chiller unit; in particular this last load is further monitored with meters at each of the 5 chillers that compose the unit. In Figure 8, it can be seen a schematic structure of this portion of the measurement system.

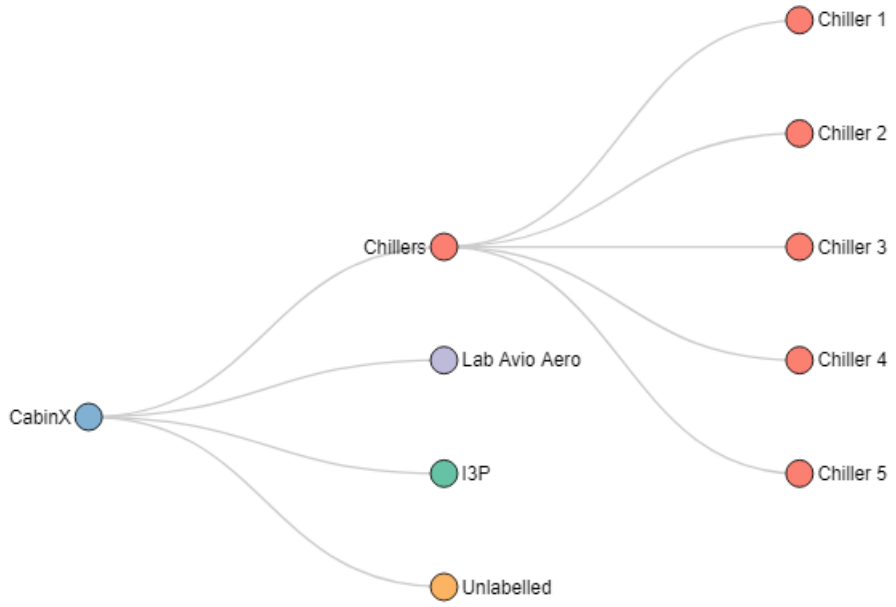

Figure 8: Structure of the monitoring system of the Cabin X.

However, even if the number of individually monitored loads is small, they cover a good share of the total energy demand.

In fact, as it can be observed from the pie-chart in Figure 9, on a yearly energy basis the unlabelled portion of the load is about the 34% while the labelled one is the largest part: the Lab Avio Aero is responsible for the 20% of the energy request, the I3P for the 18%, while the chiller unit account for the 28% of the total consumption. So, even if the chillers work for a specific season, they represent the highest load on a yearly basis.

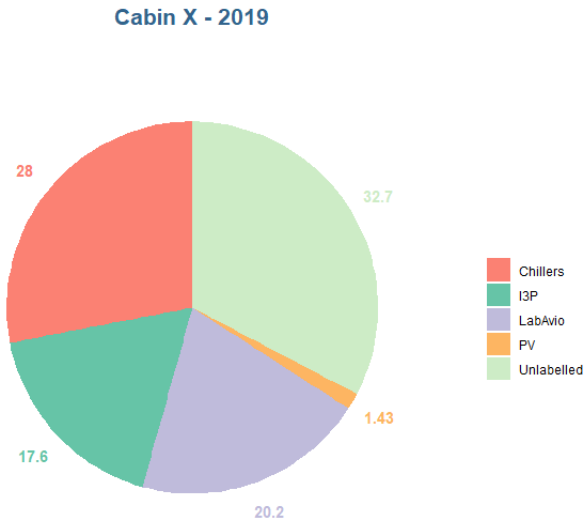

Figure 9: Percentages of consumption of labelled and unlabelled sub-loads in 2019.

In the following subsection two monitored load will be explored individually: the I3P building and the Chillers. The Avio Aero laboratory is not further explored because at the moment it is no more in operation; moreover, his behaviour is too dependent on the scheduled activity, compromising the possibility of general considerations.

4.4.1 I3P building

The first analysed load is the building in which the Innovative Companies Incubatore (I3P) has the offices.

It is an independent building, with the PV system installed on its roof and connected in parallel with the LV electric line: the photovoltaic plant satisfies part of the building's energy demand. In fact, in order to know the actual energy demand of the building, it has to be considered not only the measure of the monitoring device that is installed on site, but also the production of the PV system that is wholly used to supply the load. This can be easily noticed in figure 10 in which the orange curve represents the power measured by the electrical meter, the blue curve is the photovoltaic power

production, while the green line is the sum on the other two and it is the load profile of the building.

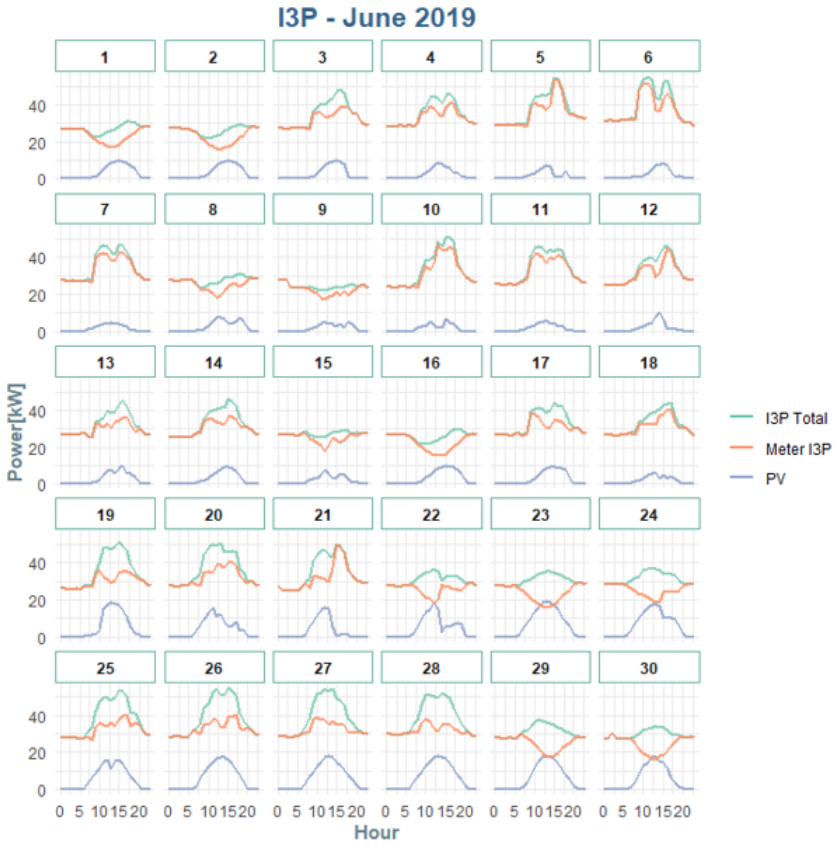

Figure 10: Power profile of the *I3P* building in June 2019.

From the figure, it can be noticed that when the PV system starts to operate, the measurement of the meter decreases: in this case, to satisfy the energy demand of the building, the self-production is used, decreasing the need for electricity from the national grid.

Dependence on external air temperature Before starting the analysis on the I3P building, it has been checked that this sub-loads is not thermal-sensitive. In fact, it is an office building whose cooling is the work of the centralised chiller unit (described in the following paragraph) that is individually monitored.

Figure 11: Energy demand variation with temperature, I3P building.

The Figure 11 shows, for each month, the hourly energy demand and the corresponding external air temperature; moreover, the different colors distinguish the tariff-related type of hours, as described in paragraph 4.1.1.

The graph confirms the assumption of thermal-insensitivity of the load. In fact, it is clear that there is not a correlation between the energy demand of the building and the external air temperature; there is, instead, a strong dependence between the energy consumption and the hour of the day: during *Off-peak hours* (purple dots) and *Mid-level hours* (orange dots) the demand of electricity is much lower than the one during *Peak hours* (green dots).

4.4.2 Chillers

The second load, in addition to the I3P building, that is individually monitored and that will be further analysed, is the chiller unit.

It can be considered both as a single load and as the sum of its components whose consumption is measured too. However, in this thesis work, it will be considered in its aggregated form also because not all the single components continuously work, as it can be observed in Figure 12 in which the energy consumption is shown for each

month and for each one of the chillers.

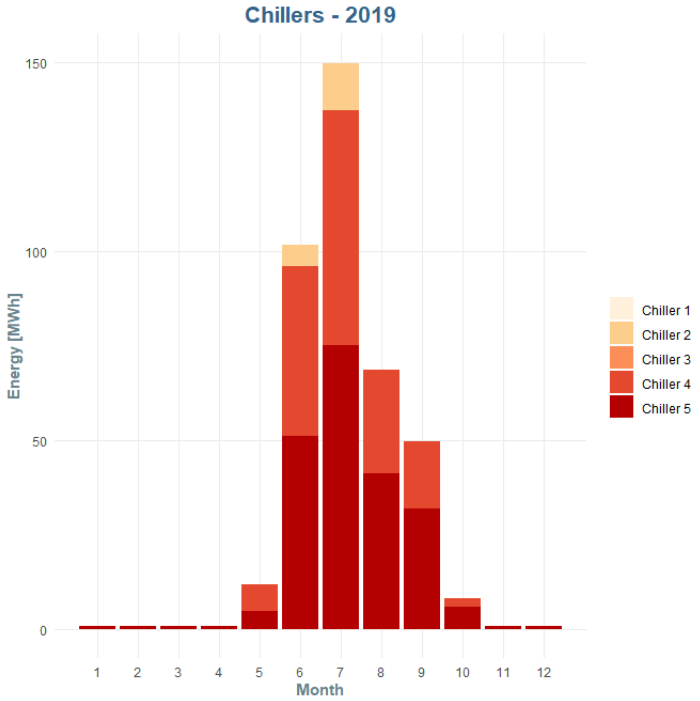

Figure 12: Energy consumption of chillers in 2019.

In general, as it can also be observed from the barplot, the operating period of the chillers is from the second half of May to the first weeks of October; during the other months, they are not in operation except for the antifreeze circuit that requires a maximum power of about 1.6 [kW] and it is always in function.

5 Results

This section deals with the results obtained from the analysis: they are reported for each one of the steps of the methodology that is described in Chapter 3.

5.1 Data pre-processing

Once the raw data are collected from the above-mentioned sources (see section 4.1.1), they need to be explored in order to check up their coherence and look for problematic values.
In the following paragraphs the results of the data pre-processing are showed separately for the production system, firstly described, and for the sub-loads.

Photovoltaic system A first cleaning up of the data begins with the identification of those measures that exceed the size of the plant and of the power values that are positive even if the solar irradiance is equal to zero. These data cannot be related to a normal operating condition of the system, so they are considered anomalous and related to a malfunctioning of the meters or of the data transmission system; once identified, they are replaced by means of linear interpolation. In the first category of anomalous data there are just the measures of the west pitch of the plant, while the east one does not present such problem. In Table 3 there is a summary per year of the number of the identified measure.

Year	Number of measures exceeding the size	% of the positive powers
2018	8	0.05%
2019	30	0.2%
2020	42	0.38%
2021	35	0.43%

Table 3: PV measure exceeding the size of the plant

With the second criterion, instead, 6165 values are identified, that is to say the 0.05% of the total number of positive powers.

The next and last phase of the data cleaning consists in the construction of boxplots with the aim of identifying possible outliers. For each pitch, the distribution of power is represented as a function of the month, for each year. As an example, it is shown in Figure 13 the boxplot related to the power of the west pitch in 2019.

Not all the identified outliers are considered as such; in fact, these points are visualized on the PV power curves in order to make considerations about their coherence: only the punctual values are substituted by means of linear interpolation, while the consecutive ones correspond only to an higher production in that month.

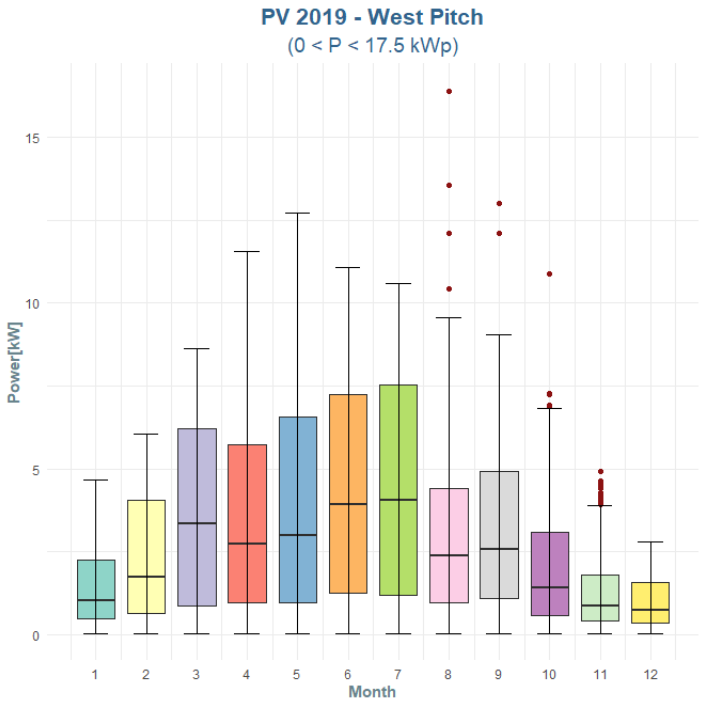

Figure 13: Boxplot of the power of the PV west pitch as function of the month.

Sub-loads The check on the data of sub-loads, unlike the one related to the production system, has not shown any particular issue: there are no missing values and none of the identified outliers needs to be replaced.

In fact, a boxplot has been constructed for each of the loads in order to show the power

distribution in 2019 as a function of the months: a great number of points are labelled as outliers but in this case it has been fundamental the further inspection of those data as part of the load profile, visualizing them on the power curves. This shows that the points that are labelled as outliers are just higher or lower value of consumption in that specific month, but with a load profile that is perfectly coherent with the load. As an example, the boxplot (Figure 14) and a power curve with highlighted potential outliers (Figure 15) are shown for the I3P building, but the same considerations have been made also for the chillers.

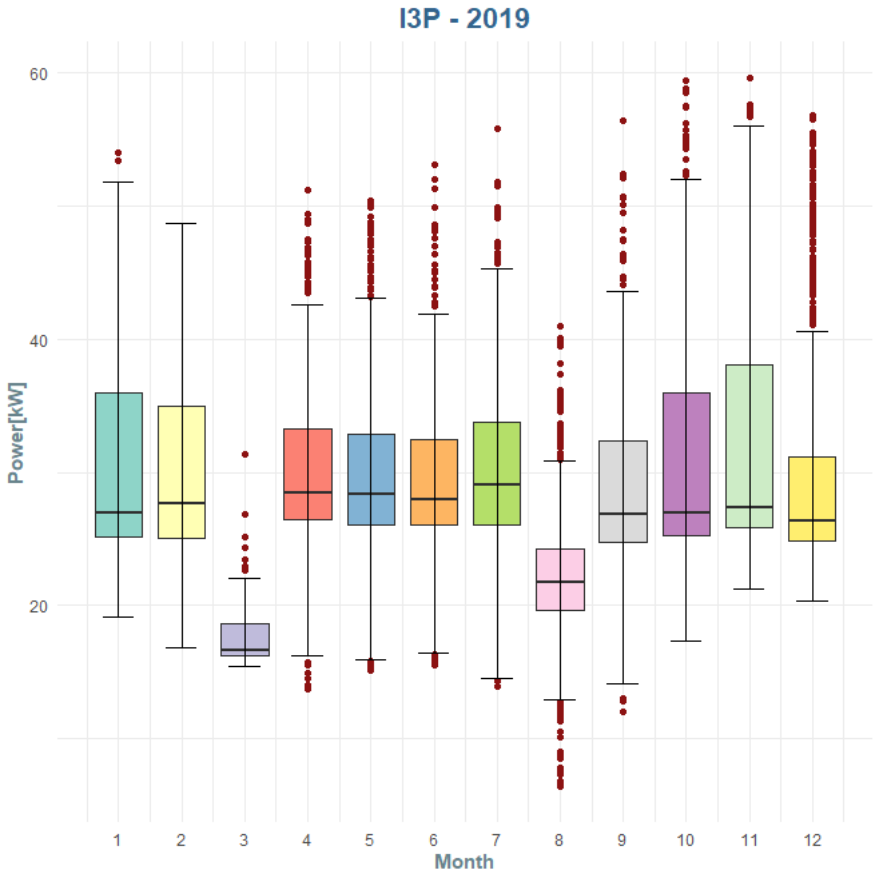

Figure 14: Boxplot of the power of the I3P as function of the month.

If December is taken as an example, as shown in Figure 15, the operating points that are labelled as outliers are those with high consumption during the office hours on working-day. However, the load profiles are perfectly coherent with the destination of use of the building and the labelling as values out-of-the-bounds is due to the Christ-

mas break that makes the profiles flat during the last weeks, decreasing the monthly average value of the power.

The same happens in August, during which the summer break has the same effect. For the other months, instead, the points correspond just to days with an higher or lower consumption with respect to the others, but always with a coherent profile.

In conclusion, there are no punctual outliers that can be considered as measurement errors or data-related issued, so the original values are kept as they are without proceeding with any substitution.

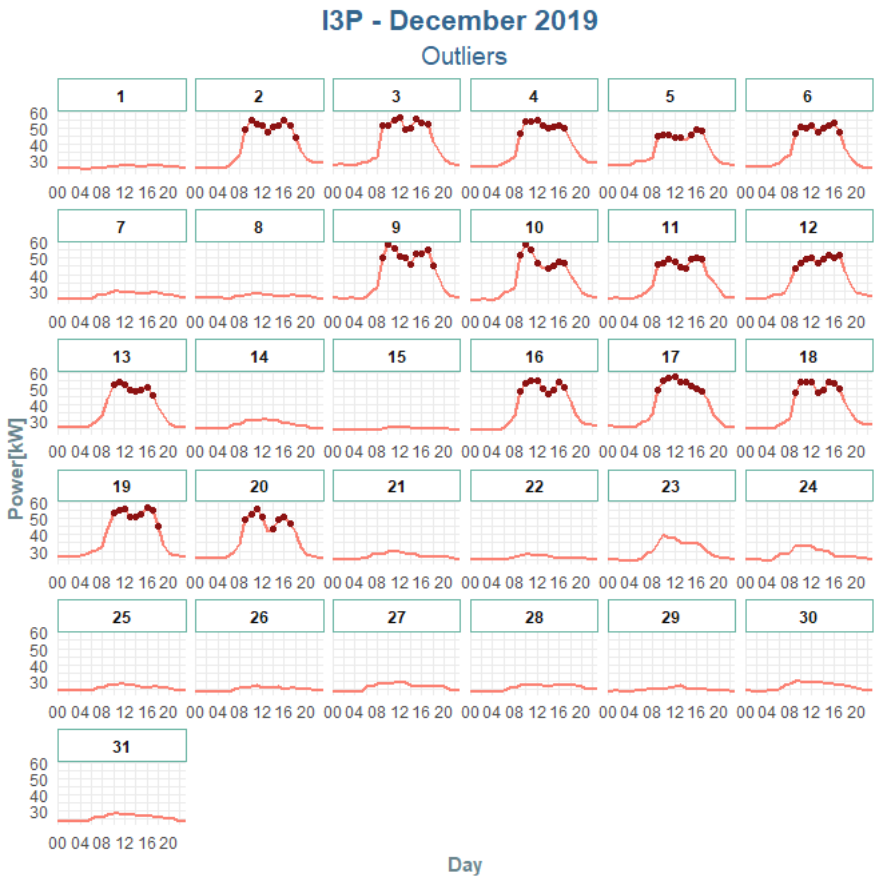

Figure 15: Potential outliers on the I3P power curve, December 2019.

5.2 Load-level analysis

In this subsection the results are reported for the part of the analysis concerning the load-side of the Cabin X.

5.2.1 Consumption data visualization

As defined in the paragraph 3.2 of the methodology, once the data have been checked, the loads can be explored in terms of their consumption, in order to have a first representation of their behaviour and periodicity.
In the following paragraphs the data are visualised firstly for the I3P building and then for chillers.

I3P building As expected from its use, the daily energy demand pattern of the I3P building is quite regular during the weeks of the year, except for the holiday period. As it can be seen in figure 16, during the working-days of a week the building requires quite a constant amount of energy that goes from 800 to 1000 [kWh], depending on the month. During weekends and holidays, instead, the demand is lower because of the lower occupancy of the building.
A particular case is the month of March: it has the lowest values of energy demand, both during weekdays and weekends; this behaviour is not peculiar for the specific considered year, in fact this characteristic was also present in March 2018. So, it has been considered as a correct scheme due to the scheduled activity of the company that works in the building.

