STAGE REPORT

Prediction of the HVAC needs of a building by means of Machine Learning algorithms

Author:
Daniel COSTERO VALERO

UPM Supervisor:
Dr. Miguel HERMANNS

ENSMA Supervisor:
Dr. Etienne VIDECOQ

Diplôme d'ingénieur
ISAE-ENSMA

October, 2018

Contents

List of Figures

List of Tables

Chapter 1

Introduction

More than 25% of the world energy consumption is due to the heating and cooling of buildings. With such a high number, the widespread adoption of sustainable construction techniques represents a key milestone in the ongoing effort of reducing mankind's impact on Earth's climate. Therefore, the European Union has recently proposed an update to its Energy Performance of Buildings Directive in order to enforce all new constructions to be nearly zero-energy buildings by 2021.

To achieve this ambitious goal, sustainable HVAC (heating, ventilation, and air conditioning) systems are required, like those harnessing low enthalpy geothermal energy. These rely on a water-to-water heat pump connected to a geothermal heat exchanger comprised of vertical boreholes with U-shaped pipes in them, and through which a liquid flows and exchanges heat with the surrounding ground. Such a geothermal HVAC system can only be considered truly renewable, if the geothermal heat exchanger and the heat injection/extraction strategy are properly designed. Otherwise, thermal exhaustion of the ground takes place, significantly affecting the performance of the system.

To ensure the long-term efficiency of a geothermal HVAC system, its thermal behavior after 25, 50, or even 100 years of operation must be known. Although a time marching of the governing energy conservation equations is possible nowadays, it is still computationally too expensive. Therefore, simplified theoretical models are used instead, which trade in accuracy and flexibility for computational speed.

The Department of Fluid Mechanics and Aerospace Propulsion of the

Technical University of Madrid (UPM) is working since 2011 in the theoretical modeling of geothermal heat exchangers, using methods of scale analysis and asymptotic expansion techniques borrowed from the research fields of fluid mechanics and aerospace engineering. The results obtained so far are very promising, and outperform the actual state-of-the-art.

The close collaboration with Sacyr Industrial (large Spanish construction company that has designed and constructed some of the biggest geothermal installations in southern Europe) and some other companies ensures the work being done at the University is perfectly aligned with the real world problem and with the actual needs of the industry.

Another important aspect to achieve this nearly zero-energy buildings goal is to optimize the control of the heat pump, that regulates the amount of energy exchanged between the building and the ground. Being able to predict the energy demand of the building allows a much better optimization of the system, both in short and long term, reducing the consumption of any external source of energy, such as electricity or natural gas.

Predicting the energy consumption of a building is an arduous task as it depends on multiple variables, such as the local meteorology, the type of work carried out in the building or the number of people that is inside the building at any time. All these features vary a lot and their relationship with the energy consumption is not obvious.

Nowadays, machine learning techniques are being used to perform this prediction as they outperform any existing theoretical model. These techniques are able to capture all the non-linear dependencies between the different variables, even if they are not clear for humans, without being explicitly programmed for it.

Among all the different machine learning algorithms, Artificial Neural Networks (ANN) and Random Forest (RF) are, by far, the most used ones, as they capture extremely well all internal relationships. The Random Forest algorithm in more recommendable because it provides some information of how the algorithm works internally. This feedback may show why the algorithm does not work properly or what should be changed to improve the performance of the algorithm. It also provides information about the importance that each feature has in the model, allowing the elimination of the useless ones and reducing the calculus time. Even if some information is provided by other algorithms such as ANN, this information has no physical or logical explanation and only makes some sense to specialists in machine learning.

This work aims to predict the power consumption of the HVAC system of a geothermal installation using the historical data stored by the system and the forecast. The prediction is done using the Random Forest algorithm that has been proven to provide good results. Different sets of variables are tested trying to optimize the accuracy of the model. This prediction will allow the optimization of the heat pump controller, improving the global efficiency of the installation.

Chapter 2

How geothermal energy works

By definition, geothermal energy is the thermal energy stored in the Earth. The thermal energy of a body is the amount of energy that it has as a result of having a temperature bigger than 0K. The Earth has a huge amount of thermal energy derived from the radioactive decay of materials in the core of the planet, formed during its creation. As the temperature outside the core is much lower than the inner one, there is an enormous gradient of temperature between the core and the surface, generating a huge heat transfer that remains constant due to the nature of the reactions at the core of the planet.

Due to this enormous heat flow, the profile of temperature inside the Earth does not vary with the time from some meters below the ground. Only in a superficial layer of around 20 meters of depth, where the interactions with the atmospheric phenomena are important, the temperature profile changes with the time. The deeper a point is placed, the closer to the core it gets and its temperature rises enormously. Neglecting the first 100 meters, then the temperature rises with a ratio of 3°C each 100 meters approximately. A typical profile temperature is shown in table 2.1. The figure in the left shows the profile of the whole Earth, down to the core, whereas the figure in the right only shows the superficial changes down to 100 meters below the ground.

Depending on the depth of the wells, some different types of geothermal energy can be distinguished (figure 2.1), as the enthalpy of the energy extracted grows with the depth of the wells.

- **Very Low Enthalpy.** This type of geothermal energy has wells with a low depth, around 100 meters. The temperature of the ground is less than 25°C, and it can be used for heating, SHW (Sanitary Hot Water) or

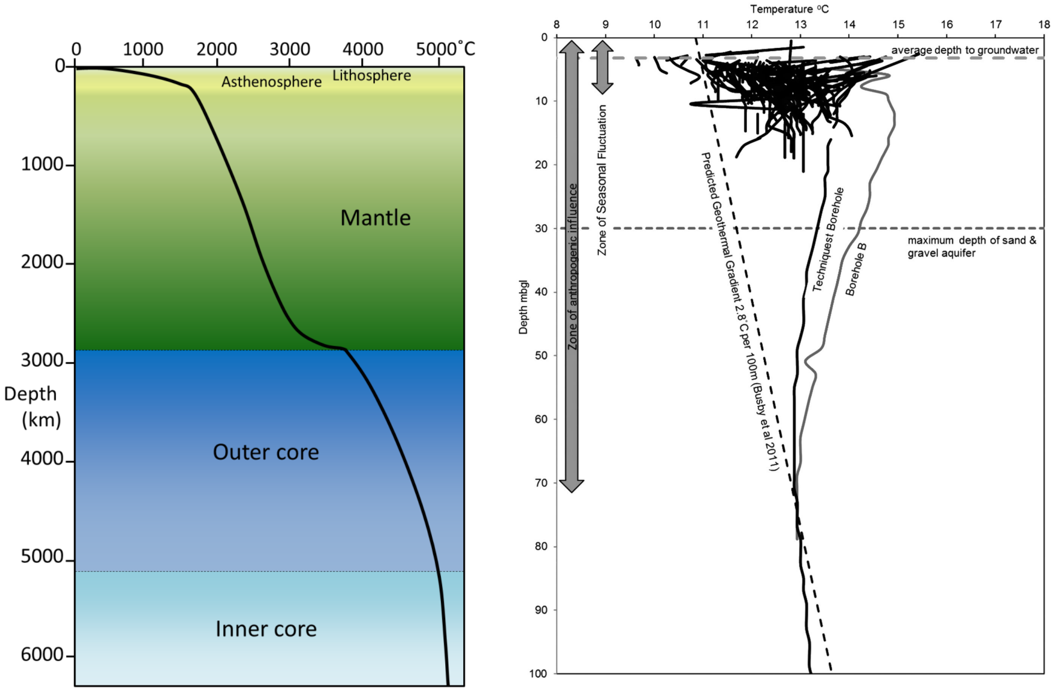


FIGURE 2.1: Temperature profile of the ground in Cardiff. (Ref [10])

air conditioning of buildings. It works as a complement for other types of energy, such as electricity or natural gas, reducing its consumption. The main advance of the wells is that the energy is exchanged with the ground, that has an almost constant temperature, instead of exchanging with the atmospheric air. During the winter, the energy is extracted from the ground, that has a higher temperature than the air. Like that, the thermal gradient between the external air and the temperature inside the building is smaller, rising the global efficiency of the heat pump. During the summer, at least in hot countries like Spain, the ground is cooler than the air and the amount of energy needed to cool down the air conditioning is less than if the air was taken directly from the exterior. On the other hand, the main disadvantage is that the ground temperature is not high enough to work on its own, and some other source of energy is needed.

- **Low Enthalpy.** It takes hot water from the ground, with temperature lower than 100ºC, so the water is always in a liquid state. This water is pumped into the buildings, heating them up, and returned into the ground, were it naturally retakes its original temperature. It requires an accessible source of hot water that may be difficult to find.