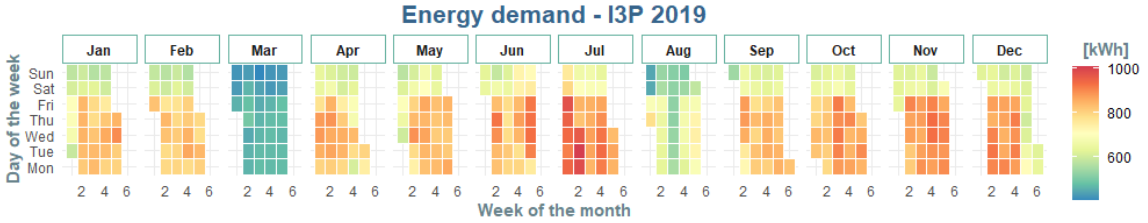

Figure 16: Energy demand of the I3P building per day in 2019.

The load can be analysed also in terms of power profiles and their dependency on boundary conditions. Observing the load profiles taken together in chronological order (see carpet plot in Figure 17), at first sight it is evident that the load follows the occupancy of the building: the power is lower during non-working hours, lunch breaks and holidays.

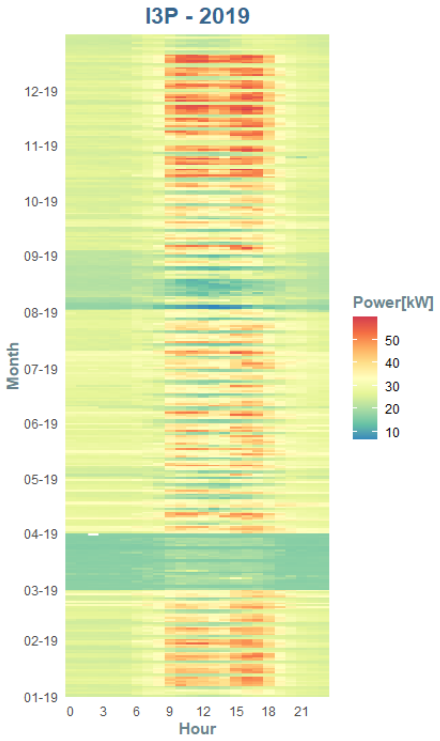

Figure 17: Power demand of the I3P building in 2019.

Chillers The chiller unit is a seasonal load and it works only during the hot months, but it has to be taken into account that it supplies buildings with a quite regular occupancy pattern. Because of these reasons, the daily energy consumption is higher during the working-days of the summer months while it is reduced during weekends, in particular on Sunday and during holidays or periods in which the spaces are less used.

As it is shown in Figure 18, which represents the energy consumption per day, the chillers start to operate during the last two weeks of June up to the end of the cooling season that is more or less the second week of October. For the most of the operating days, the energy consumption is about from 4 to 6 [MWh], with peaks up to 10 [MWh] in some days of June and July, during which the air temperature is high and the buildings are used.

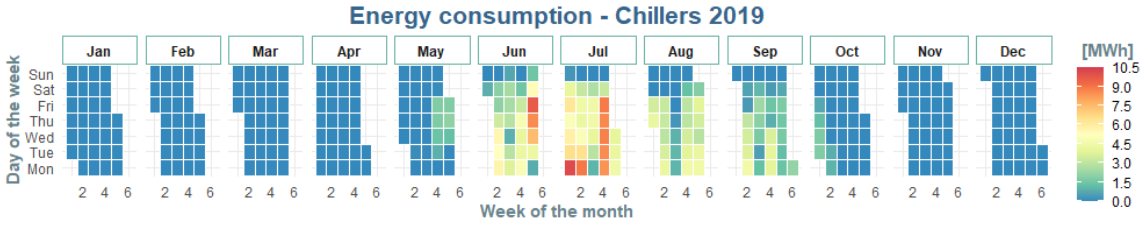

Figure 18: Energy consumption of the chillers per day in 2019.

Instead of considering a daily aggregated visualization of the consumption, it is also possible to look at the daily power profiles during the year by means of the carpet plot in Figure 19. It can be noticed the seasonality of the load and the fact that it supplies areas with scheduled working time; they are in operation for a time range that is larger than the working-hours period because of the earlier starting with the aim of reaching a comfort indoor air temperature. Even if the chillers are off during nights, there are some exceptions: in some days of June and July (the same that in figure 18 show the highest energy consumption) the chillers work at high power during night.

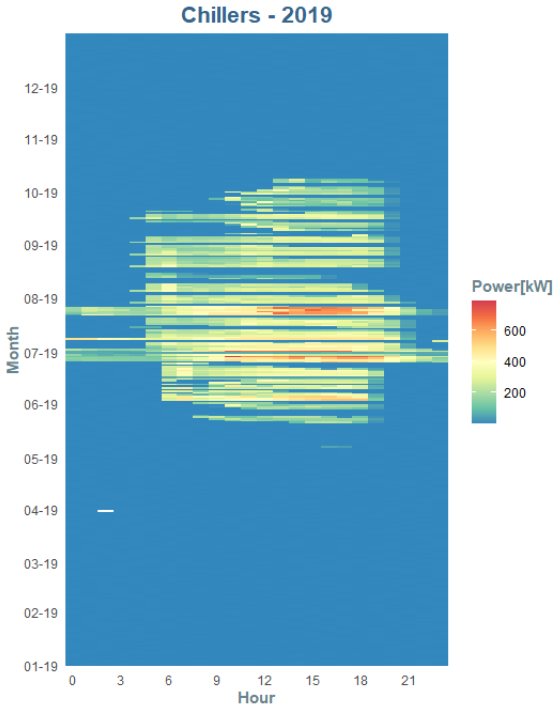

Figure 19: Power demand of the chiller unit in 2019.

5.2.2 Typical load profiles identification

As illustrated in the methodology, to find typical profiles of the considered sub-loads, a clustering analysis and a classification procedure are carried out, identifying as optimal number of cluster a trade-off value which allows to have representative profiles and a good performance of the classifier.
In the following paragraphs the results are reported for the I3P building and for the chiller unit, showing first the final outcomes in terms of identified load profiles and classification tree and then, for sake of completeness, the other attempts.

I3P building For the I3P building, following the procedure described in the methodology, the optimal number of cluster is found to be 3. This value allows to have acceptable typical profiles, that are shown in Figure 20, and a good performance of the classifier, whose resulting decision tree is illustrated in Figure 21. Moreover, in Table 4 there is a summary of the metrics that have been calculated to asses the performance

of the classifier; in particular there are the values of accuracy (Equation 2), recall (Equation 4) and precision (Equation 3).

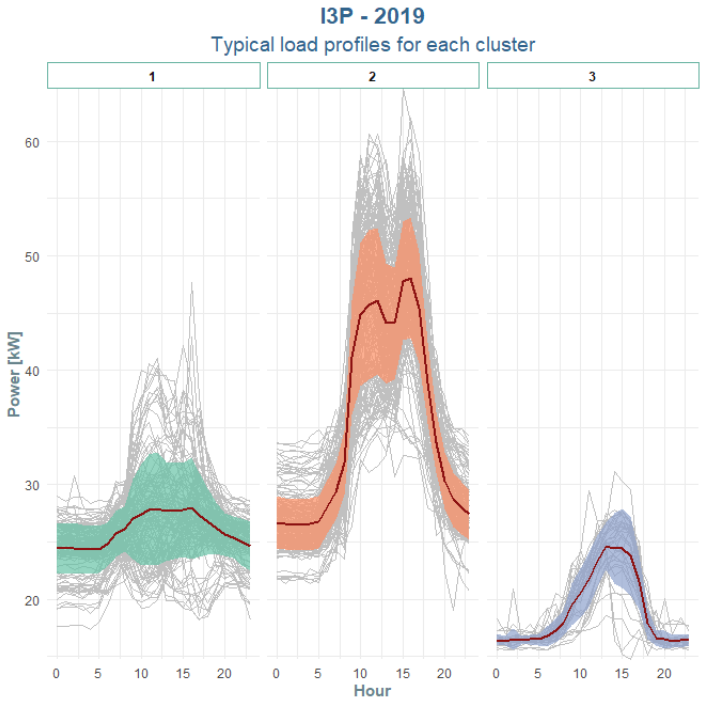

Figure 20: Typical load profiles of the I3P building.

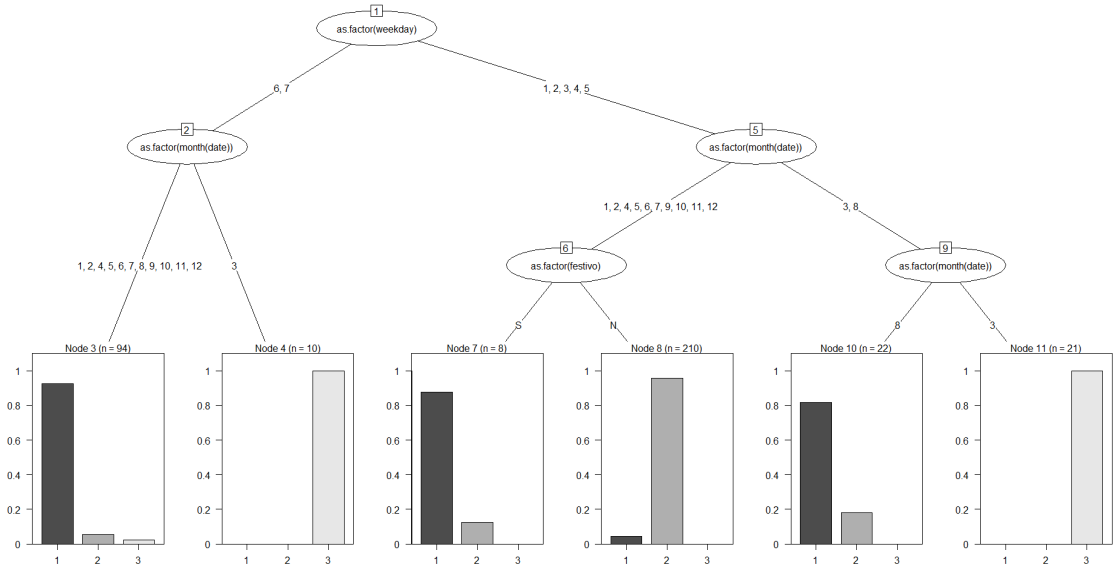

Figure 21: Classification tree for the I3P building.

Accuracy	94.25%
Recall	96.35%
Precision	93.92%

Table 4: Classifier's performance metrics with 3 clusters for the I3P buildings.

From the observation of both the figures, that summarise the results of the procedure, some consideration can be made. The profiles that represent the Cluster 1 is typical of the weekends and holidays: during these days, the energy demand is lower and the profile is flatter because the building is generally not used, except for some isolated cases. The profile of the Cluster 2, instead, is the one of the working-days: it is the one that reaches the highest values and it follows the normal occupancy pattern of an office building. In fact, the power starts to increase around at 8-9 a.m., reaching a first peak a couple of hours later, when the office is full; then, it decreases in correspondence of the lunch break. In the afternoon, there is a second peak of demand after which the power decreases, up to its lowest value at the end of the working-day, around 7 p.m.. Finally, the third profile is the one that is associated to the month of March, that has the lowest value of consumption.

A clear visualization of the clusters pattern is given by the calendar plot in Figure 22.

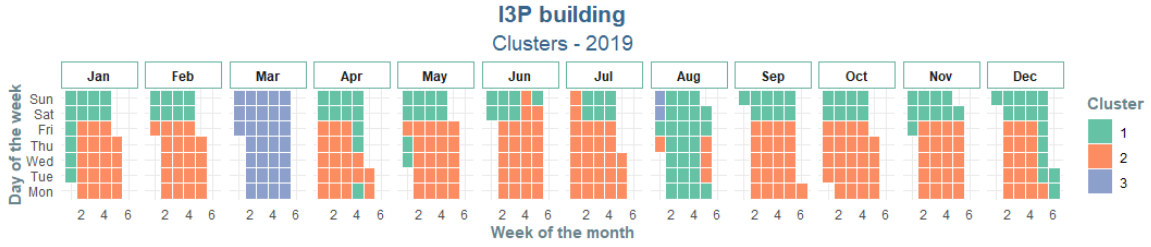

Figure 22: Corresponding cluster for each day for the I3P building, 2019.

The calendar plot allows to have a general overview about the clusters distribution and to observe, for example, which ones are the days with low consumption and flat profile (e.g. weekend and holiday); regarding these days, it can be noticed that there are situations in which a certain day, that should have flat profile (i.e. Cluster 1),

belongs to Cluster 2 with high peak powers: this is the case of some weekends of July and some days of August. This observation highlights the necessity of the above-mentioned analysis on the baseload, that will be described in the following parts of the work.

In order to reach the described results, as mentioned in the methodology, other two number of clusters are tested but they result to be not satisfying and with lower performance. In Figure 23 the typical load profiles are shown for the case of 4 and 5 clusters, respectively.

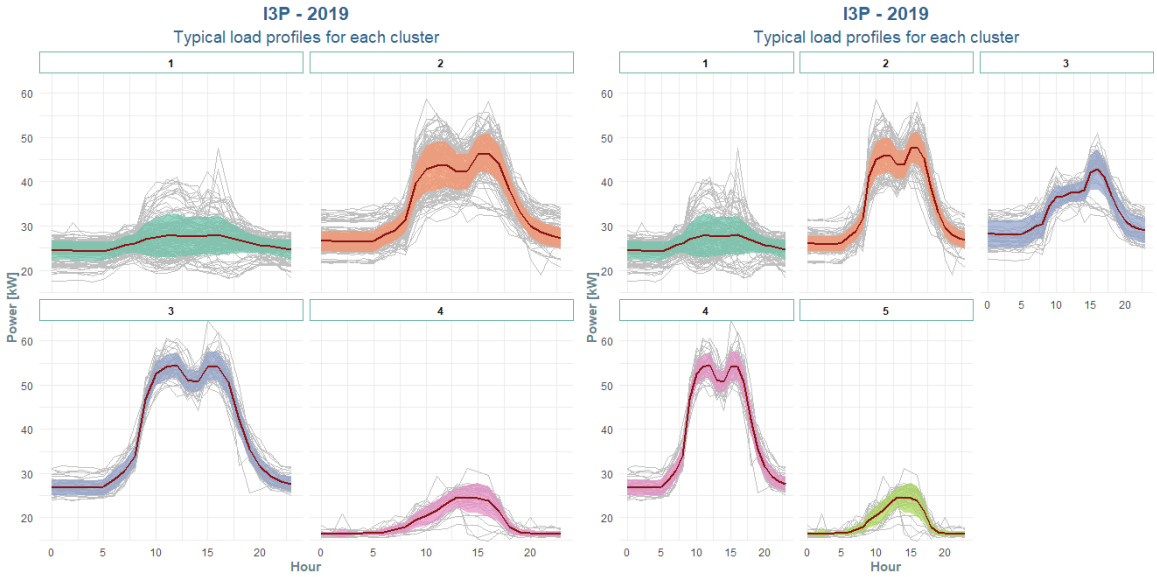

Figure 23: Typical load profiles for I3P building in the case of 4 (left) and 5 clusters (right).

They seem to be more precise and representative than the case with 3 clusters, that is finally chosen; however, they cause a low performance of the classifier (as it can be observed from the metrics summarised in Table 5) because of an over-fitting of the model on the data. In fact, some of the clusters are never predicted by the classifier and this causes a decrease of the performance metrics.

	4 clusters	5 clusters
Accuracy	80.82%	72.05%
Recall	65.86%	60.54%
Precision	68.8%	66.5%

Table 5: Classifier's performance metrics with 4 and 5 clusters for the I3P buildings.

Chiller unit For the chiller unit, the optimal number of clusters results to be equal to 4. The Figure 24 shows the typical profiles for each one of the identified clusters: they are quite representative, except for some daily profiles belonging to the first two clusters; however, this situation has been considered acceptable because they represents days in which chillers work only for few hours or have a behaviour that can be considered anomalous.

The classification with 4 clusters gives good performance, as it can be observed from the metrics in Table 6, and it results in the decision tree represented in Figure 25.

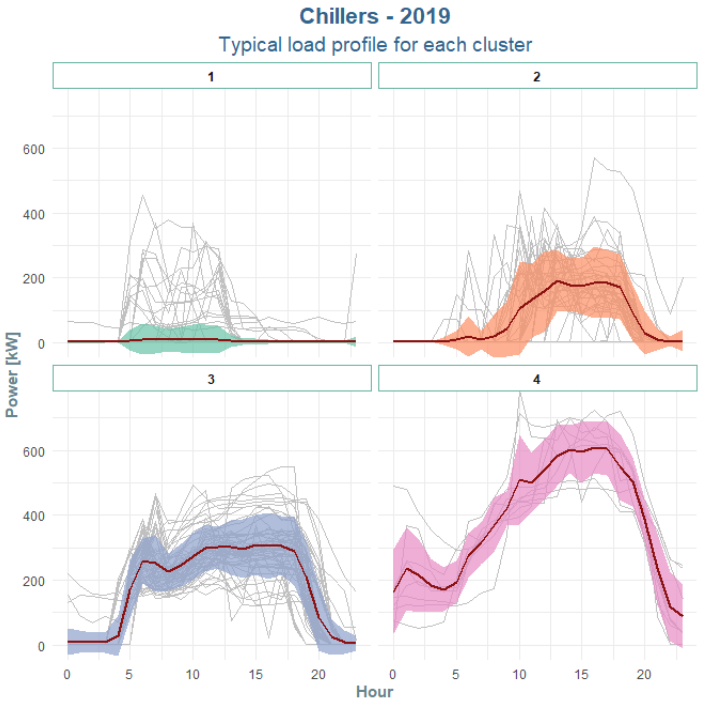

Figure 24: Typical load profiles of the chiller unit.

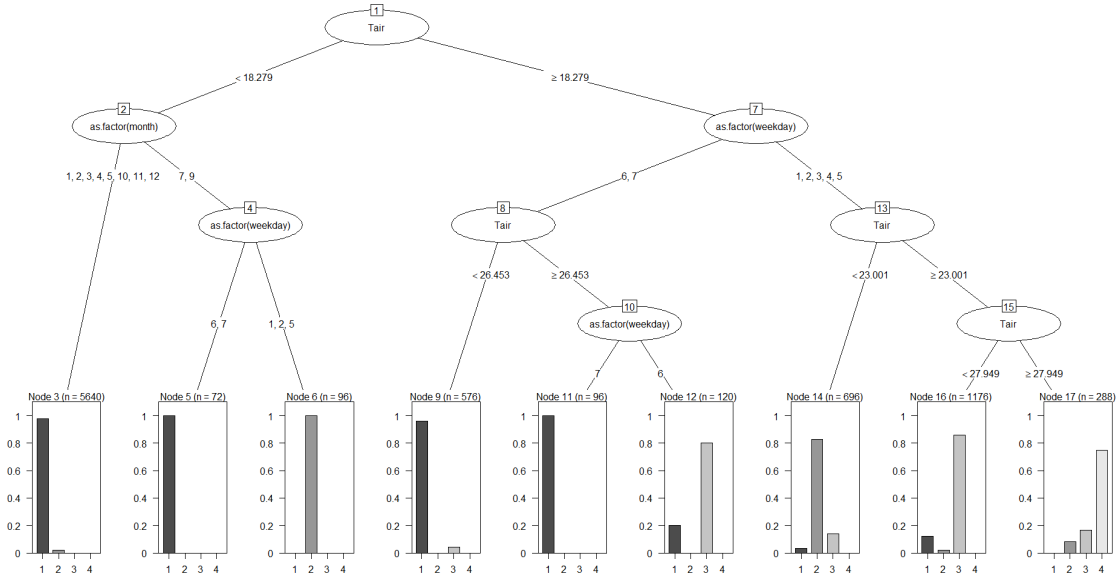

Figure 25: Classification tree for the chiller unit.

Accuracy	92.33%
Recall	83.69%
Precision	88.88%

Table 6: Classifier's performance metrics with 4 clusters for the chiller unit.

As for the I3P case, also for the chiller unit, from the observation of the results, some considerations can be made about the boundary conditions that results in a certain cluster. Unlike the I3P building, the chiller unit is a thermal-sensitive load and, consequently, also the external air temperature appears in the classification tree. The first profile is typical of days with an external temperature lower than 18 °C, weekends with temperature lower than 26 °C and Sundays: it is the case in which the chillers barely work because of a low temperature or a low occupancy of the buildings. The load profile of the Cluster 2, instead, is associated to working-days with temperatures between 18 °C and 23°C, but also to some days of July and September with low temperatures. Moreover, the third profile is the one of working-days with temperature between 23°C and 28°C and some hot Saturdays. Finally, the fourth profile is the

one with the highest power even during nights: it is typical of the days in which the average external air temperature is very high, more than 28 °C.
As it is highlighted in Figure 26, the Cluster 1, that corresponds to low power and flat profile, is the one of cold months and most of the weekends; Cluster 4, instead, represents only 8 days, all belonging to June and July that are the hottest months. Clusters 2 and 3, instead, includes the other summers days, depending on the temperature.

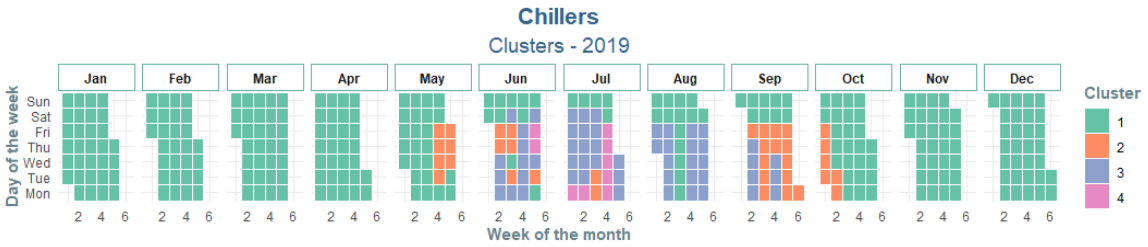

Figure 26: Corresponding cluster for each day for chiller unit, 2019.

As it has been done for the I3P building, for sake of completeness, the results are reported also for the tested number of clusters that are not chosen as optimal. In particular, the Table 7 summarise the classifier performance metrics for the cases with 3 and 4 clusters, while the Figure 27 shows the correspondent typical load profiles.

	3 clusters	**5 clusters**
Accuracy	94.79%	89.59%
Recall	85.92%	62.87%
Precision	92.2%	71.1%

Table 7: Classifier performance metrics with 3 and 5 clusters for the chiller unit.

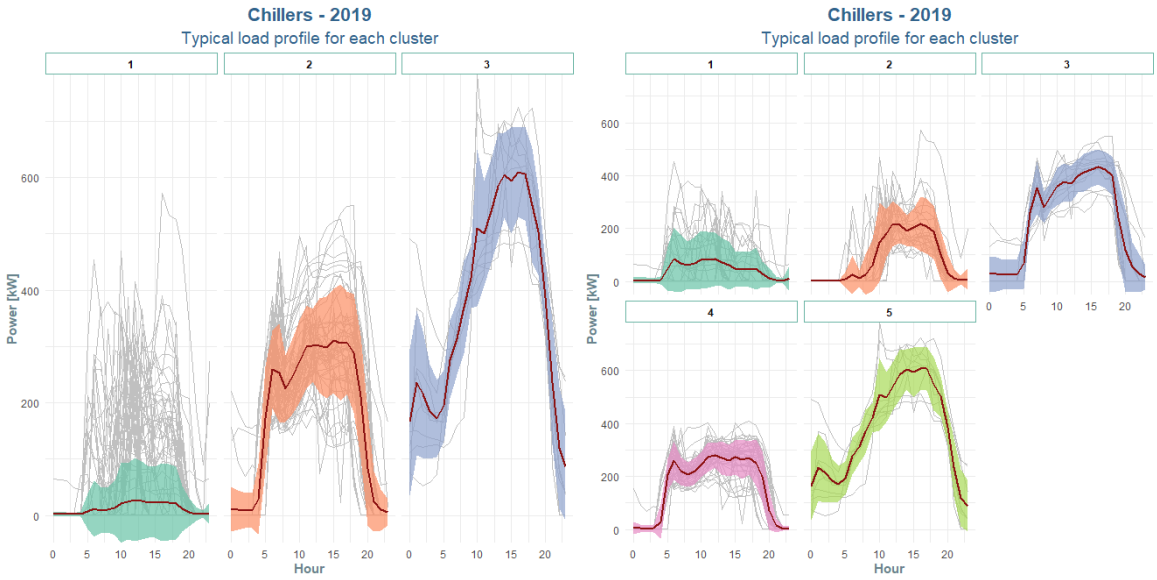

Figure 27: Typical load profiles for the chiller unit in the case of 3 (left) and 5 clusters (right).

The case with 3 clusters shows a classifier with the highest performance but this is due to the fact that each one of the clusters (and so, of the profiles) are associated to a larger number of days with a consequent ease to be predicted. The load profiles have been considered insufficiently representative and the choice was made in favor of 4 clusters that allow better profiles and good classification performance. The case with 5 clusters, instead, presents a problem of over-fitting: the profiles are too precise and they are associated to too specific boundary conditions, causing a decrease of the efficiency of the classification.

5.2.3 Baseload analysis

This section of the work deals with the results of the analysis on the baseload, described in the section 3.2.1 of the methodology. In the first part, the outcomes are summarised for the I3P building and then, in the second part, for the chiller unit. In particular, the results include the normal ranges of power for each cluster, the values of the KPI and the quantification of the energy savings after the simulated improvement.

I3P building For the I3P building, the baseload is intended as the consumption during non-occupancy hours: 24 hours for holidays and 13 hours for working days (between midnight to 8 a.m and from 8 p.m. to midnight).

The Figure 28 shows, for the time range of interest, the distribution of the average power for each cluster by means of a boxplot, whose main paramenters, expressed in kW, are summarised in Table 8.

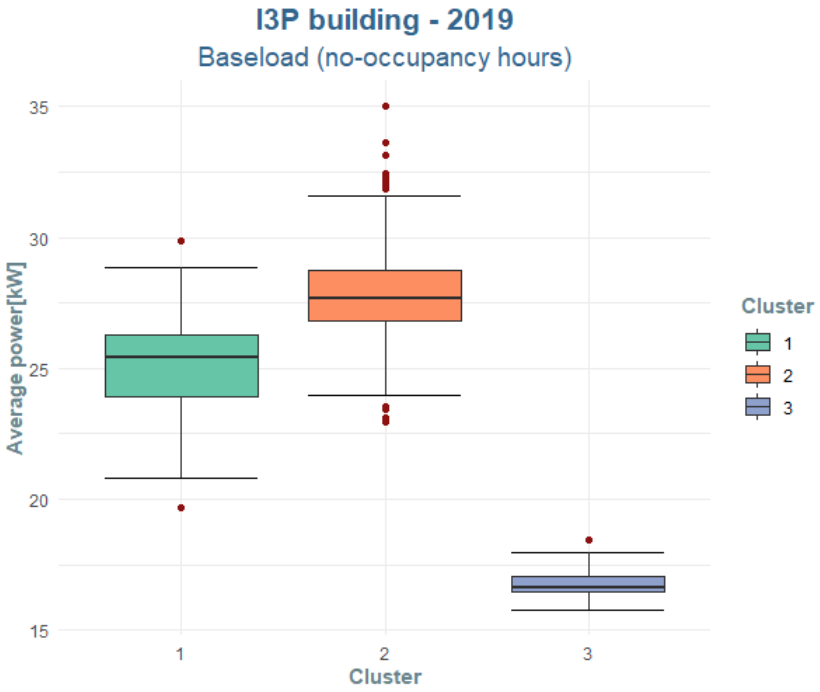

Figure 28: Ranges of the baseload average power for the I3P building.

	Cluster 1	Cluster 2	Cluster 3
Minimum	20.79	23.98	15.76
1st Quartile	23.91	26.83	16.45
Median	25.39	27.66	16.64
2nd Quartile	26.28	28.75	17.07
Maximum	28.86	31.62	17.98

Table 8: Values of the boxplot for the baseload average power of the I3P building.

The Cluster 2, the one corresponding to working-days, presents a power range with the highest value even if it is referred to hours during which the offices are empty: a median of about 28 kW during period in which the building is not used. In addition, there are days (corresponding to red dots) whose power request is even bigger, reaching values higher than 33 kW.
This situation affects the value of the KPI that is shown for each day in Figure 29.

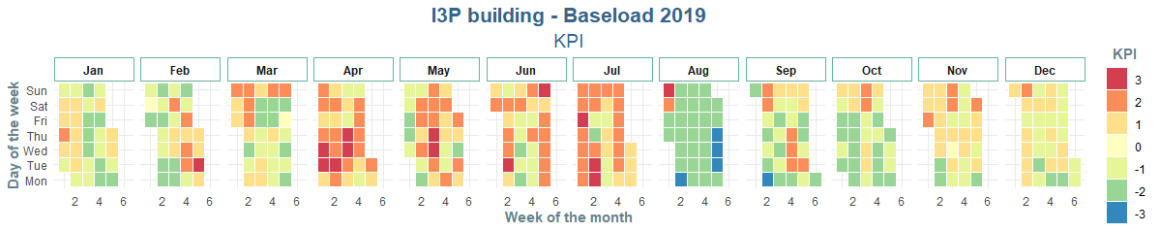

Figure 29: Values of the KPI for the I3P building.

The worst days are those with an high KPI, equal to 2 or 3, and they are concentrated mostly during months from April to July. The best month, instead, is August with a KPI that assumes negative values for almost all the time: the cause is the summer break, during which the offices are empty and mostly of the electrical devices are switched off.
In order to better understand the causes of this behaviour, the Figure 30 shows the power curves of the I3P building during April that is the month with more *worst days*: they are 5 and they are highlighted in the figure with a red rectangle. It can be noticed that during those days, the power, after the end of the working-hours, remains at high values without particular decrease with respect to occupancy hours. This might be due to the fact that the electric devices in the offices are left switched on, even if the scheduled working-time is ended.

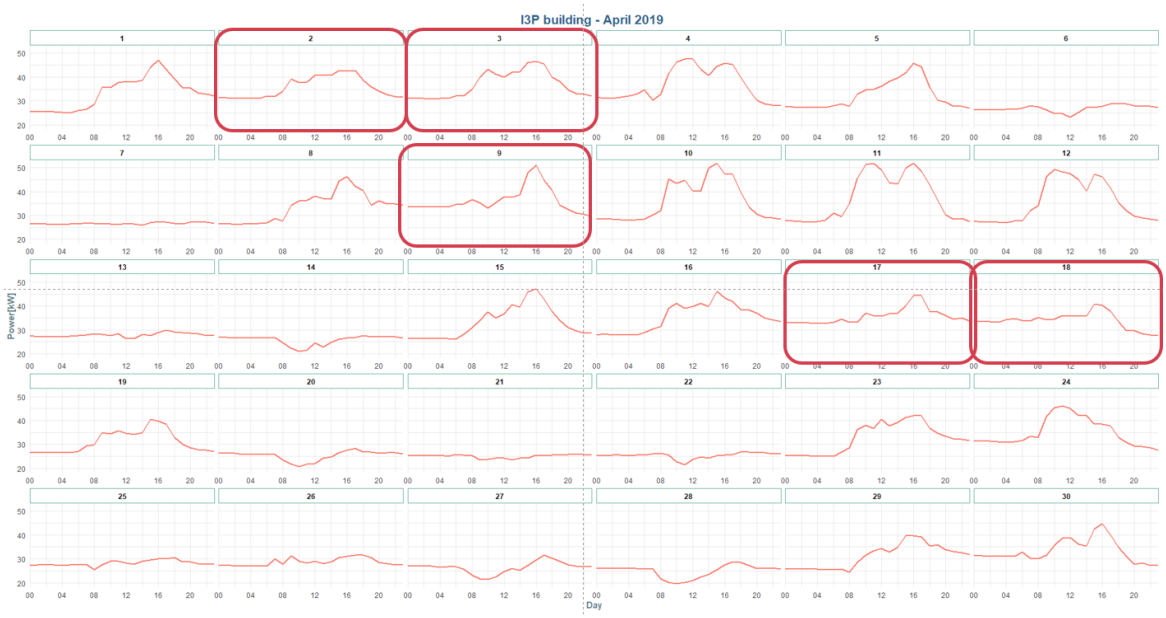

Figure 30: Power curves of the I3P building in April 2019 with highlighted days with worst KPI.

Finally, the last result regards the simulation of an improvement of the consumption during non-occupancy hours. As described in the methodology, the power of the days with an high KPI (i.e. equal to 2 or 3) is substituted with the average value of the corresponding cluster.

This operation allows to evaluate the potential baseload energy saving: from an yearly actual value of about 140.1 MWh, it is possible to decrease the demand up to 136.9 MWh that corresponds to a saving of the 2.3%.

These results, that are summarised in Table 9, can be visualised with the pie-chart in Figure 31.

Actual Demand [MWh]	**Improved Demand** [MWh]	**Saving** [MWh]	**Saving** [%]
140.1	136.9	3.2	2.3

Table 9: Summary of the results of the improvement for the I3P building.

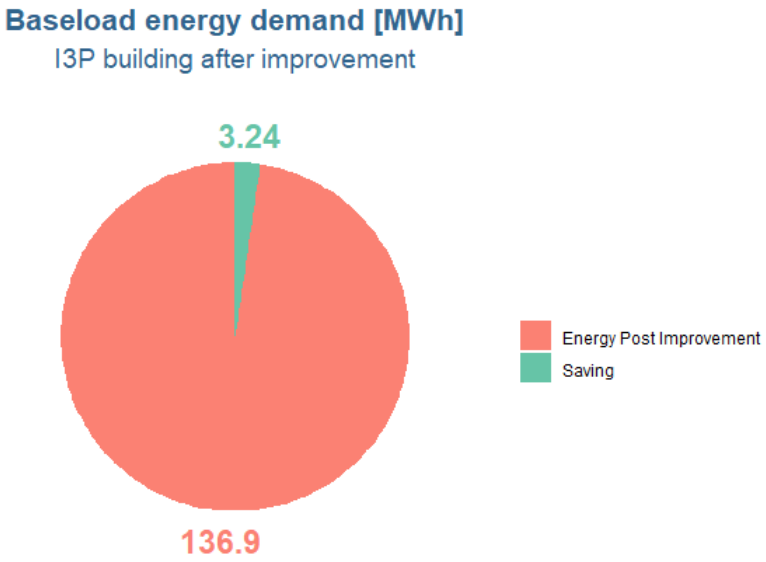

Figure 31: Yearly energy demand and saving post improvement for the I3P building in 2019.

Chiller unit For the chiller unit, the baseload is intended as the consumption during the *Off-peak hours*: 24 hours for holidays and 8 hours for the other days (between 11 p.m. to 7 a.m., from Monday to Saturday).

The distribution of the average power is represented in Figure 32 and the main parameters of the boxplots, in kW, are summarised in Table 10.

As it can be noticed, the Clusters 1 and 2 are characterised by very tight ranges, with low values of average power. It is the case of winter days and, in general, of days with a not high temperature (e.g. spring and fall): in these periods chillers are switched off during nights and holidays, so the only consumption is the one associated with the antifreeze circuit.

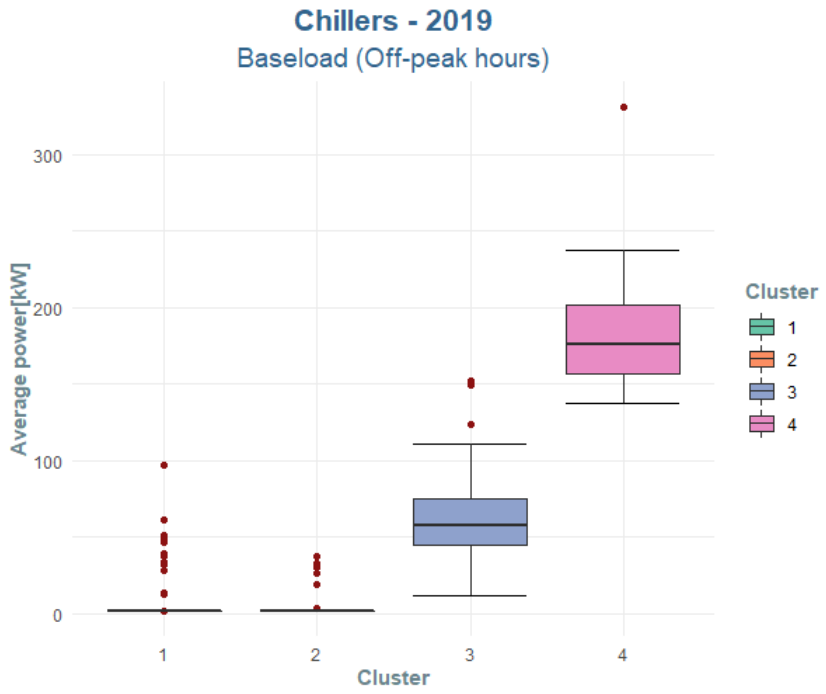

Figure 32: Ranges of the baseload average power for the chiller unit.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Minimum	1.43	1.45	11.31	137.65
1st Quartile	1.45	1.46	44.30	156.10
Median	1.46	1.46	57.34	176.16
2nd Quartile	1.48	1.49	74.69	201.49
Maximum	1.49	1.49	110.75	237.45

Table 10: Values of the boxplot for the baselod average power [kW] of the chiller unit.

Another point that has to be highlighted is the behaviour of the Cluster 4 and its power distribution. In fact, it is characterised by very high values of power, comparable with the daytime ones: the maximum power reaches values of about 238 kW, with a day in which it is even more than 300 kW.

This behaviour can be explained observing the Figure 33 which represents the power and temperature evolution during 7th and 8th July 2019; in particular this last day

is a Monday with very high external air temperature and belonging to Cluster 4. As it can be seen, the chiller unit is switched on the night before at 10 p.m., 11 hours before the starting of the working hours and 7/8 hour before the usual ramp up. This is done in order to avoid situations of discomfort in the building during the next day, because of the external temperature.

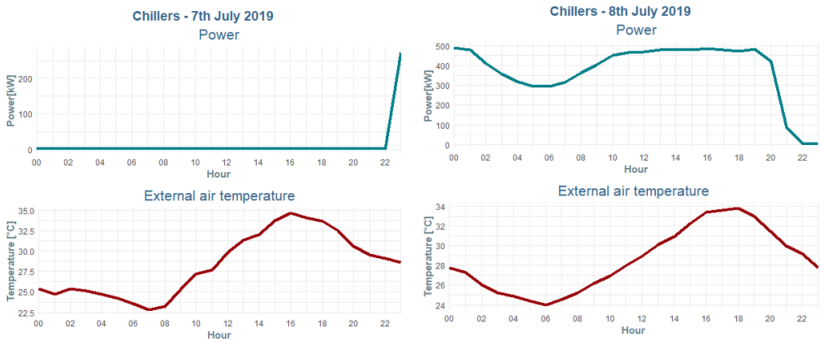

Figure 33: Power curve of the chiller unit during 7th and 8th July 2019.

Another explanation of the high consumption of the Cluster 4 can be done observing the Figure 34 which represents the power curves and the air temperature trends for the Sundays of June 2019.

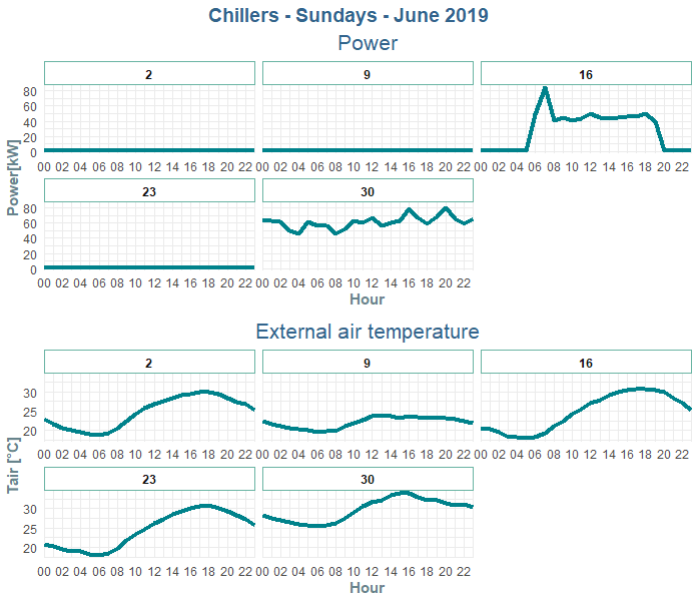

Figure 34: Power curve of the chiller unit during Sundays of June 2019

Firstly, it is evident that the external air temperature is comparable for all the days, in terms of both trend and values. However, two Sundays show an unusual behaviour, with values reaching 80 kW, completely different from the typical Sunday profiles that is flat. In fact, 16th and 30th July belongs to Cluster 4.
Considering this behaviours of the power profiles belonging to the Cluster 4, it has been chosen to classified as anomalous all the 8 days of the cluster. In fact, they are considered not acceptable and widely improvable: as a consequence, the worst and highest value of the KPI (i.e. 3) as been manually assigned.
After these considerations, the resulting daily values of the KPI are shown in Figure 35.