Enthalpy	Terrain sample	Temperature Range	Utilization
Very Low	Subsoil Underground water	$5°C < T < 25°C$ $10°C < T < 22°C$	Heating, ACS, ...
Low	Thermal waters Volcanic zones Deep sediment storage	$22°C < T < 50°C$ $T < 100°C$	SPAs, agriculture District heating
Medium		$100°C < T < 150°C$	Electricity generation Binary cycles
High		$T > 150°C$	Electric generation

TABLE 2.1: Different types of geothermal energy

- **Medium and High Enthalpy.** In the ground, the pressure is very elevated and grows with the depth. Even if the temperature is above 100°C, the water is still on a liquid state due to this high pressure. When a well is made and this liquid hot water reaches the surface, with the atmospheric pressure, it evaporates and a hot flow of vapor is obtained. It can be used to generate electricity directly with a turbine or it can be used in a binary cycle. This type of geothermal energy is widely used in countries such as Iceland or Finland, as they count with hot sources close to the surface and a ground formed mostly by granite, that facilitates the performance of extremely deep wells.

2.1 Heat Pump

In very low enthalpy installations, some external source of energy needs to be used. It is usually made through a heat pump, that uses electrical energy to control the amount of heat exchanged with the ground, as a function of the demand of the building. A diagram of a heat pump is shown in figure 2.2.

The heat pump is filled with an special fluid, usually called 'refrigerant', that follows the Rankine cycle, shown in figure 2.3. First, it is heated up with the heat coming from the wells. As the pressure here is very low, the fluid evaporates with a constant pressure. During the evaporation, the temperature remains constant, rising up once all the fluid has become vapour. Then, it is compressed with an electric compressor, obtaining a gas with high pressure and high temperature. This fluid transfers heat to the fluid that needed

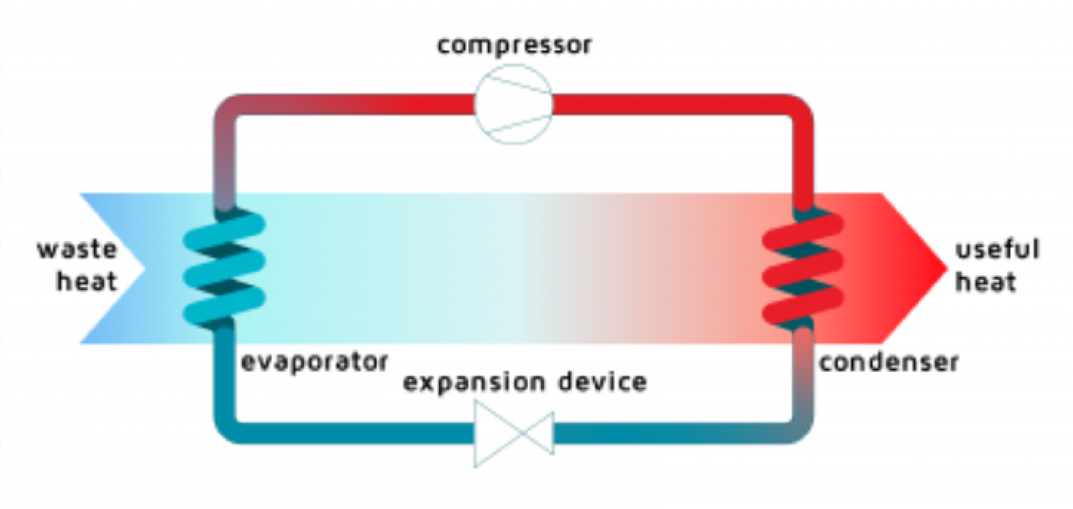


FIGURE 2.2: Diagram of how a heat pump works

to be heated up initially, condensing itself into an almost-liquid state. The fluid, with high pressure and low temperature, is expanded in a turbine, producing energy that is used to move the compressor and closing the cycle.

The main advantage of this cycle is that the electric energy consumed is smaller than the energy exchanged as heat, improving the efficiency of the installation. The fluid going out of the wells, has always a constant temperature around 17ºC. Being able to extract heat from this flow requires the temperature inside the heat pump (the refrigerant) to be smaller. On the other hand, reducing the gradient of temperature followed by the refrigerant during the cycle improves the efficiency of the heat pump. If, instead of exchanging heat with the wells, it was exchanged with the external cooler air (10ºC for example), the temperature reached for the refrigerant would be smaller and the efficiency would be worst. In addition, external air suffers weather changes, and the amount of heat exchanged and the efficiency would vary with the time.

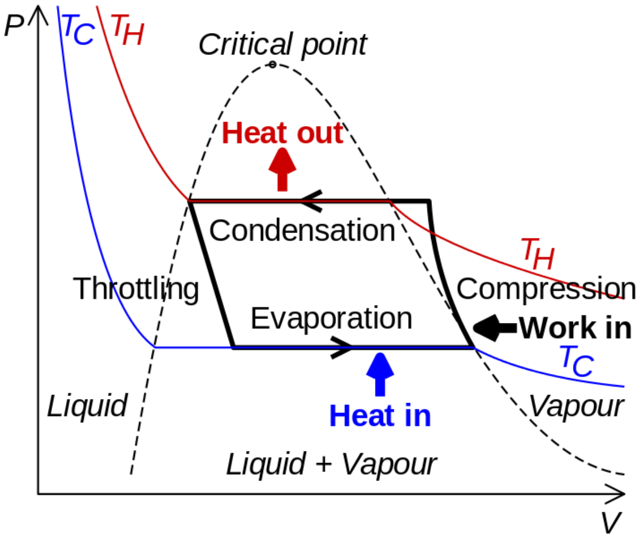


FIGURE 2.3: P-V diagram of a Rankine cycle

Chapter 3

Machine Learning

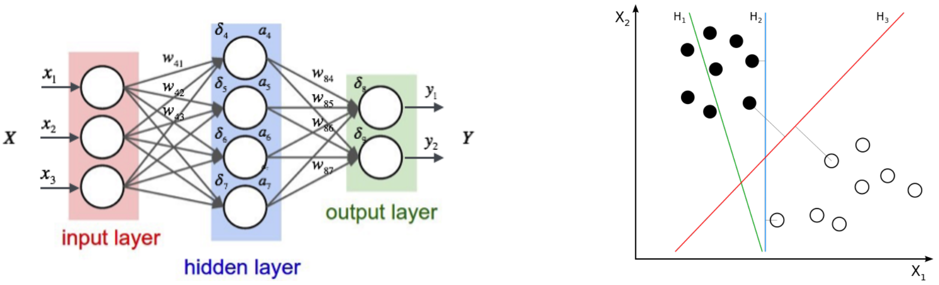
3.1 Introduction to Machine Learning

Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" (e.g., progressively improve performance on a specific task) with data, without being explicitly programmed.

A more formal definition is given by Tom M. Mitchell: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E".

All the machine learning techniques can be classified into two groups: supervised and unsupervised learning.

- **Supervised learning.** In this type of techniques, some examples and their correspondent desired outputs are given to the algorithm. With all these data, it learns by itself the different relationships between the inputs and the outputs, creating an internal model. This models is able to 'predict' the output for any given example. An example of supervised learning would be the prediction of the prize of a house given its size and its position and a collection of examples.
- **Unsupervised learning.** In this type of techniques, only a list of data is given, without any label or desired output for the examples. The algorithm is able to find by itself some inner patterns and structures. An example of unsupervised learning is the suggestions of webs such as amazon or YouTube. They know what people do in their webs: what they search, what they look, where they click... With all these data, they are able to predict your interests and use it with commercial purposes.



(A) Artificial Neural Network (B) Support Vector Machine

FIGURE 3.1: Examples of supervised learning algorithms

Machine learning algorithms are usually used for two main purposes: classification and regression. In classification problems a label is given to each point of the data, both in supervised and unsupervised learning. A typical example of that is the spam filtering, classifying emails as 'spam' or 'not spam'. In regression problems the prediction is made over continuous variables, such as the prize of a house or the power consumption of a building.

In supervised learning, the data is usually divided into three different sets: training set, cross-validation set and test set. A typical separation would be around 60%, 20% and 20% of the total set of data, respectively. The train set is used by the algorithm to find the inner relationships in the data and create a model. The cross validation set is used to tune the parameters of the algorithm and optimize the model. The data of the training set is trained with different parameters of the algorithm and its performance is analyzed in the cross validation data, choosing the set of parameters that reduces the error predicting the cross validation. Once the model is optimized, the test set is used to evaluate the accuracy of the predictions made by the model.

There are several algorithms used in supervised machine learning, such as artificial neural networks, decision trees, support vector machines or genetic algorithms:

Artificial Neural Networks (ANN) are algorithms inspired in the functioning of the brain where there are a lot of 'neurons' connected to each other, exchanging information between them and creating a complex network. A typical neuron receives information from some other neurons and evaluates a function with a linear combination of the different received information, returning a real value that is going to be used for other neurons. It is a supervised learning algorithm that 'learns' by calibrating the inner parameters of the network. A typical diagram of this type of network is shown in figure 3.1a.

Support Vector Machine (SVM) is a supervised algorithm used to classify data into two different types. It creates a boundary between the two types of data, maximizing the distance of the points to this boundary. By doing that, two regions of the space of variables are created and the algorithm predicts the label of any new example depending on which region of the space the sample is placed. A diagram of how a Support Vector Machine works is showed in figure 3.1b. As it is shown, line H1 is not able to separate correctly both labels. Even if H2 and H3 create a correct boundary, it is clear that the reliability of line H3 is bigger than the one of H2, and it is the one that is calculated by SVM.