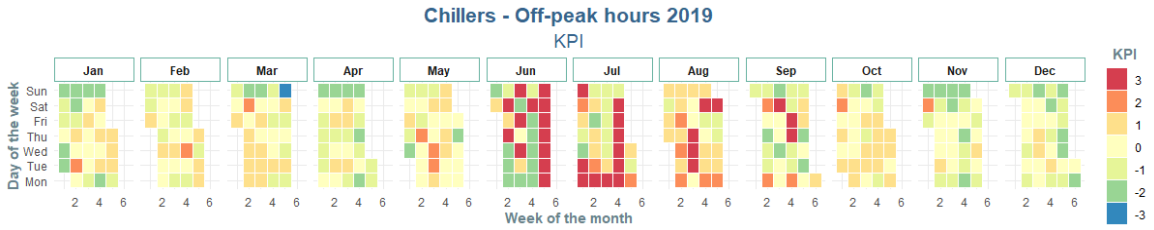

Figure 35: Values of the KPI for the chiller unit

As it can be noticed, the majority of the worst days (the ones with a KPI = +3) are in June and July that are hot months with an high level of occupancy of the buildings: the thermal comfort of the occupants has to be guaranteed and the chillers work at high power with anticipated ramp ups.
The winter days, instead, are characterised by low and almost constant values of power because only the antifreeze circuit works; the different values of the KPI for this period depend exclusively on the slightly different working pattern of the circuit, remaining in very tight power ranges. Because of this reason they are not considered anomalous or unacceptable and, in the improved scenario, their value is kept unchanged.

Regarding the simulation of improvement for the chiller unit, the following substitution criteria have been applied:

- **Cluster 3:** the power of the days with KPI=2 and KPI=3 is substituted with

the average value of the cluster;

- **Cluster 1 and 2 :** the power of the days with KPI=3 is substituted with the average value of the cluster, while the one with KPI=2 are left unchanged;
- **Cluster 4:** the power of all the days is substituted with the average value of Cluster 3.

The simulated improved scenario allows to evaluate the potential yearly energy saving during nights and holidays. Unlike the I3P building, in this case, the detected energy waste is a huge fraction of the actual consumption: from an yearly actual value of about 52.7 MWh, it is possible to decrease the consumption up to 32.8 MWh, corresponding to an energy saving of the 37.8%.
These results are summarised in Table 11 and they can be visualised by means of the pie-chart in Figure 36 that gives, at a first sight, the idea of the potentiality of the improvement.

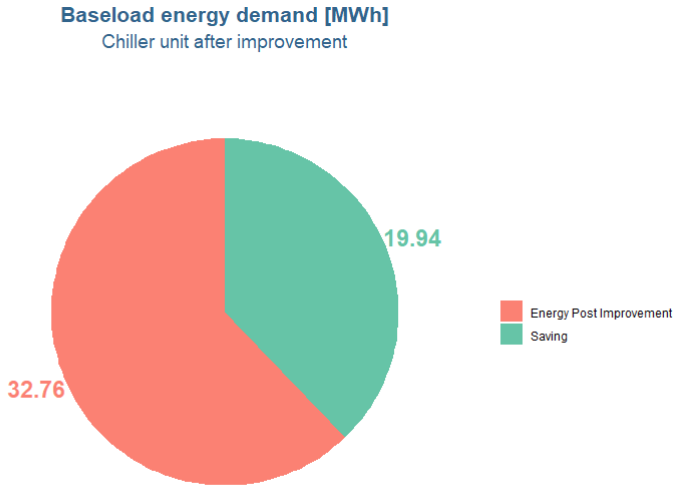

Figure 36: Yearly energy demand and saving post improvement for the chiller unit in 2019.

Actual Demand [MWh]	Improved Demand [MWh]	Saving [MWh]	Saving [%]
52.7	32.8	19.9	37.8

Table 11: Summary of the yearly results of the improvement for the chiller unit.

5.3 Production-level analysis

This section deals with the results of the production-level analysis that has been described in subsection 3.3. More specifically, the first subsection includes an overview of all the datasets that have been tested during the modelling phase, in order to find the final configuration that is described in the second subsection; finally, the last part of the section deals with the results of the anomaly detection and predictive maintenance procedure.

5.3.1 Tested data configurations for the model

In the following paragraphs there is the description of all the data configurations that have been tested during the development of the forecast model. In fact, in order to find the best solution which allows a satisfying prediction of the production of the photovoltaic plant, different meteorological data are taken from different sources and used to develop models that predict the production of a single pitch or of the total plant. The input meteorological data that have been considered, and that will be further explained, are:

- Data from the on-site meteorological station;
- Data from Solcast;
- Data from mixed sources.

Data from on-site meteorological station. As a first attempt, data from the on-site meteorological station are used with their original time granularity (i.e. 15-minutes samples). The inputs of the model are:

- Global horizontal irradiance in $[\frac{kW}{m^2}]$;
- External air temperature in [°C].

Regarding the output, two models are firstly built: one to predict the power of the east pitch and the other for the west one.
With this time granularity, for the east pitch 96015 measurements are available between the 2018 and the 2021 and the dataset has been divided in this way: the training set is composed of 61245 data (about 64.4 %) while the testing one includes the remaining samples (69.6 % of the total) consisting in January and February 2021 and from January to October of 2020. For the west pitch, instead, 92040 records can be used: the months from October to December 2019 (33.7 %) are used to test the model, while the remaining 66.3 % of the total dataset (61043 data) is used in the training phase.
The Table 12 summarises the main metrics that have been computed to evaluate the performance of the models, during both the training and the testing phase: as it can be noticed, the Mean Square Error and the Mean Absolute Error are sufficiently low but the Mean Average Percentage Error is characterised by too high results and so a further exploration of the results is carried out.

	MSE $[kW^2]$		**MAE** $[kW]$		**MAPE** [%]	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
East pitch	7.081e-03	7.447e-03	0.5692	0.5871	49.27	59.45
West pitch	8.089e-03	6.206e-03	0.5663	0.5023	76.79	82.18

Table 12: Metrics of the model with on-site meteorological data with 15-minutes aggregation.

In fact, even if the metrics have acceptable values, an additional visualization of the results shows that the models are not able to well predict the production of the pitches: considering the *scatter plots* in Figure 37, it can be seen that the points are not all along the diagonal line (that corresponds to a perfect prediction) and the linear correlation results in a slope that is lower than one.

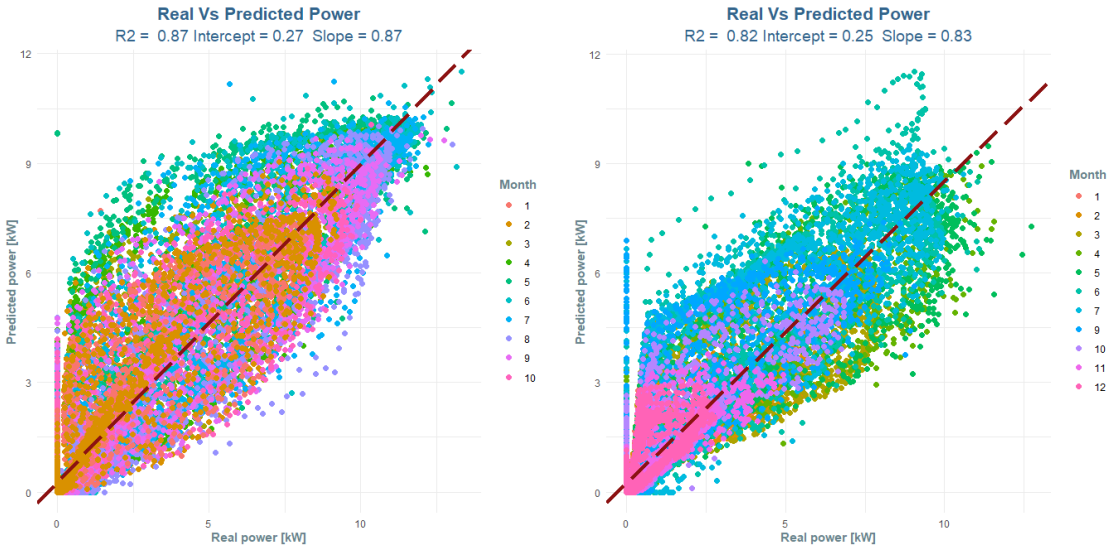

Figure 37: Real Vs Predicted Power for the east pitch (left) and for the west pitch (right)

There is not a particular difference among the months of the year, but a different visualization of the results can give a general explanation of the poor prediction capability. In fact, the Figure 38 shows, for the east pitch, the predicted value of the real power as a function of the hour of the day of each month belonging to the testing set: the values that are worse predicted, and that are far from the diagonal line in the plot, are those corresponding to the beginning or to the end of the day. During these hours, the sun is low on the horizon and this suggests that a parameter that has to be considered is the position of the sun in order to take into account the orientation of the pitches.

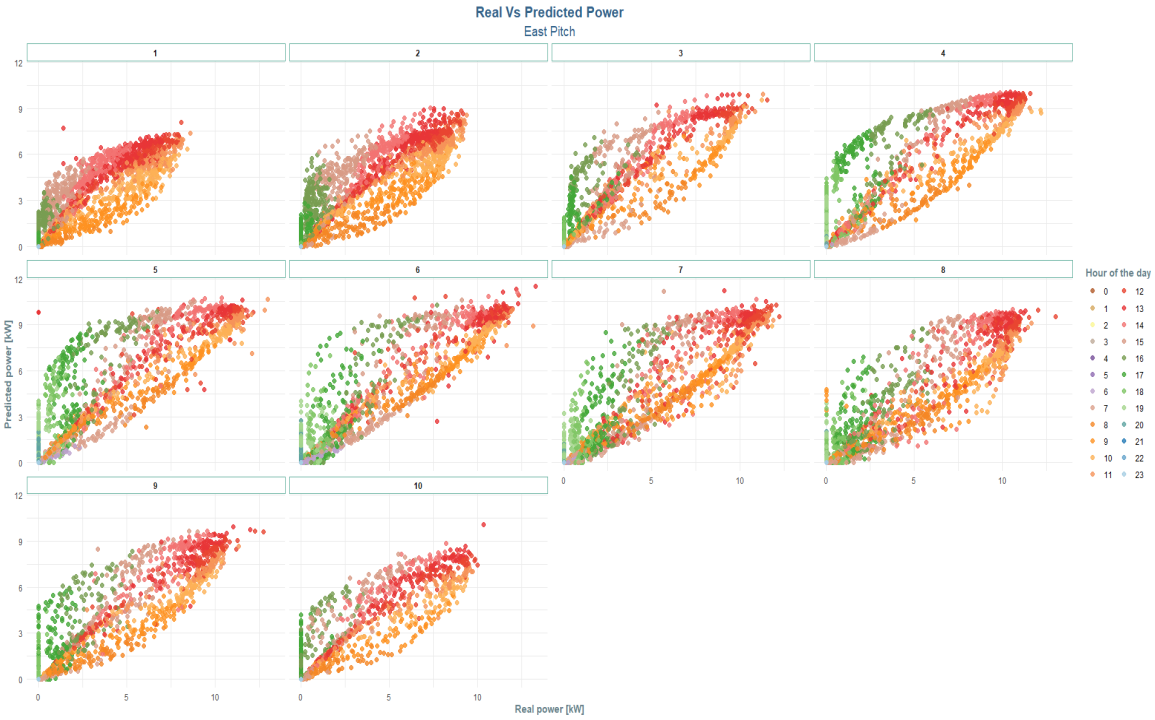

Figure 38: Real Vs predicted power of the east pitch by month.

At this point, before changing the input parameters of the model, another data configuration, consisting in the same input and output data but with a different time granularity, is used: an hourly aggregation is chosen in order to verify whether a lower level of detail can be better in terms of prediction, avoiding too much fluctuation of the measurements.

Reducing the time granularity and averaging the measurement, the number of available data decreases: 24020 record for the east pitch and 23023 for the west one; these datasets are divided in training and testing sets with the same percentages of the previous case.

However, this attempt do not give satisfying results, as it is shown by the comparison of the Figures 39 and 40 which represent the predicted (blue line) and real (yellow line) power curves in April 2020: the former refers to the model with 15-minutes samples, while the latter is the result of the model with the higher data aggregation. The higher aggregation of the data does not improve the results of the model but, on the contrary, during some days (e.g. 28th of April) the prediction is noticeably worse.

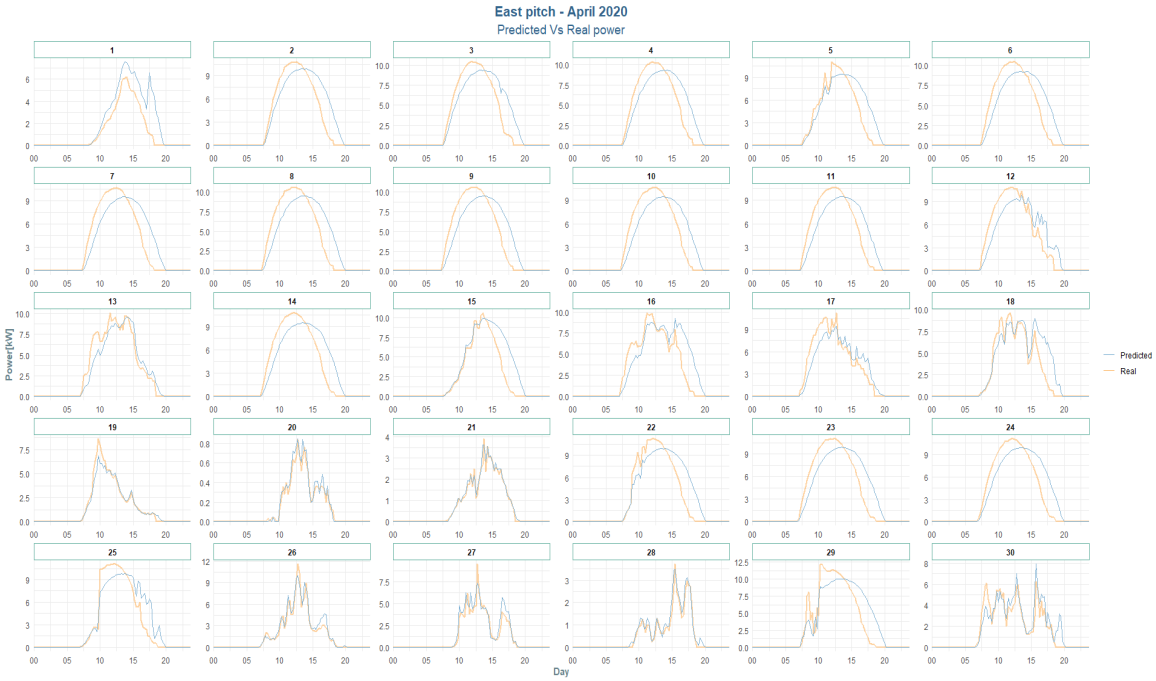

Figure 39: Real and predicted power of the east pitch in April 2020 with 15-minutes aggregated data.

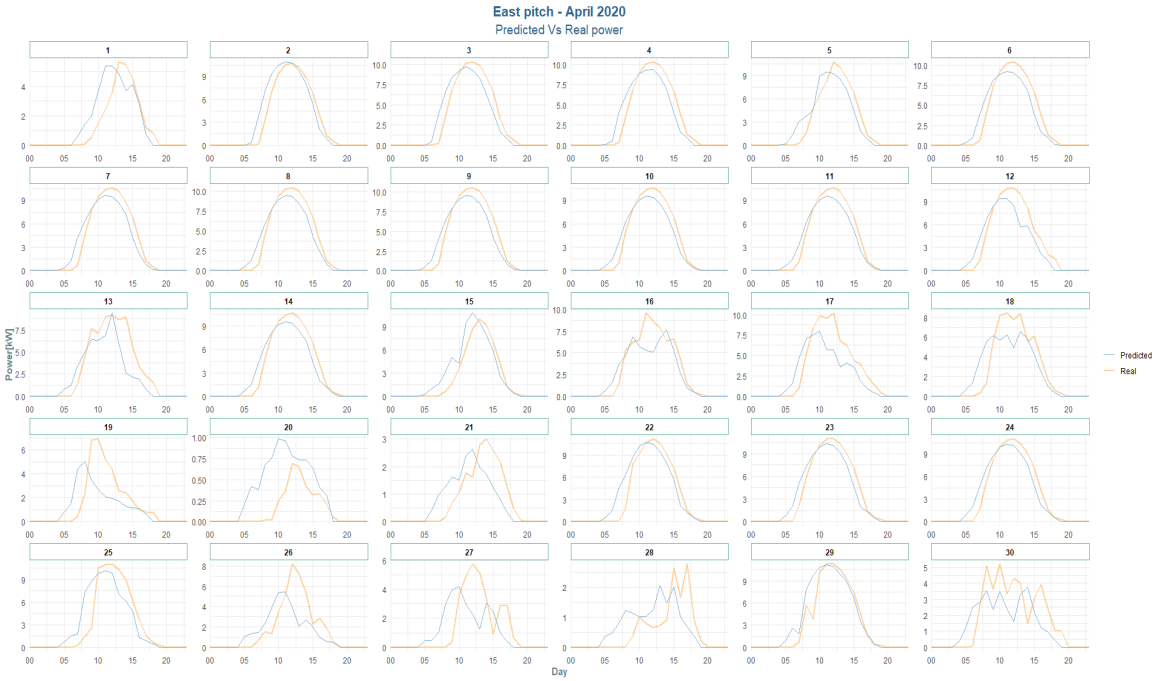

Figure 40: Real and predicted power of the east pitch in April 2020 with 1-hour aggregated data.

The worsening of the performance can be noticed also by the values assumed by the computed metrics, summarised in Table 13: all of them show an increase with respect to the previous case and in particular the MAPE reaches values that are much higher than 100%.

	MSE $[kW^2]$		**MAE** $[kW]$		**MAPE** [%]	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
East pitch	11.84-03	10.69e-03	0.7186	0.6665	266	219
West pitch	9.311e-03	8.434e-03	0.9516	0.8806	464	566

Table 13: Metrics of the model with on-site meteorological data with 1-hour aggregation.

Data from Solcast. After the above-mentioned first attempt, another set of data is used to try to overcome the limits of the previous set taking into account the position of the sun during the day. The second set of input data is taken from Solcast (see subsection 4.1.1) and it consists in:

- External air temperature in [°C];
- Global Horizontal Irradiance in $[\frac{W}{m^2}]$;
- Diffuse Horizontal Irradiance in $[\frac{W}{m^2}]$;
- Direct Normal Irradiance in $[\frac{W}{m^2}]$;
- Direct Horizontal Irradiance in $[\frac{W}{m^2}]$;
- Solar Zenith in [°];
- Solar Azimuth in [°];
- Cloud Opacity in [%].

This input dataset is used firstly with its original time granularity and then with a more aggregated one (15-minutes and 1-hour samples, respectively) in order to develop separate models to predict the power production of the east and west pitch, as for the previous case.

Differently from the previous meteorological data source, Solcast allows to have at disposal a larger number of data with a consequent more complete dataset for the model. In fact, for the case with lower data aggregation the available records are 101900 for the east pitch and 104526 for the west one that are both divided between training and testing set with percentages of about 70% and 30% respectively. Instead, the hourly aggregated set consists in 24020 measures for the east pitch and 23023 for the other one divided in about 65% for the training and the remaining 35% for the testing phase.

In Tables 14 and 15 there is a summary of the metrics calculated for each model.

	MSE $[kW^2]$		**MAE** $[kW]$		**MAPE** [%]	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
East pitch	5.907e-03	6.606e-03	0.4939	0.5123	91.17	74.27
West pitch	5.462e-03	5.851e-03	0.3777	0.3941	68.09	82.94

Table 14: Metrics of the model with data from Solcast with 15-minutes aggregation.