Genetic algorithms (GA) can also be used in machine learning. Specifically, a cost function is defined as the difference between the model to be created and the training data and needs to be minimized. As it depends on a large number of different variables, traditional minimization techniques may be taken too long and genetic algorithms can be used to reduce the calculation time.

The main disadvantage of all these algorithms is that they do not provide any information of how they work internally, making difficult to rationally improve the algorithm if it does not work correctly. For this work, decision trees have been used because they return the importance of all the variables used. In some simple cases, the whole internal procedure followed by the algorithm can be visualized, making easier to understand what the algorithm is doing at any time. The functioning of these algorithms is explained below.

3.2 Fitting problems

The main objective of any machine learning algorithm is to create a model from a set of given data points able to accurately predict new output values from new data. The capacity of extrapolation is the most desired capacity

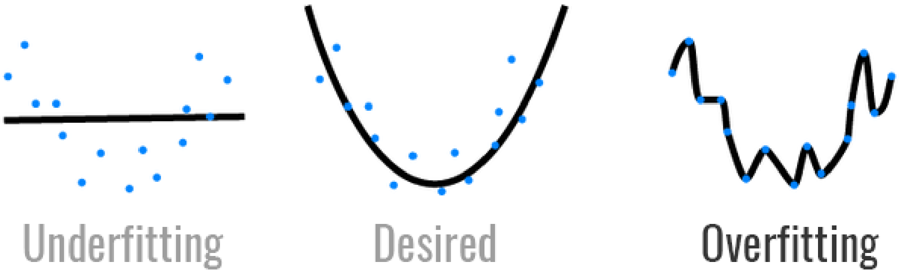


FIGURE 3.2: Fitting problems

of these algorithms and the parameters of the algorithm must be tuned correctly to achieve it.

Every machine learning algorithm has a set of parameters that must be optimized to minimize a cost function. This cost function (equation 3.1) measures the total average distance between the created model and the data set. If the value of the cost function is too big, the model does not fit the data correctly and the created model is not able to predict any new value. An example of this situation, called underfitting, is showed in the left side of figure 3.2. To correct this type of error, some parameters of the model need to be changed or some extra data need to be provided to the algorithm.

$$J(\theta) = \frac{1}{2m} \cdot \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 \tag{3.1}$$

On the opposite side, if the value of the cost function is too low, the model follows completely the data and looses it predictive capacity. This situation is called overfitting and is a very typical problem of decision trees (right side of figure 3.2). The best way to reduce this problem is reducing the number of parameters of the algorithm, making it simpler. The optimal situation is between the two previous ones, where the model only follows the trend of the data but is still able to predict correctly new values for the desired output.

Many different techniques exist to avoid overfitting. All of them are based on adding restrictions to the inner parameters of the algorithms. In ANN for example, the cost function is modified (equation 3.2) adding some extra parameters λ:

$$J(\theta) = \frac{1}{2m} \cdot \sum_{i=1}^{m} \left(h_{\theta}\left(x^{(i)}\right) - y^{(i)} \right)^{2} + \lambda \cdot \theta \tag{3.2}$$

$x^{(i)}$ and $y^{(i)}$ are each one of the training examples, h represents the created model as a function of θ, the set of original parameters of the algorithm, and λ is the new created set of parameters. λ has usually the same value for all the parameters, but it can be imposed individually. If a large value is given to λ, the optimization of the cost function is going to find small values for θ, trying to make the second part of the equation as low as possible. By doing so, some parameters will be set to very low values (or even zero), simplifying the complexity of the model and reducing the overfitting.

It is important to mention that the choice of a correct value for λ is critical to avoid overfitting. If a small value is taken, few parameters are going to be removed and the model will still be overfitting. On the other hand, if a big value is set, the second term of the equation is going to be more important than the original cost function and the model is going to underfit. To select the correct value of λ, the cross-validation data set is used. Taking only the training set of the data, several models are trained using different values of λ. These models are used to predict values from the cross-validation set, from which the output is known. The model whose predictions are closer to the values of the cross-validation set provides the optimum value of λ.

In other algorithms, like RandomForest, which have a lot of parameters, a similar procedure can be carried out iterating with a lot of different combinations of parameters. In advanced libraries, some internal procedures that affect to the way in which the splits are made are performed automatically inside the algorithm, trying to avoid the overfitting. If the model still overfits, it is usually better to change the data provided to the algorithm or the variables of the problem than to change the parameters of the algorithm.

3.3 Decision Trees

The goal of this family of algorithms is the same as any other machine learning techniques: create a model able to predict an output from a set input

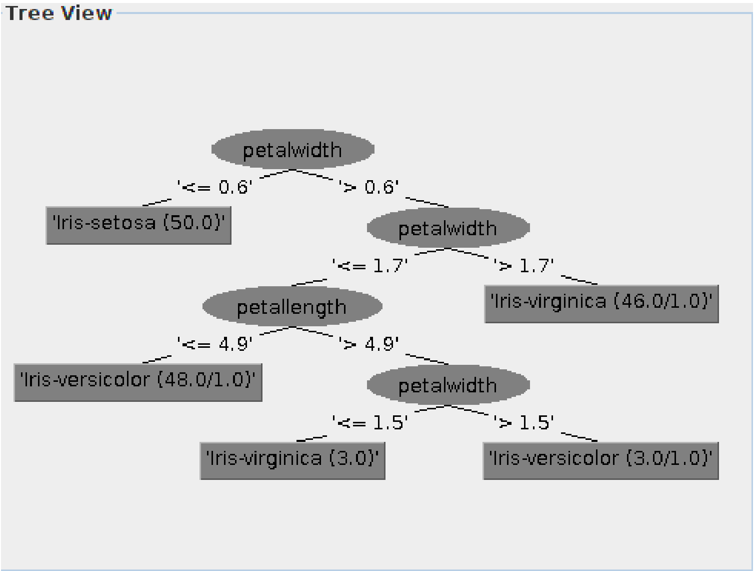


FIGURE 3.3: Decision tree diagram

variables. If the output are different labels, then it is a classification tree; if it is one (or more) continuous real number, it is called a regression tree.

To create this model, one question about a single variable is asked on each node: is variable X bigger than a value x? According to the answer, two different leafs are created. Sometimes, more than two leafs can be created, but it provides similar results than asking two consecutive questions.

This question is chosen internally by the tree, following different statistic techniques. On each node, a lot of different questions are tried and the one that provides most information (reduces the 'entropy' of the data) is chosen. The criterion to determinate which question is selected is usually the one that minimizes the root mean squared error or the mean absolute error of the answer. Sometimes, some other criteria are followed to separate concrete points of the data with more specific questions that do not minimize the error for this specific question but they do optimize the global performance. An example of the decision process made by a tree is showed on figure 3.3.

A single tree is able to separate correctly all the different data up to the point where only a single point is inside each node. This model is clearly overfitting and will not be able to predict correctly. To avoid this situation, some stopping criterion needs to be introduced. The problem is that this criterion is not easy to choose because the path followed by the tree is unknown. A similar procedure to the one explained before, using cross-validation, could be used, but the change of the parameters can affect the way on which the splits are made, obtaining a total different tree for each set of parameters. To correct this problem, several different trees are built and all the results are processed in order to choose the final one. One of these methods is called Random Forest and is explained below.

3.4 Random Forest

To solve the overfitting problem of single decision trees, a lot of different techniques have been developed. One of the most used is called Random Forest and is based on the idea of building several trees, all of them different to the others, to solve the same problem. As each tree is unique, the predicted values are going to be different. The final output is chosen to be an average of all the outputs in regression problems, or the mode for the classification ones.

To build different trees, different data sets are needed. To create this subsets of data, a technique called Bagging (or **B**ootstrap **agg**regat**ing**) is used. It divides the data set (N samples) into m subsets of n samples, by randomly sampling with replacement. If n=N, approximately 67% of the data is going to be original, and the rest is going to be copies of those ones.

Even with m different trees, if some variable is very important for one tree, it is going to be important for all the other trees and the resulting forest is going to be very similar one to another. In addition, evaluating all the variables on the split of each nodes takes a lot of time and resources. This problem is solved with a technique called Feature Bagging. On each node, only a randomly chosen subset of the variables is analyzed and the one that provides the most relevant information is chosen. This number is usually taken as $\sqrt{M}$, where M is the number of attributes or variables of the problem.

3.5 Metrics

All machine learning algorithms, once they have been trained, are able to predict outputs for a given set of new data. In order to measure how good this predictions are some metrics need to be defined. In machine learning, three metrics are the most widely used: MAE, RMSE and R^2.