	MSE $[kW^2]$		**MAE** $[kW]$		**MAPE** [%]	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
East pitch	4.2e-03	4.53e-03	0.3735	0.7474	72.17	53.98
West pitch	3.154e-03	3.017e-03	0.4818	0.9357	81.69	119

Table 15: Metrics of the model with data from Solcast with 1-hour aggregation.

From the values of the metrics it can be noticed a further improvement of the performance with respect to the previous input dataset, in terms of Mean Square Error and

Mean Absolute Error; the Mean Average Percentage Error, instead, reaches higher values in particular for the west pitch with hourly data, exceeding 100%.
Even if the first two metrics have acceptable values, the visualization of the results, in terms of comparison between real and predicted power, suggests that even this model is not able to give a fulfilling forecast with neither of the time granularity.
As an example, the results are shown for the east pitch with 15-minutes data.

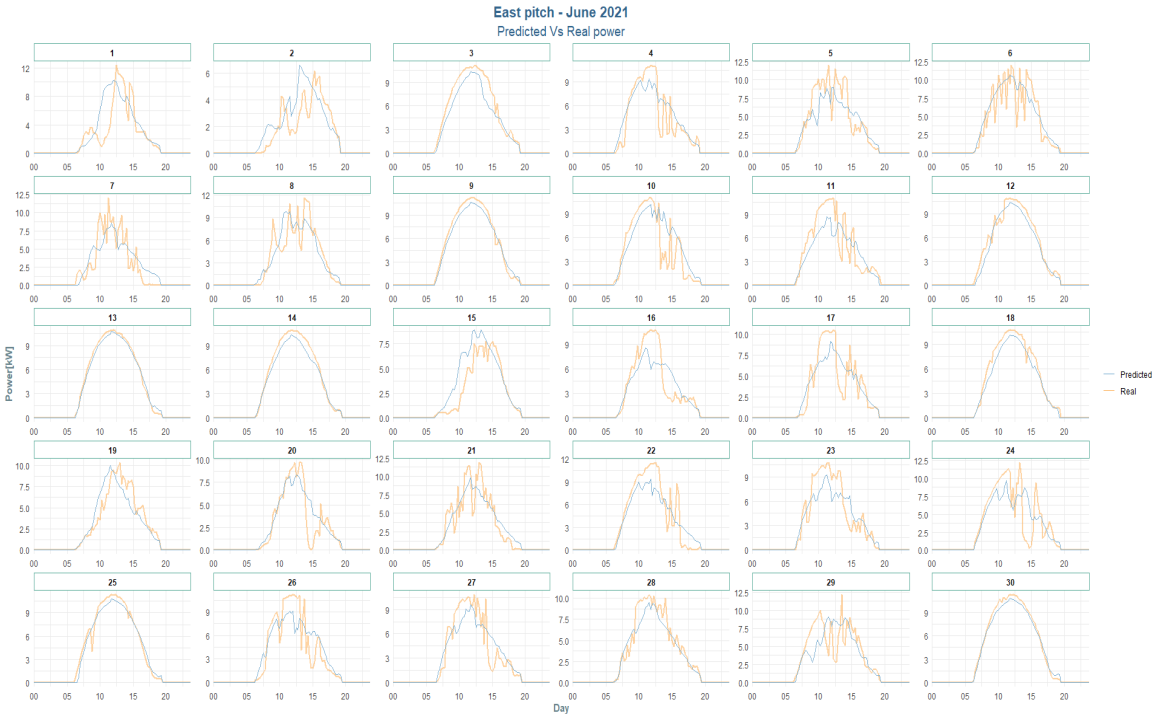

Figure 41: Real and predicted power of the east pitch in June 2021 with 15-minutes aggregated data

The Figure 41 shows the comparison between the real (yellow line) and predicted power (blue line) for June 2021: in clear sky conditions (e.g. 13th June) the two curves are almost overlapping and so the model is able to perform a realistic forecast of the production. However, the model reveals a lower prediction capability in those days in which the shading causes a fluctuation of the production.
As for the previous case, the results of the model can be visualised, by means of a *scatterplot*, in terms of predicted values of the real power by month, highlighting the different hours of the day.

The Figure 42 shows that, in this case, the prediction quality is worse during the central hours of the day: the problem related to the position of the sun is solved but both the figures suggest that the issue is linked with the information about radiation. In fact, Solcast provides meteorological data (both irradiance and cloud opacity) that are estimated and not measured, with a consequent insufficient ability to well describe the real behaviour of the irradiance on the pitches of the plant.

Figure 42: Real Vs predicted power of the east pitch by month.

Data from mixed sources. The above-described previous modelling attempt allowed to understand the necessity of information about the position of the sun, but also the importance of taking advantage of the on-site measurements about the irradiance in order to account for realistic shading issues on the PV plant.

As a consequence, a new dataset has been constructed taking data from both the available sources, considering the measurements of the on-site meteorological station and integrating them with data from Solcast.

More specifically, the dataset consists in the following parameters:

- External air temperature [°C] from the on-site meteorological station;
- Global horizontal irradiance [$\frac{W}{m^2}$] from the on-site meteorological station;
- Diffuse horizontal irradiance [$\frac{W}{m^2}$] from Solcast;
- Direct normal irradiance [$\frac{W}{m^2}$] from Solcast;
- Direct horizontal irradiance [$\frac{W}{m^2}$] from Solcast;
- Solar zenith [°] from Solcast;
- Solar azimuth [°] from Solcast;
- Cloud opacity [%] from Solcast.

Since it was clear from the previous attempts that the hourly-aggregated data do not give particular advantage in the performance of the model, this new input dataset has been considered with only an high time granularity and so 15-minutes records have been used. However, an additional case has been considered in terms of output and three separate models have been developed to predict the power production for the east pitch, the west pitch and the total plant.
The number of available data for this dataset is slightly lower than the previous one because of the necessity of considering only time-steps during which all the parameters are accessible, and in particular for the total plant, the batch of input is smaller because in this case data from both east and west pitch must be available. For the east pitch 96015 samples are available and the 63.8% of them is used to train the model, for the west pitch, instead, 92040 records are at disposal and the training set includes about the 66% of the total; finally, for the whole plant the 61.7% of the complete dataset (566238 data) is exploited to train the neural network.
The resulting values of the metrics used to evaluate the performance of each model are summarised in Table 16

	MSE $[kW^2]$		**MAE** $[kW]$		**MAPE** [%]	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
East pitch	7.664e-04	9.035e-04	0.1520	0.1622	17.92	19.02
West pitch	2.91e-03	1.877e-03	0.2484	0.2143	24.45	25.59
Total plant	1.149e-03	9.737e-04	0.3012	0.2819	19.67	21.64

Table 16: Metrics of the model with data from mixed sources.

As it can be clearly noticed from the table, all the three models show an excellent performance in terms of all the metrics. The Mean Square Error decreases with respect to the other tested models reaching very low values, especially for the east pitch for witch it is about one order of magnitude lower than the case with only on-site meteorological data; the Mean Absolute Error has good low values for all the models, in both training and testing phase. Finally, the Mean Average Percentage Error shows sufficiently low values for the first time among all the attempts that have been done with different datasets: in facts, even if it is higher for the west pitch it can be still considered acceptable.

The positive performance of the models can be observed also considering the *scatter-plots* representing the predicted value of each real power data: the Figure 43 shows this information for each one of the models, highlighting the different months that are included in the testing dataset. As it can be noticed, the models for the east pitch and for the total plant show an almost perfect prediction capability, except for some points that are far from the diagonal line but that are a very small part of the total and might be the consequence of measurement errors of the real power. Regarding the model for the west pitch, it is the one with worse results (even though they are good enough) with respect to the others: in fact, it is the one with the lower slope and it is evident that there are points (light blue ones) that are predicted with a reduced value of power.

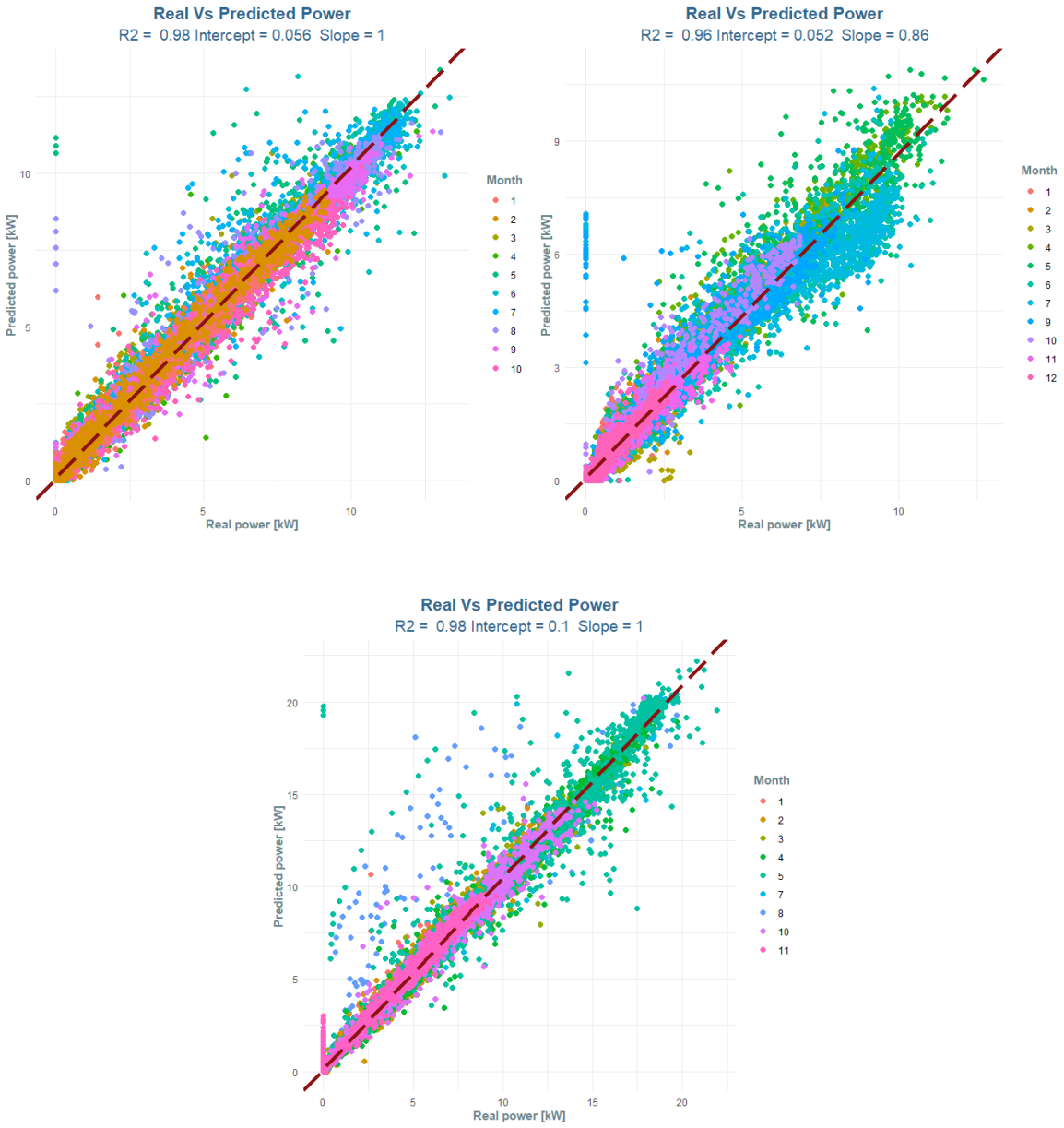

Figure 43: Real vs predicted power for the east pitch (up right), the west pitch (up left) and the total plant (bottom).

Since the model for the west pitch shows lower performances with respect to the others, further visual exploration of the results is carried out in order to find possible issues and limits of the dataset in exam.

The Figure 44 shows the comparison between the real power (yellow line) and the predicted one (blue line) in July 2019. As it can be noticed, the problem of this

forecast model is the prediction of the power when it has higher values, during the central hours of the day. However, the profiles result to be correct and quite precise and the inaccuracy regards only the magnitude of the power.

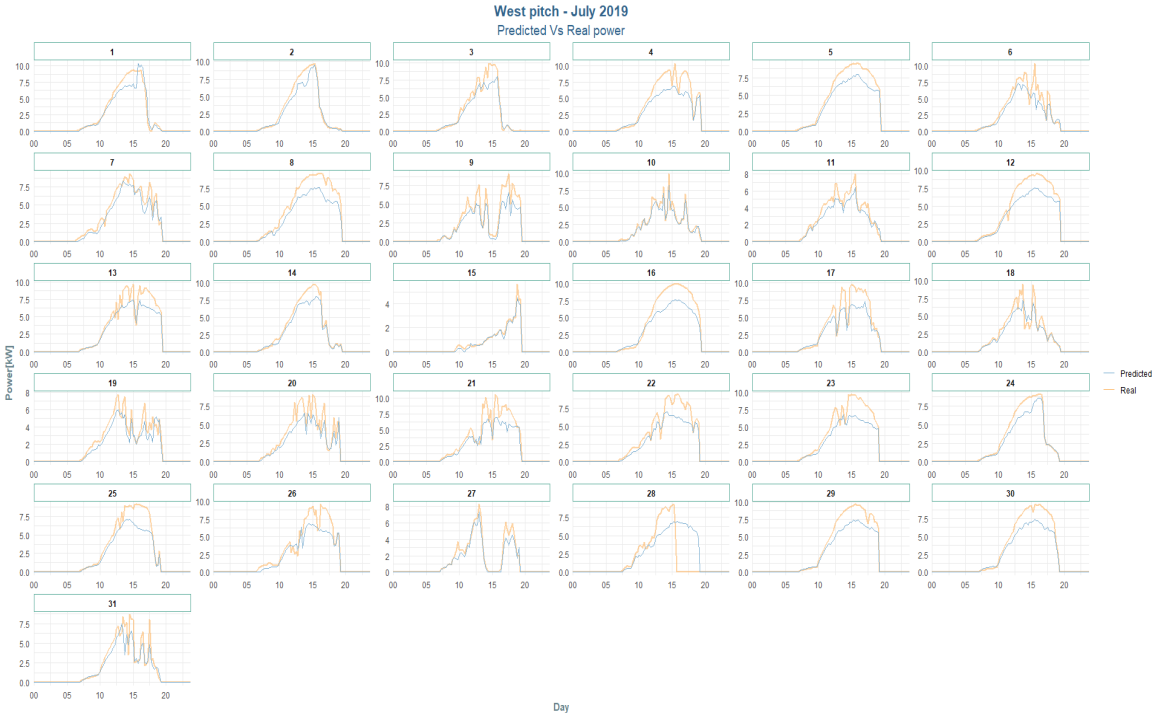

Figure 44: Real and predicted power of the west pitch in July 2019

An explanation for this behaviour has been found in the use of irradiance components from Solcast. In fact, as previously mentioned, they are estimated and not measured with a consequent possible inaccuracy with respect to real values; this might cause an incorrect learning of the model that might be due to an imprecision of the irradiance values associated to certain values of the real power production.

5.3.2 Final model

All the considerations of the previously described attempts result in the definition of the final dataset configuration for the model.

First of all, the dataset that has been used to develop the final model is composed of measurement and information with low time aggregation: 15-minutes data are considered.

Regarding the predicted output, three models have been constructed in order to forecast the electric production of the single pitches and the one of the total plant.
The inputs, on which the output depend, are taken from the two different data sources, described in section 4.1.1; in particular, the input dataset is composed of the following quantities:

- External air temperature [°C] from the on-site meteorological station;
- Global horizontal irradiance $[\frac{W}{m^2}]$;
- Solar zenith [°] from Solcast;
- Solar azimuth [°] from Solcast;

The measurements of external air temperature and global irradiance are taken from the on-site monitoring station in order to guarantee a realistic representation of the meteorological condition and take into account a reliable characterization of the shading on the system. The solar zenith and azimuth, instead, are selected to consider the sun's position in the sky in order to increase the prediction performance at the beginning and at the end of the day, when the sun is low on the horizon. Concerning the last two inputs, a consideration can be made: they are taken from Solcast, an on-payment data source, because it has been used to provide also other information for the other tested model. However, both solar zenith and solar azimuth can be obtained from open databases. making the construction of the model possible with all easily accessible and free data.
Even if the input parameters change, the final dataset is the same of the previous case (i.e. model with data from mixed sources) in terms of time period and, consequently, of amount of data. In fact, for the east pitch, the available data are 96015 and they are divided in 63.8% and 36.2% between training and testing, respectively; the set for the west pitch, instead, consists in 92040 samples whose 66.3% is used to train the model. Finally, the dataset referred to the total plant includes 566238 data: the 61.7% train the neural network, while the remaining 38.3% is used to test the model. The datasets have been created trying to select suitable time periods that allows the

model to properly learn periodicity and seasonality of the photovoltaic power production. For this purpose, among the available samples, data from different months and different seasons are selected to train and test the model on all the possible conditions.

In Table 17 there is an overview of the different months that are included on each dataset, distinguishing between training and testing sets of each one of the model.

	TRAINING SET	**TESTING SET**
East pitch	from Jan. to Oct. 2018 from Jul. to Dec. 2019 from Mar. to Jul . 2021	from Jan. to Oct. 2020 Jan. and Feb. 2021
West pitch	from Jan. to Dec. 2018 from Jan. to May 2020 from Mar. to May 2021	from Jan. to Jul. 2019 from Sep. to Dec. 2019
Total plant	from Jan. to Oct. 2018 Jul. and Aug. 2019 (partially) Sep. and Oct. 2019	from Jan. to May 2020 Oct. and Nov. 2020 Jul. and Aug. 2019 (partially)

Table 17: Months included in the datasets for the three forecast models.

This data configuration, in terms of time aggregation, time domain and input variables, has been chosen because it allows the development of a forecast model that shows an excellent capability to predict the power production of both the single pitches and the total plant.

In fact, the calculation of the error metrics results in satisfying values of MSE, MAE and MAPE for all the output powers. As it is reported in Table 18, most of the results are lower for the testing phase that for the training one and it suggests the absence of an over-fitting issue of neural network. The lowest values of the Mean Square Error and Mean Average Error are estimated for the east pitch, while the lowest Mean Average Percentage Error is reached with the model for the Total Plant. In general, all

metrics' values are considered sufficiently low and as indicators of a good performance.

	MSE $[kW^2]$		**MAE** $[kW]$		**MAPE** [%]	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
East pitch	8.25e-04	9.69e-04	0.151	0.172	17.42	19.75
West pitch	2.85e-03	1.54e-03	0.242	0.203	28.65	31.09
Total plant	1.62e-03	8.11e-04	0.259	0.224	16.56	16.53

Table 18: Metrics of the final model.

As it can be noticed from the comparison between the Table 16 and Table 18, the metrics have comparable values for both the final model and the one that consider also the other components of the irradiance.

However, the excellent performance of the final model can be evaluated by means of the observation of the power curves of the production that highlight the better prediction capability of the final model, with respect to the other one.

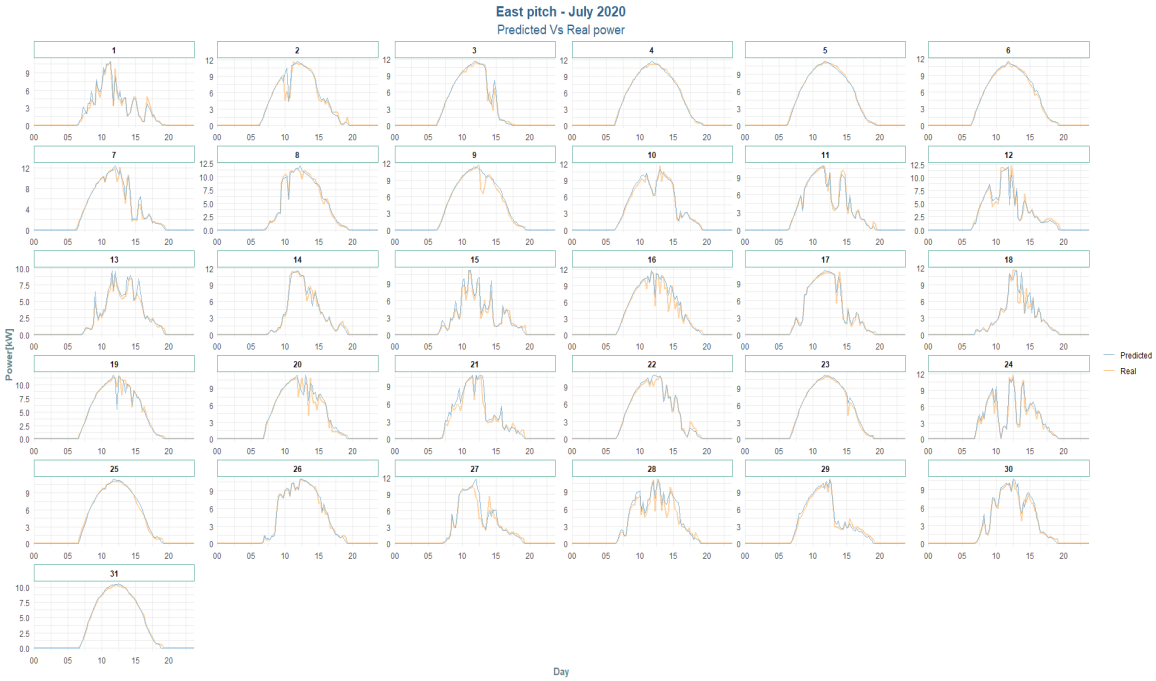

Figure 45: Real and predicted power of the east pitch in July 2020 with the final model.

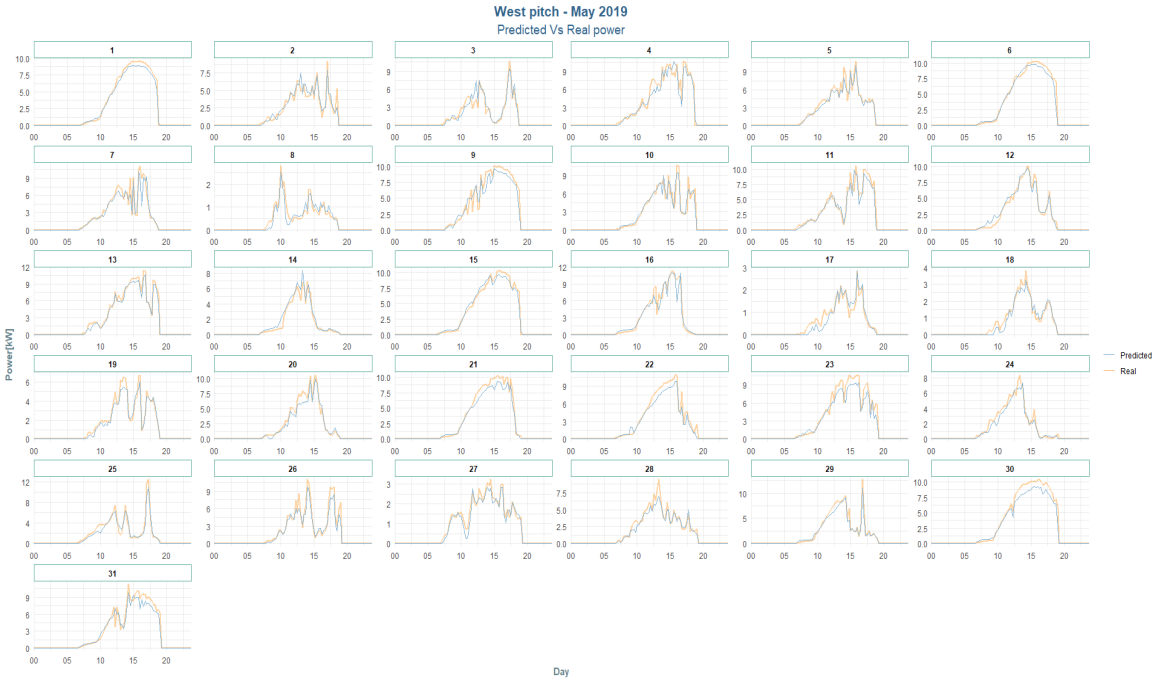

Figure 46: Real and predicted power of the west pitch in May 2019 with the final model.

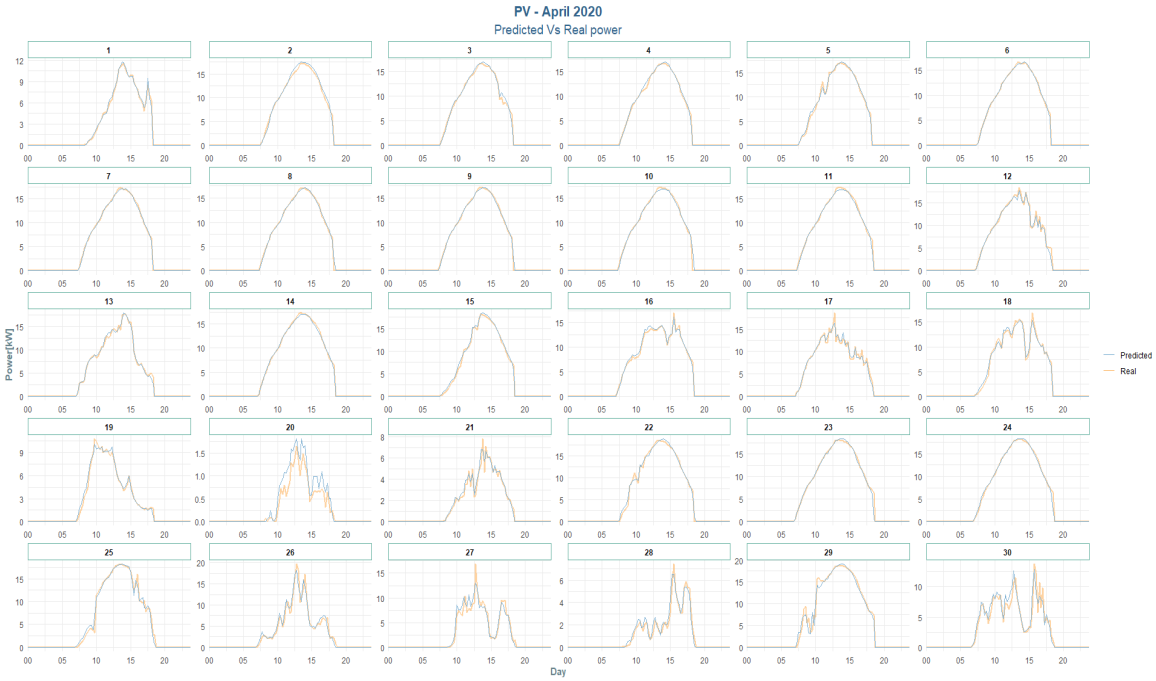

Figure 47: Real and predicted power of the total plant in April 2020 with the final model.

The Figures 45, 46 and 47 show the comparison between the predicted power (blue line) and the real one (yellow line): in particular, the first is about the east pitch during July 2020, the second regards the west pitch in May 2019 and the third figure represents the power of the total plant in April 2020.

The model for the east pitch is able to provide a a prediction of the power that is almost perfectly overlapping with the real one, not only in clear sky condition (e.g. 4th and 25th July) but also in shading conditions (e.g. 15th and 24th July).

The performance is satisfying also for the west pitch model: the real and predictive curves are really close with all the weather conditions. The forecast gives acceptable results for any magnitude of the power: both high power values (e.g. 15th May) and lower ones (e.g. 8th and 27th May).

The model is accurate also considering the total plant: the real and predicted curves can be considered coincident both in clear sky condition, with high values of power (e.g. 3r and 24th April), and in presence of shading, with different magnitudes of power(e.g. 17th, 28th and 30th April).

Finally, the quality of the performance of the model can be appreciated also observing the *scatter plots* in Figure 48 that represent the prediction of a certain value of the real power, for each one of the models. It can be noticed that the majority of data points are very close to the diagonal line, especially for the case of west pitch (up right) and total plant (bottom). Just few points considerably deviate from the condition of satisfying prediction, but they can be either associated to situation of malfunctioning of the system or neglected because of their small number.

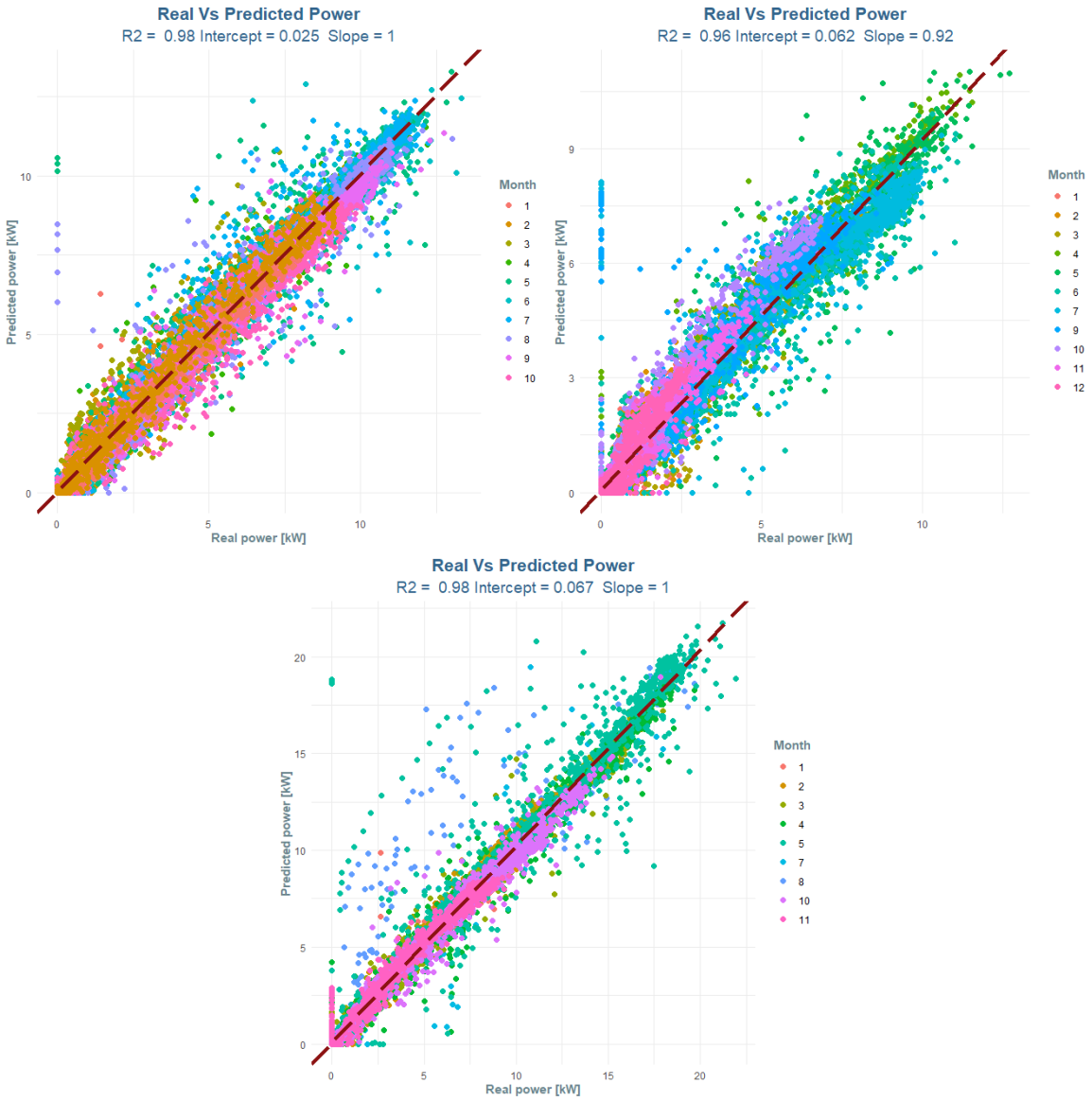

Figure 48: Real vs predicted power for the east pitch (up right), the west pitch (up left) and the total plant (bottom) with the final prediction model.

5.3.3 Anomaly detection

This section deals with the results obtained with the anomaly detection procedure, following the steps described in the methodology, in section 3.3.2.

Datasets for the anomaly detection The anomaly detection procedure is carried out, for each one of the physical domain (i.e. for each one of the models), considering

specific datasets that include relevant anomalous days and that are different from the ones used to train and test the corresponding neural networks.
In Table 19 the dataset is summarised for each model, in terms of considered months.

East pitch	November 2018 June 2019 December 2020 August and September 2021
West pitch	August 2019 October and November 2020 from January to March 2021
Total plant	July 2019 (partially) August 2019 (partially) from January to June 2021

Table 19: Datasets for the anomaly detection.

Sub-hourly residuals and comparison with thresholds The sub-hourly residuals are computed considering the 15-minutes samples and calculating the difference between the predicted power and the real one: positive residuals indicate situations in which the actual power is lower than the prediction.
The residuals are than compared to the lower and upper limit, as defined in the section 3.3.2 of the methodology, defined as multiples of the MAE calculated for positive powers in the testing phase of the ANNs.
The values of the Mean Absolute Errors and of both the limits are summarised in Table 20

	MAE (P>0) [kW]	Lower Limit [kW]	Upper Limit [kW]
East pitch	0.4703	1.18	2.35
West pitch	0.3971	0.995	1.99
Total plant	0.5336	1.34	2.67

Table 20: Values of the limits for the sub-hourly anomaly detection.

In the following figures the results of the sub-hourly anomaly detection are graphically shown: the Figures 49, 51 and 53 represent the computed residuals and their comparisons with the limits, indicated by the two dotted lines; the figures are referred to the east pitch (orange lines), the west pitch (green lines) and the total plant (blue lines), respectively.

The Figures 50, 52 and 54, instead, present the output values of the detection, for each one of the model.

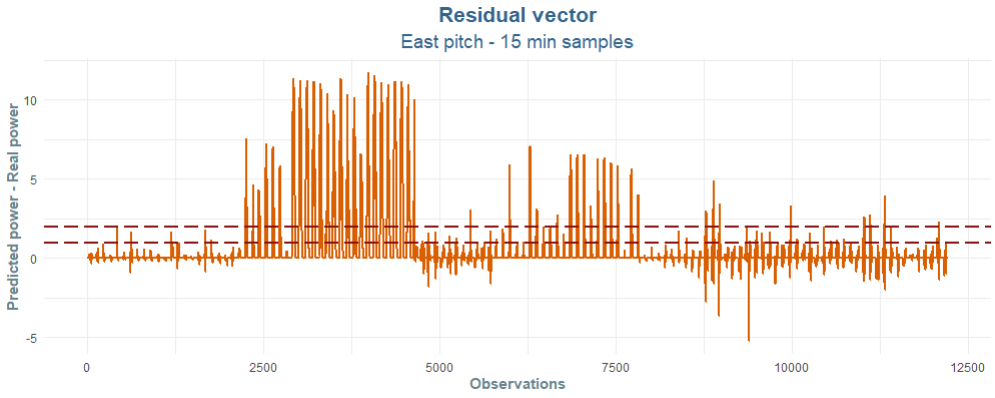

Figure 49: 15-minutes residuals and comparison with limits for the east pitch.

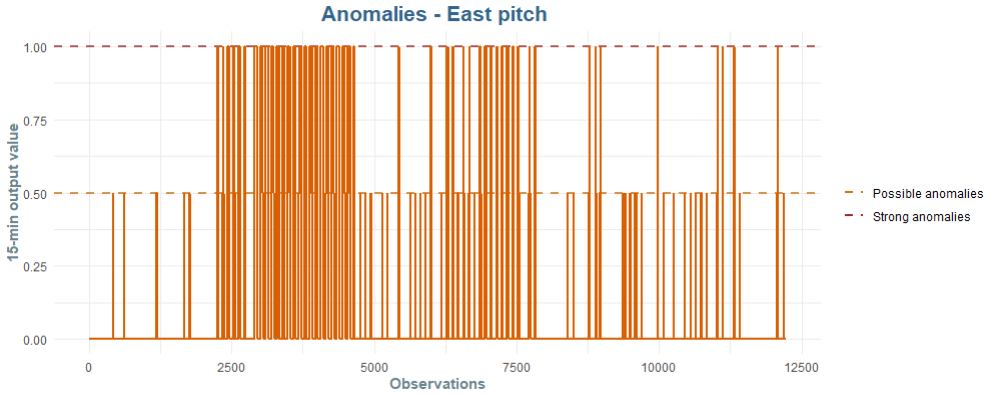

Figure 50: Output values of the sub-hourly anomaly detection for the east pitch.

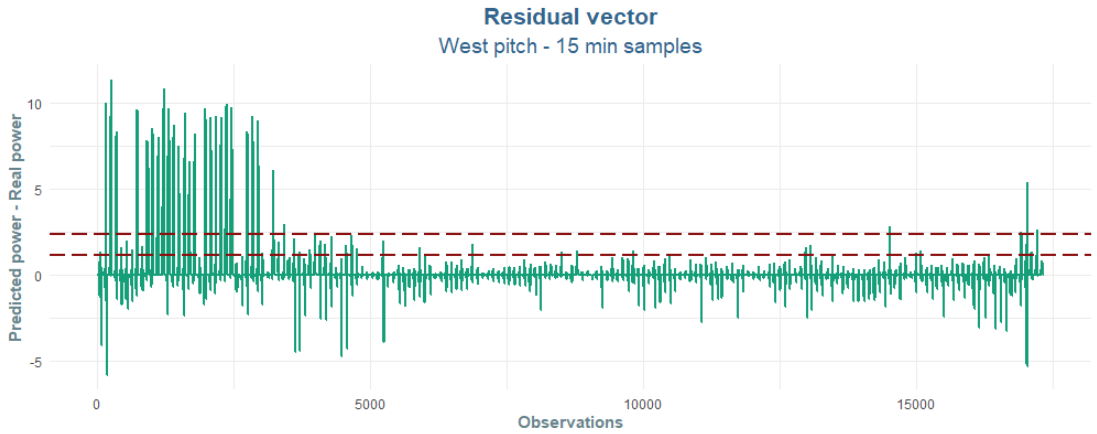

Figure 51: 15-minutes residuals and comparison with limits for the west pitch.

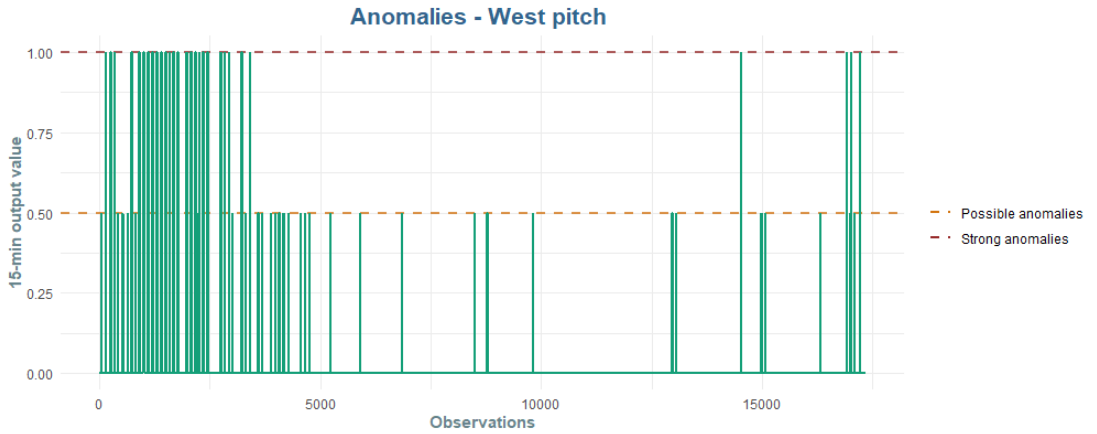

Figure 52: Output values of the sub-hourly anomaly detection for the west pitch.

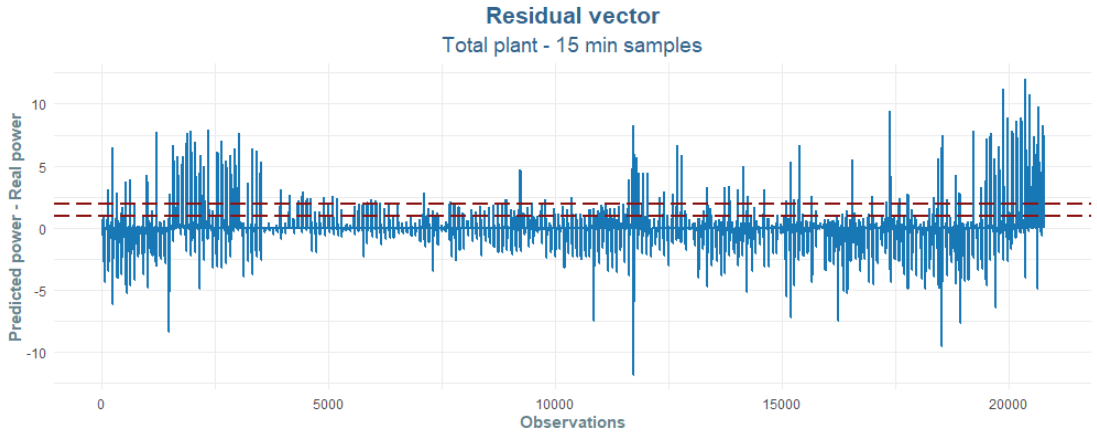

Figure 53: 15-minutes residuals and comparison with limits for the total plant.