- **Mean Absolute Error (MAE):** is defined in equation 3.3. It measures the difference between the model prediction y_i and the real value x_i of an example i.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{3.3}$$

- **Root Mean Squared Error (RMSE):** is defined in equation 3.4. The behaviour is similar as the one in MAE, but averages the square of the difference. This metric gives more importance to points with a high deviation, rising its value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - x_i)^2}{n}} \tag{3.4}$$

Both MAE and RMSE are absolute metrics and give information about the global error. Sometimes, relative measures of this parameters are used, comparing them with the maximum value of the measures.

- **Coefficient of determination (R^2):** is defined in equation 3.5.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{3.5}$$

Where:

$$SS_{tot} = \sum_i (y_i - \overline{y})^2 \tag{3.6}$$

$$SS_{res} = \sum_i (y_i - h_i)^2 \tag{3.7}$$

$\overline{y}$ is the average of all the data, y_i are the real values and h_i are the predicted values. In summary, R^2 compares how good is the model (blue area on the right side of figure 3.4) in relation to a model consisting on a horizontal line with the mean value of the considered data (red area on the left side of figure 3.4).

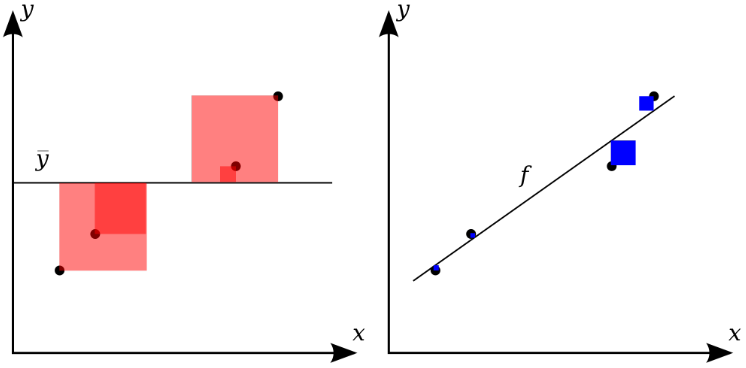


FIGURE 3.4: Coefficient of Determination

> A model that fits perfectly the data would have a value of 1 for the coefficient of determination. If it fits as well (or bad) as the horizontal line, it would have a zero value. It could even has a negative value if the model is worst that the horizontal line.

R^2 is usually the best parameter to measure how good a model adjusts to the real values. But it has a problem that advise against its indiscriminate use in some particular cases, such as this work. As it compares the model to a horizontal line with the mean value, this mean value has an enormous importance in the parameter.

To illustrate this, a cosine function is considered. The mean value of a period is 0, but if another different period of time is taken, the mean value changes a lot. For example, if the initial 1% of the period is considered, the mean value is around 1. If it is taken next to π, is going to be close to -1. If the first quarter of the period is considered, the mean value is close to 0.6.

A set of random points that belong, or are very close, to the cosine function are provided as training set. With all these data, that have a mean value of approximately zero, the model is created. If the R^2 of the complete model is calculated, the prediction is compared to a horizontal line at value 0.

If the next quarter of period wants to be predicted, the model is going to be compared to a horizontal line at 0.6, the mean value of this zone, and not at 0. The R^2 obtained on this prediction is comparing different things than the R^2 calculated for the training set. This will always happen, independently of the chosen sets, but its effect is negligible for models in which the local average is always similar to the global one.

In this work, the R^2 parameters has only been taking in consideration when the samples have been randomly shuffled, making the local average always constant. Otherwise, only absolute metrics have been considered, such as MAE and RMSE.

3.6 Scikit learn: RandomForestRegressor

In this work, the RandomForestRegressor method of the library Scikit-learn has been used. It is one of the most used libraries in machine learning because it provides really good results in a short computational time. In addition, it is easy to use and it has a complete user guide with a lot of different examples available. There are some other more advanced libraries that provide better results, but they are usually more difficult to use and its use it is not recommended for amateur users.

This algorithm has some parameters that can be defined by the user [11]. The default values of the algorithm are valid for most of the examples and changing them does not improve drastically the results. If a model does not work correctly with the default values, it is usually due to the provided data, not to the parameters of the algorithm. Nevertheless, the parameters of a model that works correctly can be tuned to optimize the model.

Chapter 4

Applying ML techniques to a real-world scenario

4.1 State of the art

The scope of this work is the power consumption prediction of the HVAC systems in buildings with geothermic installations. In the literature, no similar work has been found. Similar approaches have been followed to predict energy consumption in buildings without geothermic wells. A review of the different algorithms that have been used can be found in [15].

In [13], an hourly prediction of energy consumption is performed using Random Forest and the results are compared with single regression trees and SVM. Random Forest was proven to provide the best results when the optimum set of parameters is chosen. It is worth mentioning that, in this paper, the energy consumption on the hour before the considered time is taken as an extra variable, improving significantly the results. In [12], a similar study is performed at the University of Florida, confirming the superiority of Random Forest over regression trees and SVM.

Some other authors have preferred to use ANN on its predictions, obtaining good results as well. For example, in [4] a prediction using ANN has been made for a passive solar building. In [3] the authors have proven the superiority of ANN over detailed model simulations, that traditionally had been used.

Improving the traditional Neural Network has also been tried by several authors. For example, in [6] the authors have used an extreme deep learning approach to train the network, obtaining good results. Some other authors

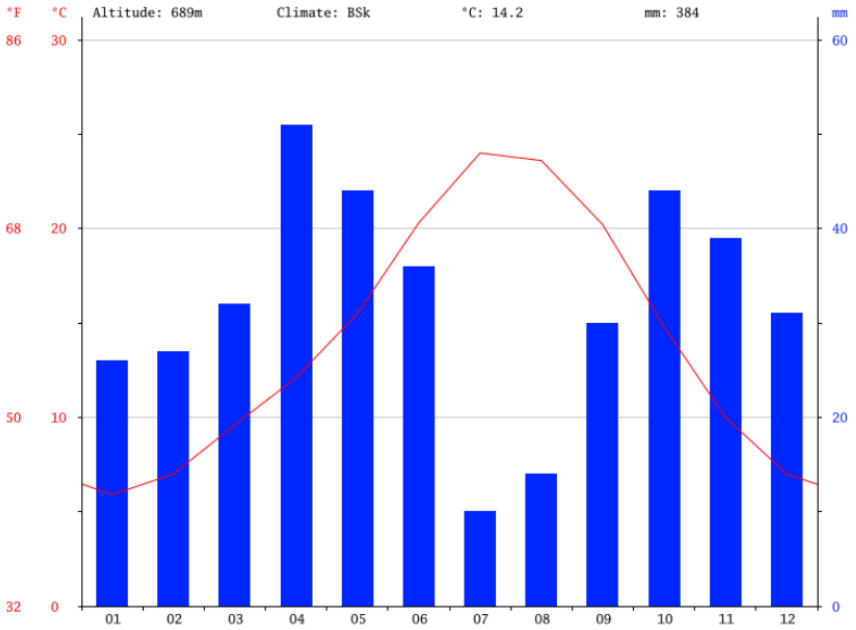


FIGURE 4.1: Diagram of the climate in Albacete

have focused on retraining the network adding new data, trying to improve the long term predictions ([8] and [14]).

4.2 Description of the geothermal installation

The installation used for this work is responsible for the HVAC needs of an enterprise building placed next to the Airport of Albacete, Spain (38º56’31.7"N 1º53’07.3"W). The company has two buildings: the office, with two floors, and the warehouse.

Albacete, situated in the south-east of Spain at 689 meters of altitude, has a dry Mediterranean weather. It is characterized by the low amount of precipitations during the year and the strong temperature variations, both during the day as during the year. The mean temperature is 14,2ºC and the mean precipitation is 384mm (Figure 4.1).

This enormous variations in the temperature make Albacete an excellent place to install geothermal systems. The ground temperature near the surface is around 17ºC, making it possible to use wells both for heating and refrigeration and reducing the thermal exhaustion of the ground. During the summer, heat is going to be transfer to the ground whereas in summer it is going to be extracted.

A diagram of the geothermal installation of the building is shown in figure 4.2. The system has two working modes: during the winter, it heats up the building and, during the summer, it cools it down. It is composed by two different subsystems: one provides the Sanitary Hot Water (SHW) and, the other one, the air conditioning, using a radiating floor and an AHU (Air Handling Unit).