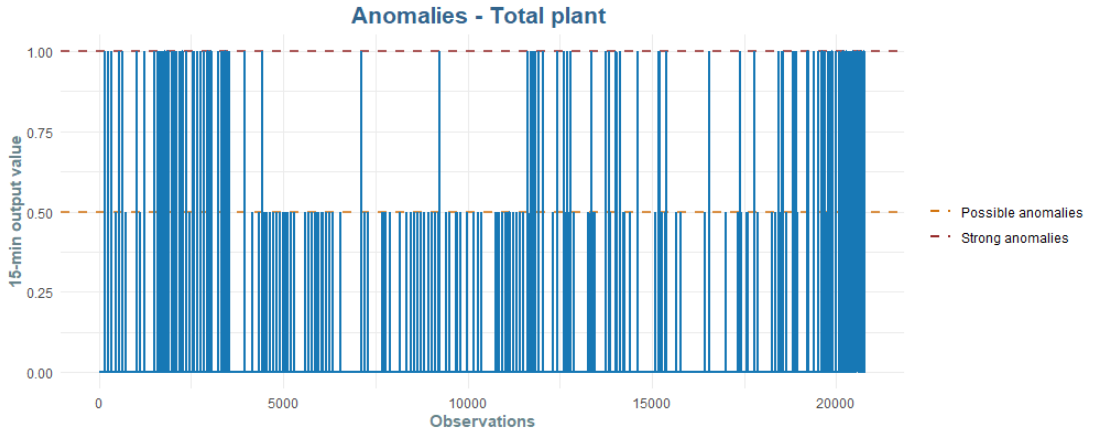

Figure 54: Output values of the sub-hourly anomaly detection for the total plant.

As it is described in the methodology, and as it can be noticed from the figures, the samples (i.e. the power values) are labelled with an output value depending on whether they represent strong anomalies, possible anomalies or normal operating conditions. In particular, the labelling criterion is the position of the residuals with respect to the limit value: if they are lower than the LL, the observation is considered normal ad the value 0 is assigned. Instead, if the residual is between the LL and the UL, the sample is a possible anomaly and it is labelled with the value 0.5. Finally, if the difference between the predicted and real power is higher than the UL, this situation represents a strong anomaly with a consequent output value equal to 1.

Daily anomaly detection After the sub-hourly anomaly detection, the residuals are aggregated on a daily basis: during this operation, the previously obtained outputs are summed up for each day and the resulting value is compared with the two daily thresholds τ_U and τ_L, that are summarised in Table 21 for the east and west pitch and for the total plant.

	Lower threshold τ_L	**Upper threshold** τ_U
East pitch	3	9
West pitch	3	10
Total plant	5.5	9

Table 21: Thresholds for the daily anomaly detection

As it has been done in the previous paragraph, the results are graphically shown in the following figures for the east pitch (orange dots), west pitch (green dots) and total plant (blue dots). In particular, the Figures 55, 57 and 59 present the daily values of the sums, corresponding to daily outputs that are represented in Figures 56, 58 and 60.

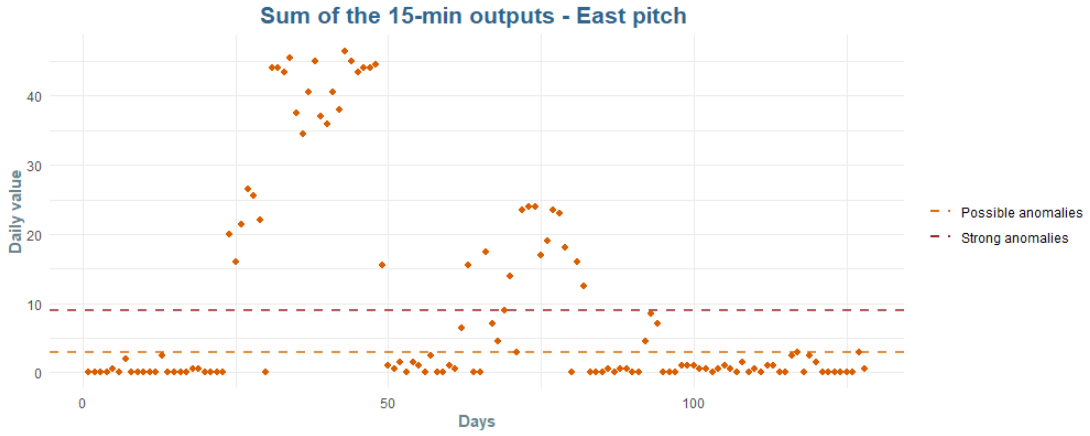

Figure 55: Sum of the 15-minutes output for the east pitch.

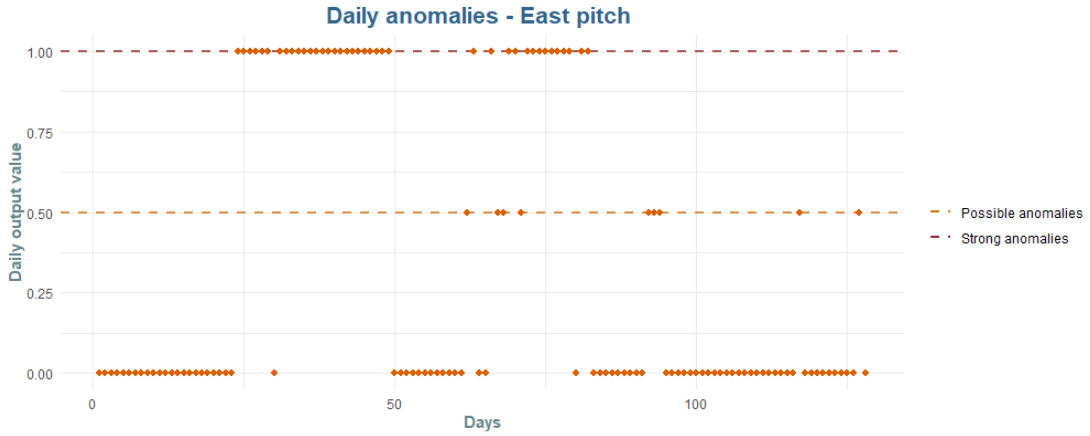

Figure 56: Daily outputs of the anomaly detection for the east pitch.

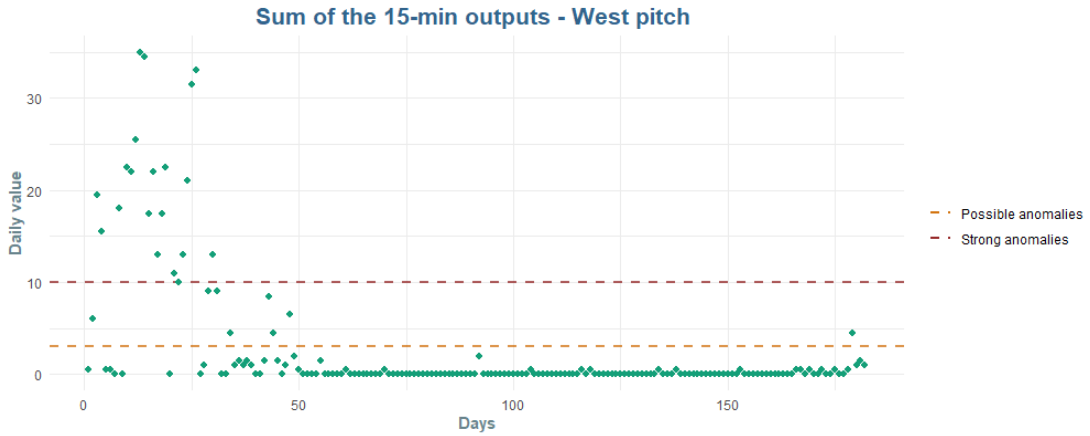

Figure 57: Sum of the 15-minutes output for the west pitch.

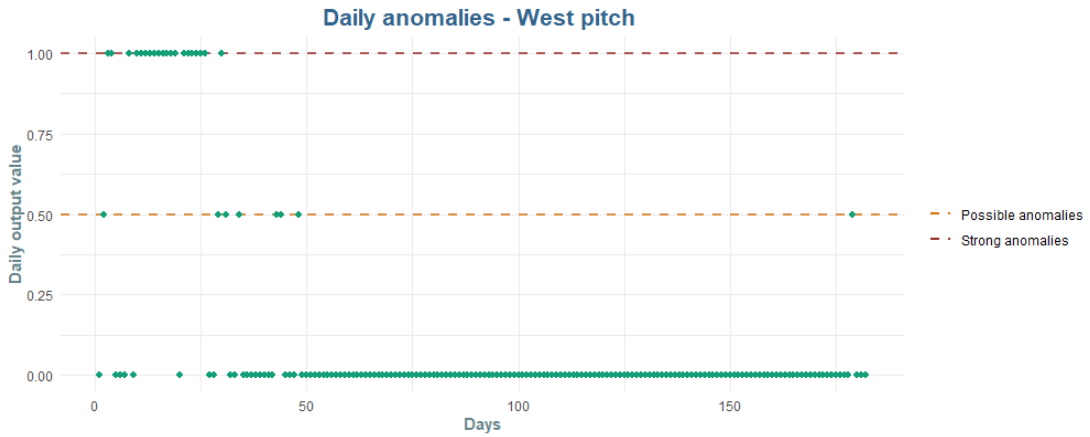

Figure 58: Daily outputs of the anomaly detection for the west pitch.

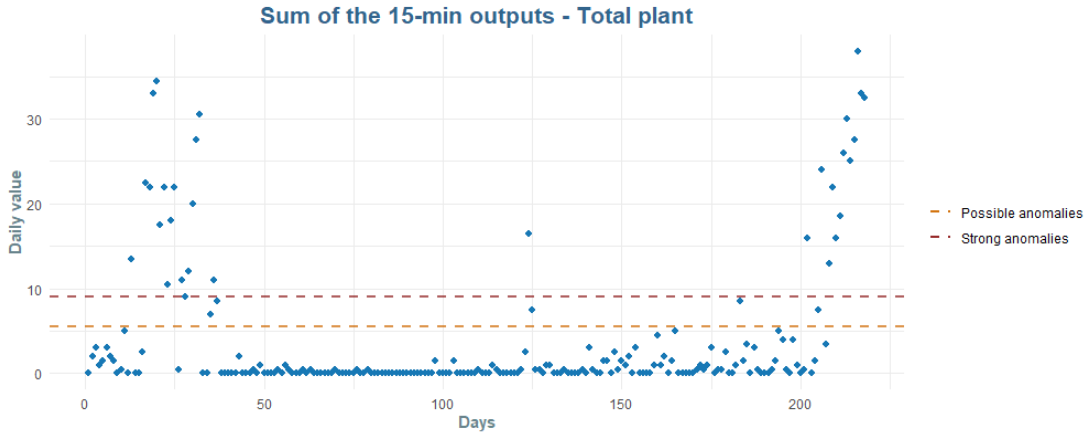

Figure 59: Sum of the 15-minutes output for the total plant.

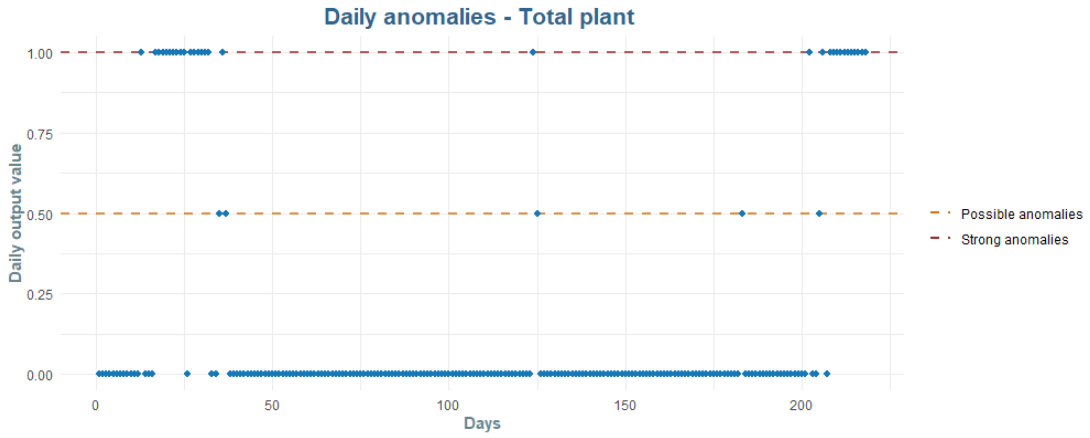

Figure 60: Daily outputs of the anomaly detection for the total plant.

From the observation of the figures for each model, it can be clearly noticed the association criterion of the day with the output value: depending on the position of the sum with respect to the thresholds, the consequent label is assigned.

Applying the described procedure, a certain number of anomalous days has been detected. In particular, for the east pitch, 39 strong anomalies and 9 possible ones are detected, corresponding to the 30.5% and 7 % of the total days, respectively. With these results, the east pitch is the one that presented more days labelled as anomalous, consisting in more than a quarter of the considered period. Instead, the west pitch shows strong malfunctions for 20 days (about 11%) and possible ones for 8 days (4%

of the considered period); finally, for the total plant, 218 day are considered and, among them, 31 results to be strongly anomalous, while 5 are identified as possible malfunctions.

Assessment of the quality of the detection In this paragraph the results are shown referring to the assessment of the quality of the anomaly detection procedure. In particular, in a first part, the values of the performance metrics are presented for the chosen daily thresholds; then, the estimation of the quality is summarised also for other two combinations of daily limits.
The following tables contain the resulting values of True Positive Rate (Equation 17), False Positive Rate (Equation 18), Accuracy (Equation 2), Precision (Equation 3) and Area Under Curve (Equation 19) of the applied procedure with the chosen thresholds, distinguish between the detection of strong anomalies, possible anomalies and total anomalies. The Table 23 refers to the west pitch, the Table 22 to the east pitch, while the Table 24 regards the total plant.

	Strong	**Possible**	**Total**
TPR	0.79	1	0.88
FPR	0	0.056	0.039
ACCURACY	0.92	0.95	0.93
PRECISION	1	0.22	0.94
AUC	0.89	0.97	0.92

Table 22: Performance metrics of the anomaly detection for the east pitch

	Strong	Possible	Total
TPR	1	0.88	0.96
FPR	0.006	0.006	0.013
ACCURACY	0.99	0.99	0.98
PRECISION	0.95	0.88	0.93
AUC	0.99	0.94	0.98

Table 23: Performance metrics of the anomaly detection for the west pitch

	Strong	Possible	Total
TPR	1	0.71	1
FPR	0.02	0	0.01
ACCURACY	0.99	0.99	0.99
PRECISION	0.90	1	0.97
AUC	0.99	0.99	0.99

Table 24: Performance metrics of the anomaly detection for the total plant

As it is evident, the detection procedure reaches an excellent quality in the identification of the anomalies on all the three considered domain and for both the types of faults.

The lower performance can be observed in the case of the detection of strong anomalies for the east pitch: the AUC in this case is 0.89. This is due to the fact that many strong faults occurred during winter in which the production can be also very low (in the order of 1 kW): even if the real power is zero, while the predicted one is not, the difference remains small and, as a consequence, the malfunction of the pitch is not detected by the model. However, an increase in the sensitivity of the detection would have caused a great number of *false positives* so normal behaviours identified as anomalous.

Before choosing the final daily thresholds, three attempts have been made for each system, consisting in three different combinations of the daily limits. The Table 25 summarizes these combinations and the Figure 61 shows the value of the Area Under Curve for each of them.

	COMBINATION 1		**COMBINATION 2**		**COMBINATION 3**	
	τ_L	τ_U	τ_L	τ_U	τ_L	τ_U
East pitch	3	8	3	9	3	10
West pitch	3	8	3	9	3	10
Total plant	3	8	3	9	5.5	9

Table 25: Tested combination of daily thresholds.

The Figure 61 represent the evolution of the Area Under Curve changing the combination of daily thresholds. As a first consideration, it can be noticed that the quality of the detection of the generic anomaly (i.e. *Total* case, green line) does not change much with the variation of the limits, while the one that is more affected by the the combination is the detection of possible anomalies. In fact, the choice of the final daily limits has strongly depended on this value, trying to obtain an acceptable and satisfying results also for the possible malfunctions.

The chosen combination of thresholds for the east pitch is the number 2, corresponding to a maximum of the AUC for possible anomalies: after that, increasing the upper value, there is a worsening of the performance for both the type of anomalies.

For the west pitch, instead, the Combination 3 has been selected because of an improving of the performance for the detection of all the anomalies.

Finally, the daily limits for the total plant are those belonging to last combination: an increase in the lower limits has been necessary to detect possible anomalies, even with a slight decrease of the quality for strong ones.

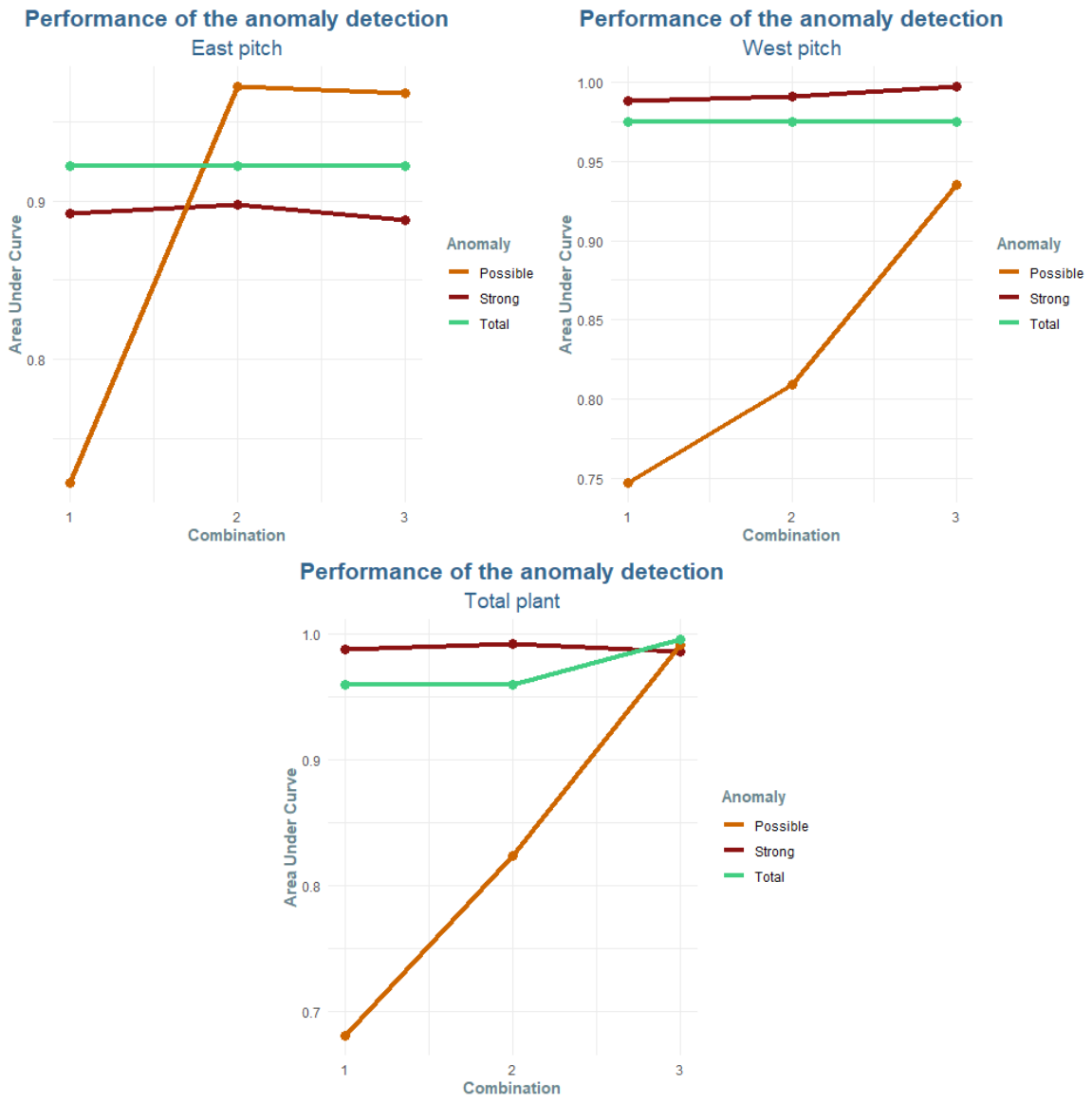

Figure 61: Values of the AUC with different combination of daily thresholds for the east pitch (up right), the west pitch (up left) and the total plant (bottom).