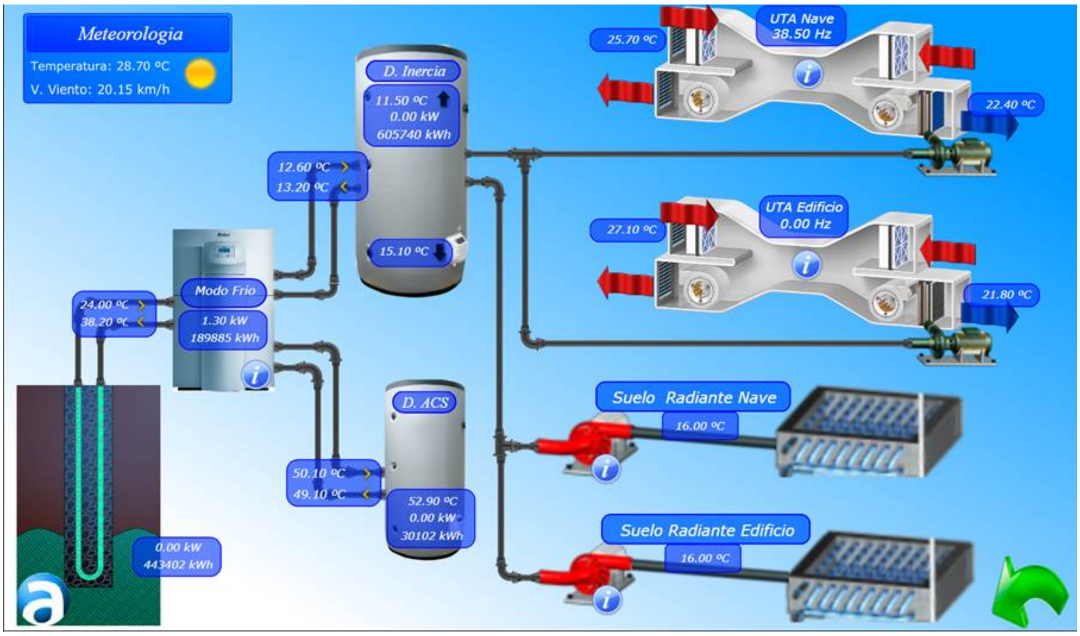


FIGURE 4.2: Diagram of the thermal installation

Four main elements compose the HVAC system: the wells, the heat pump, the inertia tank and the radiating floor and the AHU. Different fluids go through all these elements exchanging heat and changing its temperature. The functioning of the HVAC system during the heating mode is considered from now on.

This installation has nine 105 meters-depth wells, separated 10 meters one from each other. In the area, the ground below 20 meters has a constant temperature of around 17ºC. [9] As the ground temperature remains (almost) constant, the amount of heat exchanged with the ground is controlled by the flow that goes through the wells.

The heat pump is powered by electrical energy and is responsible for the control of all the heat exchanges between the installation and the ground. It works following some pre-programmed logic, trying to maintain the temperatures of the tanks into the desired interval.

The inertia tank has a capacity of 1000 litres and a temperature around 60ºC in the heating mode (and around 11ºC in the cooling mode). This temperature is abnormally high for a geothermal installation, because it reduces the global efficiency of the system. A temperature around 45ºC is high enough to heat up the place, rising the global efficiency of the system, as the difference of temperature with the ground is lower.

This hot water reaches the AHU and the radiant floor heating up the building. After the heat exchange, the cooled fluid reenters into the inertia tank, reducing its mean temperature. In order to avoid the temperature of the tank to be too low, some heat needs to be given to this deposit. This is done by the heat pump, that uses electrical energy to extract geothermal energy from the ground, using a Rankine cycle.

The building is separated in two different areas: the office, that has two floors, and the warehouse. Each area has an AHU and a radiant floor connected directly to the inertia tank. Each device can be configured individually, providing different temperatures for each area of the building.

As the SHW system is independent from the inertia one, the energy consumption used by the inertia tank can be calculated. This energy is only used to keep the building with the desired temperature and varies with the atmospheric conditions. Being able to predict this energy consumption, using the meteorological predictions, can be used to optimize the functioning of the heat pump and reducing the total electric consumption of the building.

4.3 Data of the building

The control system of the HVAC needs some information in real time in order to work properly. Specifically, it needs to know the temperatures and the flows entering and leaving the heat pump towards the tanks and the wells in order to control the functioning of the heat pump. In addition, the exterior temperature and the total electrical energy consumption (including all the pumps of the system) are also known.

To measure the flow, several flow meters are placed along the pipes. They measure the amount of fluid that has gone across the instrument in a period of time, so they do not provide a continuous signal. With these measures, the energy consumption can be calculated with the equation 4.1.

$$E = \dot{m} \cdot \Delta t \cdot C \cdot \Delta T \tag{4.1}$$

Where $\dot{m}\Delta t$ is the mass of fluid that has gone through the heat pump, experimenting a variation of its temperature ΔT. C is the heat capacity of the fluid (a mixture of water and ethylene-glycol).

As the main interest of this work is to minimize the energy provided by the heat pump to the inertia tank, this energy can be calculated using the flow leaving the tank towards the heat pump. This amount of fluid multiplied by the difference of the temperature between the entrance and the exit of the heat pump, provides the consumed energy used to heat up the building.

4.3.1 Dates

All the data measured in the building have been collected on different instants, starting on May 2015 and finishing on July 2018. The matrix to be provided to the algorithm needs the variables to be on the same time instant, so a common reference of time is needed. This reference has been taken as every hour on the hour in the UTC reference: one o'clock, two o'clock, three o'clock and so on.

The meteorological measurements provided by the AEMET are already on this time reference so no interpolation is needed. Nevertheless, all the measurement of the building, such as the temperatures or the energies, have been measured on the local time of Albacete (UTC+1). In addition, the Daylight Saving Time (DST) is active on Spain, so another hour needs to be added during the summer. Once all the data is placed in the UTC system, they need to be interpolated into the hours on the hour in order to have all the measurements on the same time instants. The way on which this has been made is explained later.

In addition, some variables related to the time instants have been taken into account. The month and the day of the month contain information about the annual meteorological weather. The hour of the day is related to the working hours of the employees and the daily changes. In addition, a variable indicating if the considered date corresponds or not to a working day and a variable indicating the day of the week, from Monday to Sunday, have also been considered. Another variable that indicates if the DST is active or not has also been taken into account as it is related to the working mode of the installation.

4.3.2 Energy of the inertia tank

The original provided data is plotted in figure 4.3. It can be seen that is grows all the time, with some plain zones. There are also some zero values that do not represent the physic reality. The plain zones may represent two things. On the one hand, they may mean that no energy has been consumed during this time, corresponding to periods where the heat pump has been off. This would correspond to short periods of time and the energy consumption once this period is over would be of the same order that any normal consumption.

On the other hand, there are some plain zones originated by instrumental failures. This would correspond to plain zones followed by an enormous peak, many times bigger than the normal consumption for its period of time. This failures may be produced by a failure storing the data. During the time the instruments are not working, no energy consumption is added and a plain zone appears. When the system starts working again, all the energy consumed during this time is suddenly added to the measures, producing a jump on the energy that corresponds to an enormous peak of power. This kind of behaviour can be seen in April 2016 on the figure 4.3.

The first type of plain zones correspond to the reality and provide information about the system whereas the second are clearly wrong and must be eliminated from the data set. In order to identify and eliminate these unreal points, each point has been classified into one of four types:

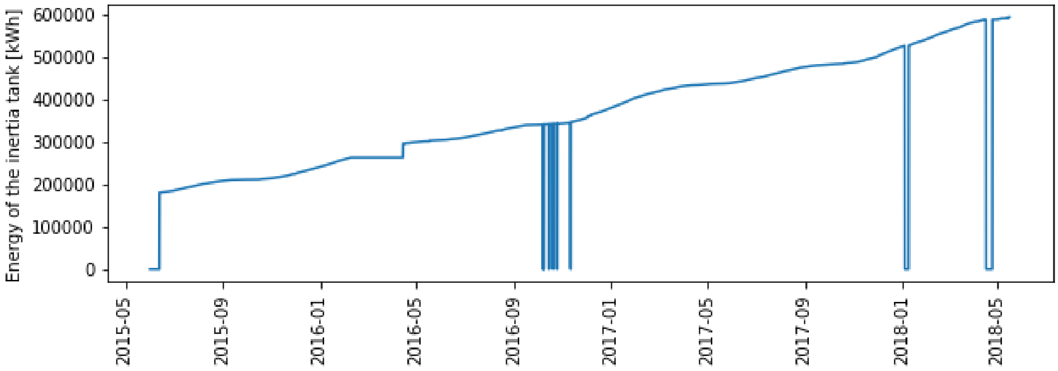


FIGURE 4.3: Inertia tank energy without preprocessing

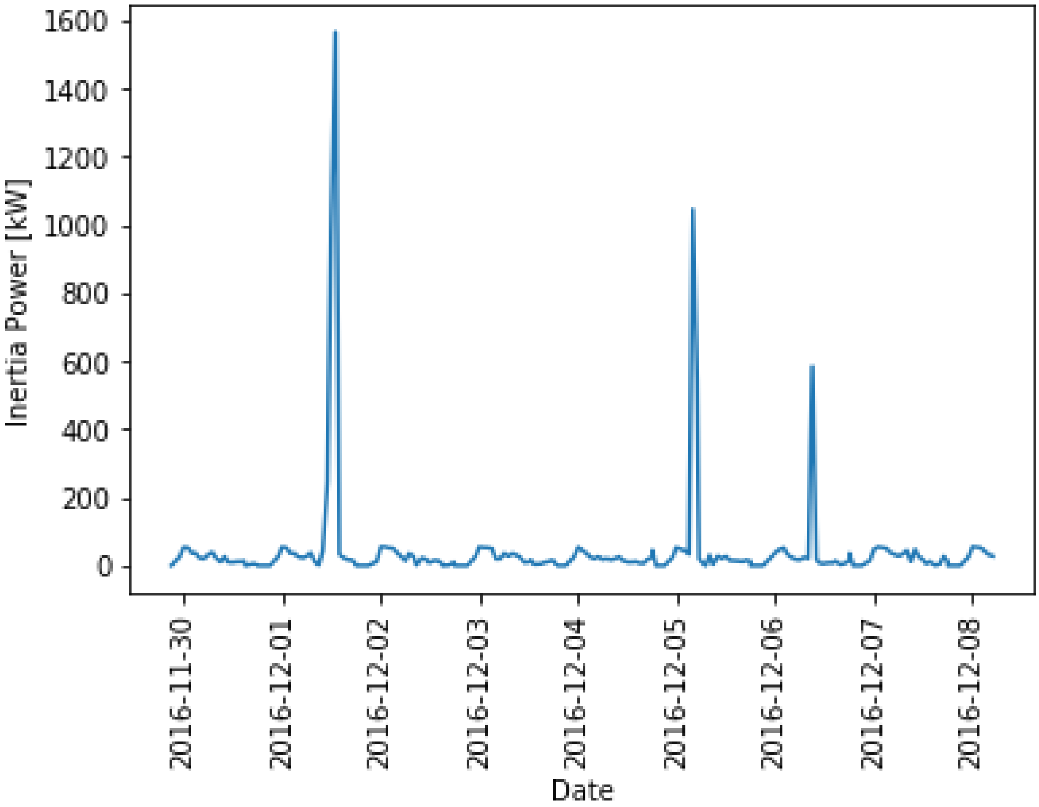


FIGURE 4.4: Random peaks classified as type 4

- **Type 0:** this category contains all the correct points that are going to be used to create the model.
- **Type 1:** points where the energy is almost zero (a minimum value of 50 kWh has been set to avoid numerical problems).
- **Type 2:** plain zone that precedes a peak classed as type 3.
- **Type 3:** peak after a plain zone. The criterion to separate correct peaks from the unreal ones is: if the peak of power just after a plain zone is 4 times bigger than the power on the next measure, two measures after the peak, then the peak is not correct and belongs to the type 3.
- **Type 4:** the rest of the points where the instant power is bigger than the maximum admissible by the machine (50kW). These type of points are shown in figure 4.4.

Working mode

Albacete suffers really high differences in temperature during the year, from some negative degrees in winter to more than 40ºC in summer. The geothermal installation can be used in places like this both for heating and cooling. Both working modes work differently and the energy consumption differs from one to another.

The change of the working mode is made manually by an employee and is chosen according to the local weather. The dates of change are not preprogrammed and can not be predicted. The main difference between both working modes is the temperature inside of the inertia tank, showed in figure 4.5. During the winter, the temperature is around 60ºC, and around 15ºC during the summer. 60ºC is a really high temperature for this type of installation, where temperatures around 45ºC are enough to heat up the building, improving the global efficiency of the installation.

To calculate the exact date of the mode change, intervals of 5000 samples (50 days approximately) have been taken. The number of samples that are higher than a reference value (27.5ºC, approximately the mean value) have been counted. If there are more points above this reference value than below, this point corresponds to the heating mode and vice versa. As 50 days have been taken as the sample, little local variations in temperature are overseen.

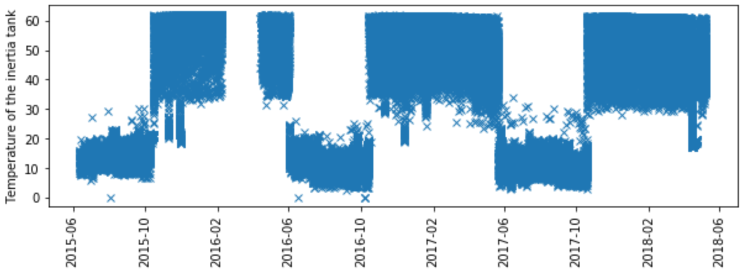


FIGURE 4.5: Inertia Temperature

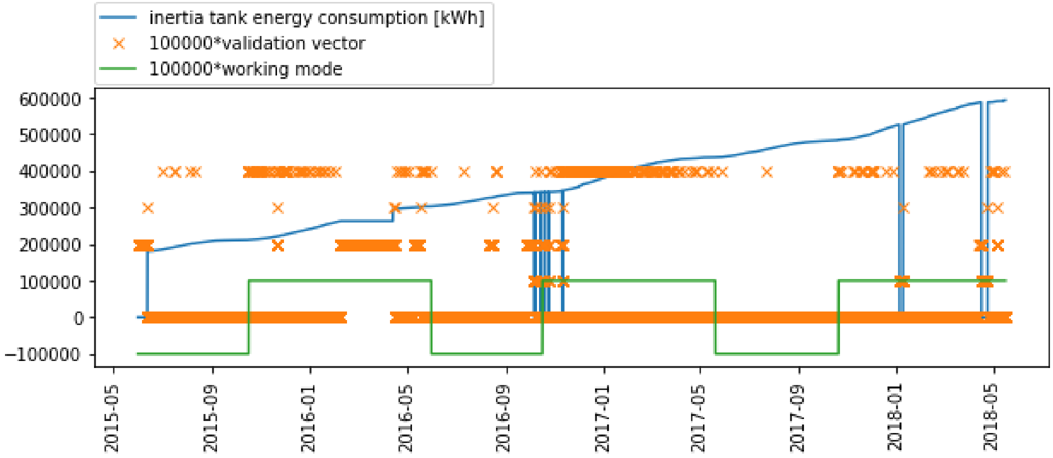


FIGURE 4.6: Summary of inertia tank data

During the summer time (cooling), the sign of the power has been changed and considered negative.

Summary of the data of the inertia tank.

Figure 4.6 shows a summary with all the data of the inertia tank. The blue line corresponds to the original energy consumption. The orange one is 100000 times the validation vector, formed by an integer from 0 to 4 indicating the type of point. The working mode multiplied again for 100000 is showed on the green line.

The total number of points is 25936, with 22516 (86%) belonging to type 0. The rest of points are mostly type 2 (2514, that corresponds to 6,69%), with very few for the other categories. From the figure, it may look like there are a lot of type 4 points, but they are only 536 (2%). It is because they are separated to one another whereas the type 2 points are all consecutive points.

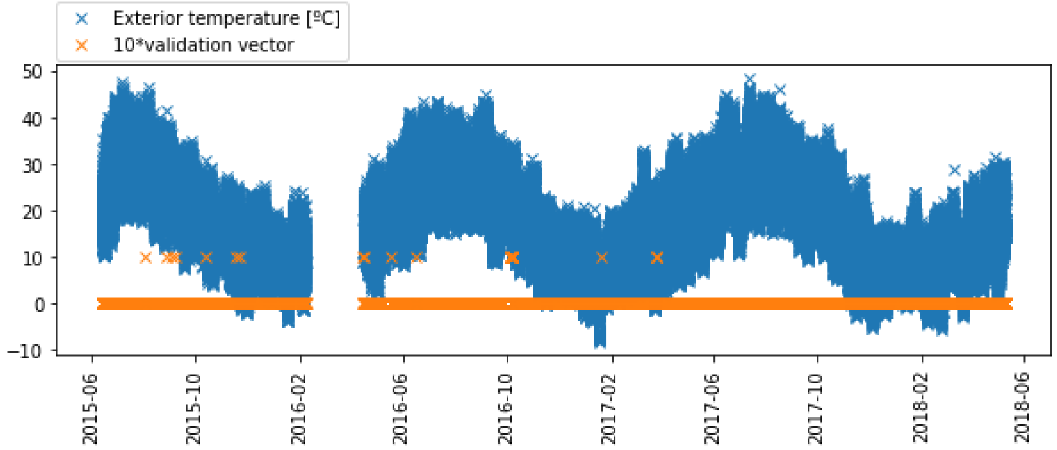


FIGURE 4.7: Exterior temperature

4.3.3 Exterior Temperature

The building also provides the measurement of the exterior temperature. Data are taken each 15 minutes approximately and are shown in figure 4.7. Most part of the data provided is valid, and only a 0.1% of the data has been classified as type 1, due to some different reasons.

The temperatures measured after a long time without collecting data are not correct. It looks like it is the last temperature measured before the break the one that is put on the first measure after the break. There are also some points where the temperature is exactly 0 degrees and that may correspond to instrumental failures. In addition, some particular zones where the temperature remains constant for a long period of time have also been removed.

4.3.4 AEMET data

AEMET (National Agency of Meteorology) is the Spanish agency responsible for providing the national meteorological services. It is also responsible for the weather forecasts and the diffusion of studies related to the weather. AEMET counts with almost 800 stations all around the national territory, compiling information such as temperature, pressure, precipitations, wind speed and direction, radiation, humidity...