5.3.4 Predictive maintenance

This last section deals with the results of the proposed predictive maintenance procedure. As described in the methodology, the trend of the daily residuals is extracted by computing the exponential moving average and it is used to highlight the necessity of an extraordinary maintenance: if a daily anomaly is in correspondence of a degradation trend, a maintenance action is suggested by means an output different from

zero, and whose value is referred to the type of anomaly.

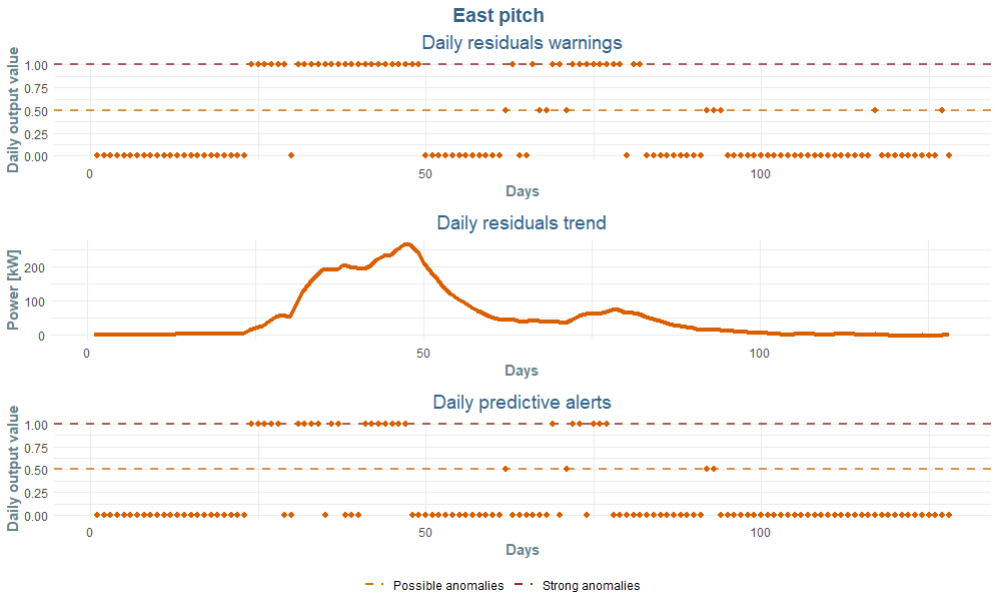

Figure 62: Daily anomalies, residual trend and predictive maintenance alerts for the east pitch.

Figure 63: Daily anomalies, residual trend and predictive maintenance alerts for the west pitch.

Figure 64: Daily anomalies, residual trend and predictive maintenance alerts for the total plant.

The Figures 62, 63 and 64 show, for the single pitches and for the total plant, the daily anomalies, the trend of the residuals and, finally, the predictive maintenance outputs. Their observation is useful to understand how the procedure works: if a daily anomaly warning is in correspondence of a ascending section, it is converted in a maintenance alert. On the contrary, if the warning is referred to a day belonging to a descending trend, there is not a predictive alert.

In this way, it is possible to identify situations in which the plant is having a degradation of its performance that does not regard only one day and that might be not immediately noticed; maintenance actions are consequently possible in order to restore condition of correct operation of the system.

An example of predictive maintenance alerts is given by the Figure **??**: it is referred to the production of the total plant in June 2021 and highlight the days during which an alert is expected. The labelled days are those for which a maintenance alert is given and the color is referred to the type of anomaly: orange for possible anomalies and red for strong ones. As it can be noticed, the highlighted days are not isolated faults, but they belongs to a time period during which more than one day is characterised

by a lower production with respect to the expected one.

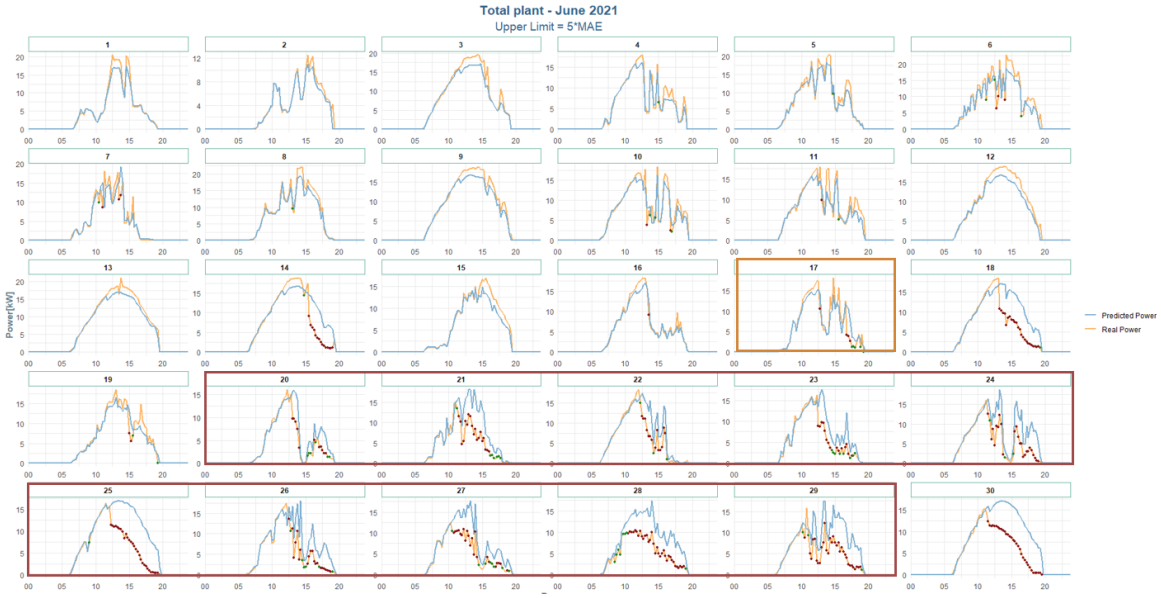

Figure 65: Example of predictive maintenance alert for the total plant.

6 Conclusions

This thesis work defines a methodology to enhance the energy management of an educational building, the Polytechnic of Turin, by means of Data Analytics techniques: it proposes a procedure to assess the amount of energy waste on the load-side and detect anomalies and malfunctions in the PV production plant.
The analysis is based on power-related data from the on-site monitoring systems and, under this aspect, the first and main difficulty has been encountered. In fact, the used techniques are based on data and their performance strongly depend on the quality of the input measurements: the models learn relations from the inputs and, as a consequence, if the inputs are not coherent or missing, the accuracy of the results will be affected. Regarding this issue, the on-site measurement network presents more than one problem: first of all, it is not arranged into a well-defined structure making it difficult to understand the meaning of each measure and its relation with the others; moreover, the switched-off, either temporary or permanent, of some of its devices is not reported and the database is not updated. Finally, most of the meters present malfunctions and transmission errors, also because of their age and of the absence of maintenance: this causes incorrect records of the measured quantities. As a consequence, a consistent part of the work is used in order to create an usable dataset: even after the arrangement of all the meters in an hierarchical structure, an issue remains regarding the quality of the data. The meters that resulted to be more problematic are those associated to the PV power plant: there is a lack of data for large periods of time and the recorded measures need to be checked and cleaned.
This first part of the thesis work highlights the crucial importance of the quality of the data and the need for an accurate pre-pocessing in order to guarantee a satisfying performance of the analysis.

The second part of the thesis work allows to better understand the behaviour of the loads and detect cases of energy waste in their operation. However, a first issue that can be emphasized, is the poor level of detail of the measurement system on the

load-side: only two sub-loads are individually monitored, missing the opportunity of a complete analysis on the consumption of other relevant domains, such as laboratories and lecture rooms.

The thesis work focuses on the baseload consumption because it has been considered of particular interest in the evaluation of energy wastes and opportunities of improvement. In fact, this share of demand is the one that it is always present, even if the device or the building is not used, so it represents the minimum value of consumption from which the electricity demand starts. As a consequence, its reduction can positively affect the total consumption of the system and it can be an excellent starting point for a better energy management of the university.

The thesis work provides a procedure for an efficient prediction of the baseload value, as a function of external conditions: it can be used to make considerations about energy savings and to be prepared for higher demand in the case of certain boundary conditions. The analysis allows to make considerations about the actual energy management of the sub-loads and proposes a quantification of the energy wastes.

The I3P-related part of the analysis identifies as one of the causes of energy waste, the high electricity demand during non-working hours, when the building should not be used. This share of consumption may be due to the fact that some electric devices are left switched on in the offices, even if the working-time is ended. This might be due to effective work-related necessities, such as the need for a continuous elaboration of data, even during night, that make it necessary to left devices in function. However, in a view of efficient energy management, it can be suggested to early communicate an higher consumption in order to not detect that day as an anomalous demand value.

The chillers-related part, instead, reveals situation of strong energy waste due to a manually early switching on of the machines in very hot days: the chillers are in function even 10 hours before the beginning of the working day. As expected, this procedure is strongly energy-intensive, causing consumption that are considered not beneficial and that have to be reduced. A solution can be identified in the installation of time-switches in order to fix the ramp ups of the chillers at a certain time that, in the case of expected hot temperatures, can be set only 1 or 2 hours before the normal

schedule.

Finally, the simulation of an improved solution shows the potential energy saving associated only with a reduction of the power during non-working periods: it is possible to reach an overall reduction of about 36 MWh, cutting the 2.3% of the I3P demand and about the 30% of the chillers consumption.

The last part of the thesis work develops a procedure for the forecast of the PV production, the anomaly detection and the predictive maintenance of the plant. Both the parts of the analysis have good performances: the artificial neural network shows very low values of MSE for all the outputs and the anomaly detection is able to identify malfunctions with excellent values of the metrics. In fact, the algorithm is able to identify, with good quality, days in which the PV plant presents anomalous behaviour and even a null production.

The step of the predictive maintenance has the aim to highlight those situations in which an extraordinary maintenance is suggested because of degradation trends of the performance of the PV plant. The general idea is to detect anomalies in the system and, when a series of them happens consecutively, provide an alert to technical professional figures to recommend a further inspection. The output of this procedure is exclusively informative and the final decision on the effective measures has to be taken also on the basis of technical experience.

The procedure has been applied to both individual pitches of the plant and to the total system: all of the three algorithms show good performances, however the advantage of considering the pitches as separated is to make a step forward and add the relevant information about the location of the anomaly.

In conclusion, the present thesis work defines a procedure that can be used to improve the energy management of a multi-purpose educational building, providing results concerning the applicability and the consistence of the algorithm. It highlights actual critical issues regarding the present situation, increasing the knowledge about the system and suggesting possible correcting actions.

Regarding the development of the algorithm, the work highlights some issues: the first

one is the crucial importance of the quality of the data. In fact, the model performance will always be influenced by the kind of inputs that it receives, and even if it is well constructed, it will give an incoherent result if it learns from incoherent data. This point is linked, to a certain extent, to an other issue related to the algorithm: it is necessary a good knowledge of the system of interest in order to effectively interpret the results of the process and use them. In fact, Data Analytics can be a powerful tool to enhance the energy management of a system but it is not based on physical correlations, so it requires an expert to evaluate the results and wisely use them.

Regarding the implementation in real-life of the algorithm, opportunities and limitations can be found. First of all, the developed methodology allows a management of both loads and production with an economic investment that is very low: it is not required additional and expensive hardware and the procedure can be applied with the actual resources (i.e. monitoring system). For the anomaly detection on the PV plant, the whole dataset is composed of quantities that are free and easily accessible: the on-site measurement network provides information about power, external temperature and solar irradiance, while the position of the sun can be collected from open-access sources. Moreover, even for the load-side of the analysis, the implementation of the actions of energy savings potentially requires not expensive devices: it has been shown that a reduction of the average baseload power can results in a relevant decrease of the consumption, and it can be achieved with a delay in the switching-on of the chiller units. This solution needs time-switches that can be easily installed at a low cost.

However, the effective application of the described solutions might meet some difficulties. The most evident one is related to the monitoring network: it is a planned maintenance to guarantee its correct operation in order to allow a real-time analysis of the system. Especially for the PV plant, the metering devices have to provide continuous measurements in order to make it possible to monitor the pitches and finds faults: if the meters does not acquire measurements, the algorithm will warn about a malfunction of the plant that it is actually associated to the monitoring system.

Another complication in the applicability of the energy saving strategies is related to the main loads. For instance, the delay in the ramp-ups of the chiller might be

cause of complaints among the occupants and further analysis should be carried out in order to verify if the thermal comfort can be achieved with the proposed values of power. Moreover, for the I3P building it has been proposed to early communicate an higher electrical consumption due to work-related necessities, in order to not identify a certain day as anomalously energy-intensive; however, this solution is not so easy to apply because there is not a reference figure who can manage these communications. As future development of the analysis, it might be useful to implement the algorithm in-real time - with all the above-mentioned difficulties- in order to continuously monitor the system (both loads and production) and give daily feedback on its operation. An improvement that results to be necessary regards the monitoring system: it should be periodically checked in order to find malfunctions and, more importantly, it should be extended in order to get a complete dataset of the consumption of the various sub-loads. At this point, it might be interesting to apply the proposed methodology to other types of loads such as heat pumps, rooms and laboratories.

Acknowledgements

First of all, I would like to thank my supervisor, prof. Alfonso Capozzoli, for giving me the opportunity to study in depth the new and interesting topics explored in this project, and for the great willingness shown during these months.
I am also extremely grateful to my co-supervisor, PhD Marco Savino Piscitelli, for all the help and continuous advice and for constantly challenging me to improve and refine my work.
Finally, I would like to extend my sincere gratitude to Eng. Giovanni Carioni, Eng. Fabrizio Tonda Roc and the Living Lab of the Politecnico di Torino for their support and for providing the data related to the case studies analysed in this thesis.

References

[1] United Nations Framework Convention on Climate Change, *Glasgow Climate Pact*, Advance Unedited Version, 2021.

[2] United Nations Framework Convention on Climate Change, *Adoption of the Paris Agreement*, 2015.

[3] IEA, *Global Status Report for Buildings and Construction*, 2019.

[4] IEA, World energy balances, online database.

[5] Taskgroup, I. B. E. E. (2019). Building Energy Performance Gap Issues: An International Review.

[6] Van Dronkelaar, C., Dowson, M., Burman, E., Spataru, C., Mumovic, D. (2016). A review of the energy performance gap and its underlying causes in non-domestic buildings. Frontiers in Mechanical Engineering, 1, 17.

[7] Molina-Solana, M., Ros, M., Ruiz, M. D., Gómez-Romero, J., Martín-Bautista, M. J. (2017). Data science for building energy management: A review. Renewable and Sustainable Energy Reviews, 70, 598-609.

[8] Himeur, Y., Ghanem, K., Alsalemi, A., Bensaali, F., Amira, A. (2021). Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. Applied Energy, 287, 116601.

[9] Fan, C., Xiao, F., Li, Z., Wang, J. (2018). Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. Energy and Buildings, 159, 296-308.

[10] Deb, C., Schlueter, A. (2021). Review of data-driven energy modelling techniques for building retrofit. Renewable and Sustainable Energy Reviews, 144, 110990.

[11] De Benedetti, M., Leonardi, F., Messina, F., Santoro, C., Vasilakos, A. (2018). Anomaly detection and predictive maintenance for photovoltaic systems. Neurocomputing, 310, 59-68.

[12] Harrou, F., Dairi, A., Taghezouit, B., Sun, Y. (2019). An unsupervised monitoring procedure for detecting anomalies in photovoltaic systems using a one-class support vector machine. Solar Energy, 179, 48-58.

[13] Garoudja, E., Harrou, F., Sun, Y., Kara, K., Chouder, A., Silvestre, S. (2017). Statistical fault detection in photovoltaic systems. Solar Energy, 150, 485-499.

[14] Le, M., Nguyen, D. K., Dao, V. D., Vu, N. H., Vu, H. H. T. (2021). Remote anomaly detection and classification of solar photovoltaic modules based on deep neural network. Sustainable Energy Technologies and Assessments, 48, 101545..

[15] Garoudja, E., Chouder, A., Kara, K., Silvestre, S. (2017). An enhanced machine learning based approach for failures detection and diagnosis of PV systems. Energy conversion and management, 151, 496-513.

[16] Garoudja, E., Chouder, A., Kara, K., Silvestre, S. (2017). An enhanced machine learning based approach for failures detection and diagnosis of PV systems. Energy conversion and management, 151, 496-513.

[17] Benkercha, R., Moulahoum, S. (2018). Fault detection and diagnosis based on C4. 5 decision tree algorithm for grid connected PV system. Solar Energy, 173, 610-634.

[18] Taghezouit, B., Harrou, F., Sun, Y., Arab, A. H., Larbes, C. (2021). A simple and effective detection strategy using double exponential scheme for photovoltaic systems monitoring. Solar Energy, 214, 337-354.

[19] Zhao, Y., Yang, L., Lehman, B., de Palma, J. F., Mosesian, J., Lyons, R. (2012, February). Decision tree-based fault detection and classification in solar photovoltaic arrays. In 2012 Twenty-Seventh Annual IEEE Applied Power Electronics Conference and Exposition (APEC) (pp. 93-99). IEEE.

[20] Madeti, S. R., Singh, S. N. (2018). Modeling of PV system based on experimental data for fault detection using kNN method. Solar Energy, 173, 139-151.

[21] Xiao, F., Fan, C. (2014). Data mining in building automation system for improving building operational performance. Energy and buildings, 75, 109-118.

[22] Amasyali, K., El-Gohary, N. (2016). Building lighting energy consumption prediction for supporting energy data analytics. Procedia Engineering, 145, 511-517.

[23] Sendra-Arranz, R., Gutiérrez, A. (2020). A long short-term memory artificial neural network to predict daily HVAC consumption in buildings. Energy and Buildings, 216, 109952.

[24] Lin, J., Fernández, J. A., Rayhana, R., Zaji, A., Zhang, R., Herrera, O. E., Liu, Z., Mérida, W. (2022). Predictive analytics for building power demand: day-ahead forecasting and anomaly prediction. Energy and Buildings, 255, 111670.

[25] Capozzoli, A., Piscitelli, M. S., Brandi, S. (2017). Mining typical load profiles in buildings to support energy management in the smart city context. Energy Procedia, 134, 865-874.

[26] Yang, J., Ning, C., Deb, C., Zhang, F., Cheong, D., Lee, S. E., Sekhar, C., Tham, K. W. (2017). k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. Energy and Buildings, 146, 27-37.

[27] Alam, M., Devjani, M. R. (2021). Analyzing energy consumption patterns of an educational building through data mining. Journal of Building Engineering, 44, 103385.

[28] Ma, Z., Yan, R., Nord, N. (2017). A variation focused cluster analysis strategy to identify typical daily heating load profiles of higher education buildings. Energy, 134, 90-102.

[29] Govender, P., Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). Atmospheric Pollution Research, 11(1), 40-56.

[30] T. Soni Madhulatha, *An overview on clustering methods*, Article in *IOSR Journal of Engineering*, April 2012, Vol. 2(4), pp: 719-725.

[31] Loh, W. Y. (2011). Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery, 1(1), 14-23.

[32] De'ath, G., Fabricius, K. E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology, 81(11), 3178-3192.

[33] De Sangro A., (2019). Virtual Sensor for Human Safety Monitoring in Factory 4.0 Applications, Master thesis, Department of Control and Computer Engineering, Politecnico di Torino.

[34] Marsh Risk Consulting, (2017). Stima Preventiva Danni Diretti Aggiornamento Valori, Politecnico di Torino.

[35] Wickham H, (2016). *ggplot2: elegant graphics for data analysis*, Second Edition, Springer.