The data used on this work have been collected in the station of Los Llanos, in Albacete, less than a kilometre away of the studied building. All

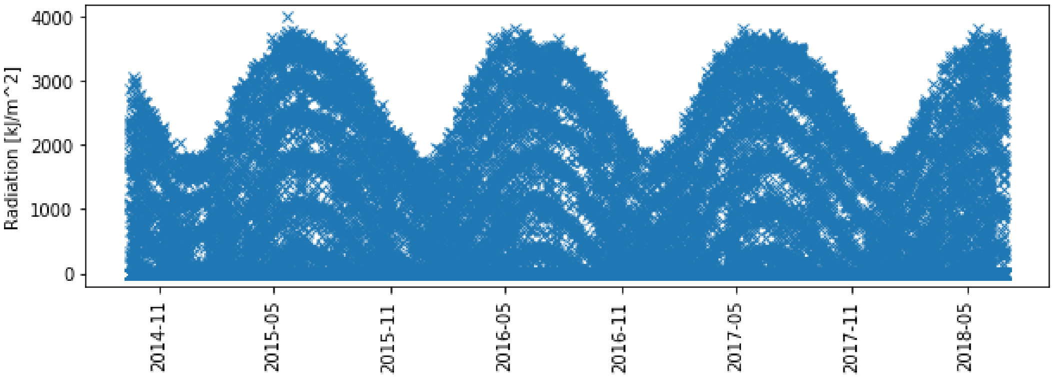


FIGURE 4.8: Radiation

these data is provided on the UTC time every hour on the hour, so no adaptation to the common date is needed. The original data contained a lot of missing points on the measurements, so all these points have been classified as type 1.

The provided data by the AEMET are the temperature in degrees Celsius, the radiation in kJ/m^2, the percentage of relative humidity, the speed and direction of the wind, and the dew point information. The temperature has not been taken into account because this data is already provided by the building. The humidity and the dew point information have a lot of missing values, around the 30%. This information is considered less important than the rest of the variables and is not going to be used in the first place. If a more precise model is needed, this data could be added. The radiation data has almost none missing values but it has to be noticed that during the night its value is zero. The radiation is showed in figure 4.8, where the annual variations are obvious.

4.4 Power calculation

Up to this moment, all the data has been collected and each point has an assigned integer number indicating if the data is correct or not. Each variable is measured in the UTC time but each one has the measures on different time instants. In this section, the way on which the different variables have been put together on the same time is explained.

The final goal is to create a matrix where each column represents the different values of a variable on the different time instants and where each row represents a variable, being the power consumption the last one. With this matrix, the machine learning algorithm is able to create a model for the problem.

To create this matrix, some considerations need to be made. First of all, every data appearing on the final matrix needs to have valid values for all the considered variables. If a variable has a wrong value for a time instant, that instant is not taken into account for the final matrix. Changing a value from one time instant to another requires an interpolation, that needs the two points around the considered time to be correct. To do that, the following logic process has been followed:

- A loop through all the dates on which the power wants to be known is performed. If a value of power is found for this instant, the values of all the variables interpolated to this time are stored in the final matrix. If there is a value of any variable that is not correct for this instant, the loop exits the current iteration and passes to analyze the next time instant.

- The first variable to be checked if the most important one, the energy consumption. A nested loop is performed trough all its values and the one that is right after (or equal) in time to the considered instant is taken. The power has been calculated as:

  $$P(t_i) = \frac{E(t_i) - E(t_{i-1})}{t_i - t_{i-1}} \tag{4.2}$$

  where the subscript 'i' makes reference to the considered instant of time. The first thing to do is checking that the values of energy in time 'i' and 'i-1' are valid (type 0). In addition, on time 'i-1' data of type 3 is also valid because the peak has been generated on time 'i-2'. These requirements are not enough to guarantee the validity of the point. If the time between 'i' and 'i-1' is too big (on this work bigger than 3 hours), the calculated power will not be representative of this instant and has not been considered. If all the requirements are fulfill, the power is calculated and the corresponding sign is added: positive for heating and negative for cooling.

- If the power is calculated, the rest of variables need to be checked. The first one is the exterior temperature, following a similar procedure. The first point right after (or equal) to the considered date is taken. The validity of this point and the previous one are inspected and, if both are correct, a linear interpolation is performed. In addition, a condition if a time difference shorter than 3 hours has also been added to reduce the interpolation error. If these conditions are not satisfied, the considered time instant is rejected and the next time instant of the energy is studied.

- Points with correct values of energy and temperature have been calculated so far. They have been stored in a matrix that can already be introduced to the machine learning algorithm. However, some additional variables can be added to improve the accuracy of the model.

- The radiation seems to be another important variable and must be included too. To do so, a similar procedure has been followed, checking the validity of the point in that time instant and storing all the variables in a new matrix. This matrix is going to be smaller than the previous one because it has more restrictions coming from the validity of the radiation data.

4.4.1 t-1 power calculation

As the geothermal installation is commanded by a heat pump, that follows a programmed logic, is seems interesting to add the power consumption on the previous hour as a feature of the model. This variable contains information about the logic of the pump and its internal operation. Variables containing information on previous time instants has been used previously, like in [13], significantly improving the results. In this work, models with and without this variables have been made and a comparison between both results maintain the superiority of model with this extra variable.

Chapter 5

Obtained results

In order to find the set of variables that provides the best results, some different sets have been tested. All the models have been performed with the default values for the parameters of the RandomForestRegressor algorithm of the Scikit-learn library [11].

As the parameters are not going to be optimized, the total set has only been separated into two different sets: train set, with the 80% of the data, and test set with the 20% remaining. In addition, two different types of models have been conducted. On the one hand, the first 80% of the ordered data, from the beginning of the data in may 2015 to November of 2018, is taken as the train set, trying to predict the remaining 20%, that is the test set and evaluates the accuracy of the model. This case corresponds to the real case scenario, where the future consumption is predicted from meteorological forecasts.

On the other hand, taking the ordered data may induce bias on the train set and the model will not predict correctly because it would have been created using bias data. In order to correct this error, it is an usual and recommendable practise to shuffle the data to be provided to the algorithm, eliminating this bias. In this case, as the local average is similar to the global one, the R^2 parameter can be used and provides a good information about the accuracy of the model.

5.1 Model with exterior temperature

The results found using the external temperature as the only variable are shown in figure 5.1 and tables 5.1 and 5.2.

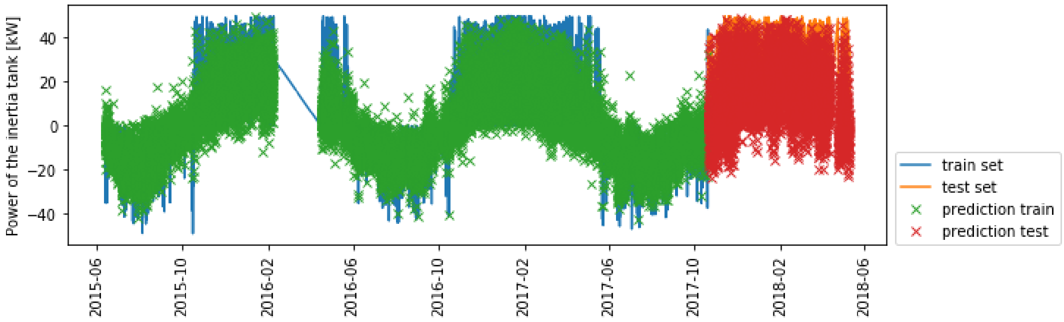


FIGURE 5.1: Model with the external Temperature as the only variable

MAE train	4.3	RMSE train	6.3
MAE test	11.8	RMSE test	14.9

TABLE 5.1: Results with External Temperature as the only variable

MAE train	4.6	RMSE train	6.5	R^2	0.88
MAE test	9.5	RMSE test	12.5	R^2	0.55

TABLE 5.2: Results with External Temperature as the only variable, with shuffled data

It can be seen how the model does not follow correctly the power curve and it is not able to predict correctly. It is caused because the model has not enough information. A temperature of around 15ºC can correspond to midday of a winter day or midnight of a summer day and the model is not able to make the difference between them. To correct that, some additional information needs to be added.

5.1.1 Model with exterior temperature and radiation

The radiation provides the model with information related to the hour of the day and the day of the year. The results found are shown in figure 5.2 and tables 5.3 and 5.4.

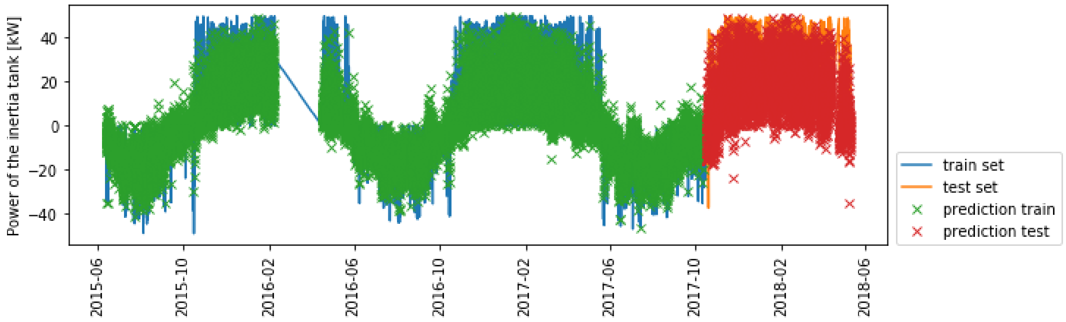


FIGURE 5.2: Model with the external Temperature and Radiation

MAE train	3.6	RMSE train	5.3
MAE test	10.5	RMSE test	13.8

TABLE 5.3: Results with External Temperature as the only variable

MAE train	3.7	RMSE train	5.5	R^2	0.92
MAE test	8.2	RMSE test	11.2	R^2	0.66

TABLE 5.4: Results with External Temperature and Radiation, with shuffled data

The model has slightly improved its performance but it is not good enough. The tree is giving around 89% of the importance of its training to the exterior temperature while only the remaining 11% is being given to the radiation. The tree fits correctly the train set but the created model is not able to predict correctly future values. To correct this problem, some other variables related to the dates have been added to the tree.

5.2 Model with exterior temperature, radiation and time related variables

Trying to improve the accuracy of the model, some time related variables have been taken into account. As the tree is able to neglect the contribution of useless variables, all the representative ones have been introduced.

Firstly, the hour of the day, that provides information about the time employees are in the building, as well as daily variations. The month has also been added to represent yearly variations as well as the week of the day for the weekly ones. The variable representing if the considered day people worked at the office has been neglected as the geothermal system follows the same logic every day and is not switched off during the holidays. A variable indicating if the DST is active (during the summer) or not has also been provided to the algorithm as it is closely related to the working mode of the heat pump. The results are resumed in figure 5.3 and tables 5.5 and 5.6.

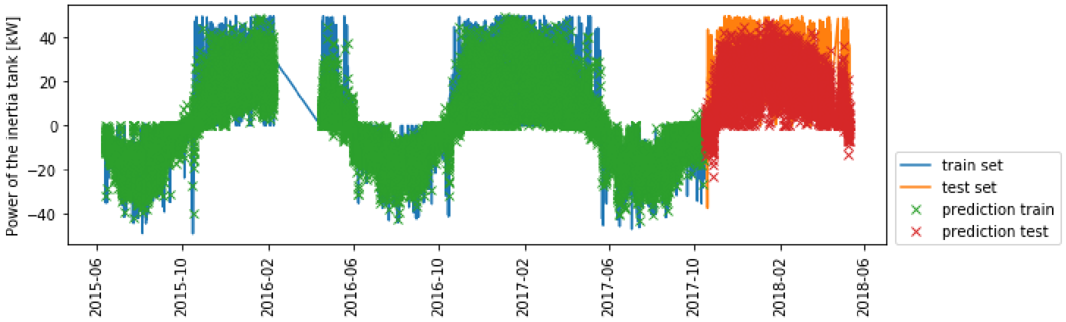


FIGURE 5.3: Model with the external Temperature, Radiation and time related variables

MAE train	2.4	RMSE train	3.7
MAE test	9.7	RMSE test	13
Importance Temperature	19.5	Importance Radiation	4.4
Importance Hour	5.0	Importance Month	17.1
Importance Weekday	3.9	Importance DST	50.2

TABLE 5.5: Results with External Temperature as the only variable

MAE train	2.5	RMSE train	3.9	R^2	0.96
MAE test	6.5	RMSE test	9.2	R^2	0.77
Import. Temperature	65.9	Import. Radiation	3.6	Import. Hour	4.3
Import. Month	9.9	Import. Weekday	3.5	Import. DST	12.8

TABLE 5.6: Results with External Temperature and Radiation, with shuffled data

As reflected in the features importance, the model with the ordered data is acting differently than the shuffled one. Shuffled data provide a better model and the results are better than the ordered one. This is mainly caused by the way on which the training has been made, explicitly, by the importance given to the DST variable. The ordered data model is giving more than the 50% of the importance to this variable, which is not intuitive and incites to think that is not working correctly. Some changes have been made in the inner parameters of the algorithm, trying to improve the performance of the model, but the results are always similar. No explanation has been found to this difference of behaviour between both models, but a bias in the data. Trying to avoid that requires the use of more variables that can capture this bias.

5.3 Results with Power on 't-1'

Adding the power consumption on the previous time instant provides information to the system about the logic control of the heat pump. To calculate it, a new restriction has been added when the difference in time between two measures is bigger than 3 hours. Like that, only if the power in the previous time instant has been calculated, the point is given to the algorithm. If there is a jump in point 'i' longer than 3 hours, the power on this time is not calculated as 'i-1' is too far. In point 'i+1' the power can be calculated, but not the one on the previous time. It is the point 'i+2' the first to be given to the algorithm, as the power in 'i+1' can been calculated.

Once the data is given to the algorithm, the train set is used to train the model. But, in this case, a new problem appears. In order to predict the power consumption on a given time, the power on the previous hour is required. That implies that the predictions have to be made hour by hour, calculating the power in the previous time in order to add it as a feature for the next hour. The main disadvantage of that is that the power consumption within x hours can not predicted directly, and all the previous x-1 values also need to be calculated.

5.3.1 Model with power on t-1 given

The first step is to confirm if the addition of this feature actually improves the accuracy of the model. To check it out, instead of predicting a value and use it as a feature for the next instant, the measures of power are taken directly as a feature for every time. If the power at time i wants to be predicted, the measured power at time i-1 is provided instead of obtaining it as a previous prediction. Like that, the error does not accumulate and the accuracy of the model for short-term predictions is tested. The results are shown in figure 5.4 and tables 5.7 and 5.8.

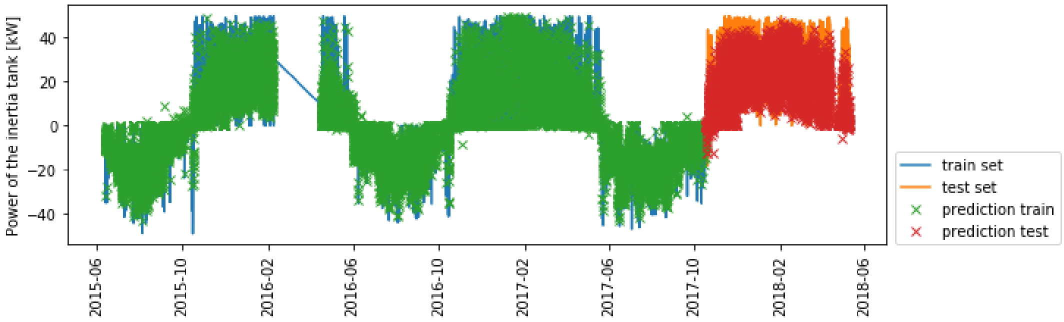


FIGURE 5.4: Model with the previous power consumption

MAE train	1.7	RMSE train	2.8
MAE test	6.9	RMSE test	9.3
Importance Temperature	10.5	Importance Radiation	2.0
Importance Hour	2.5	Importance Month	1.5
Importance Weekday	1.5	Importance DST	0.5
Importance Power t -1	81.5		

TABLE 5.7: Results with t -1 power consumption

MAE train	1.8	RMSE train	3.0	R^2	0.97
MAE test	4.6	RMSE test	7.1	R^2	0.85
Import. Temperature	12.4	Import. Radiation	1.6	Import. Hour	2.1
Import. Month	1.3	Import. Weekday	1.3	Import. DST	0.4
Import. t -1 power	80.8				

TABLE 5.8: Results with t -1 power consumption with shuffled data

The results show that this additional feature is really characteristic, as the tree gives it more than a 80% of importance. The results have also improved, achieving the best results so far. In addition, both trees have given the same importance to the features, which means that both trees have been trained similarly.

The fact of finding better results with shuffled data proves that they are bias and that some extra variable that contains information about this bias needs to be added in order to improve the results of the tree. The main difference in the training is the importance given to the exterior temperature, so maybe a variable containing the temperature on the previous instant could improve the results.

5.3.2 Model with power on t-1 predicted

A real functioning model must be able to predict the power consumption on a period of time only from meteorological data, in this case exterior temperature and radiation. To achieve this goal using the power consumption on the previous time as a feature, each time instant has to be predicted one by one, in order to use this prediction to calculate the power consumption in next time instant.

The main problem of this model is that the error gets accumulated. To calculate the power consumption in time i, the power on i-1 is needed. This one has been calculated from the prediction made for i-2, that used the one made for i-3 and so on. With this procedure, the consecutive values are predicted using a non-correct value of the previous power.

If the error is small, the prediction will not be too far from the real value and the next prediction either. On the other hand, a big error in one prediction will cause the next one to be calculated from a very different set of variables, raising the error over and over. This restricts this type of model to short term predictions.

MAE test	8.6	RMSE test	11.0

TABLE 5.9: Results with power on t -1 predicted

The results showed in table 5.9 are worst than the previous ones in figure 5.7 as expected. Nevertheless, the results are better than the ones that have not got the power consumption in t -1 as a feature, proving the good influence of this parameter to the model. The importance of the model are exactly

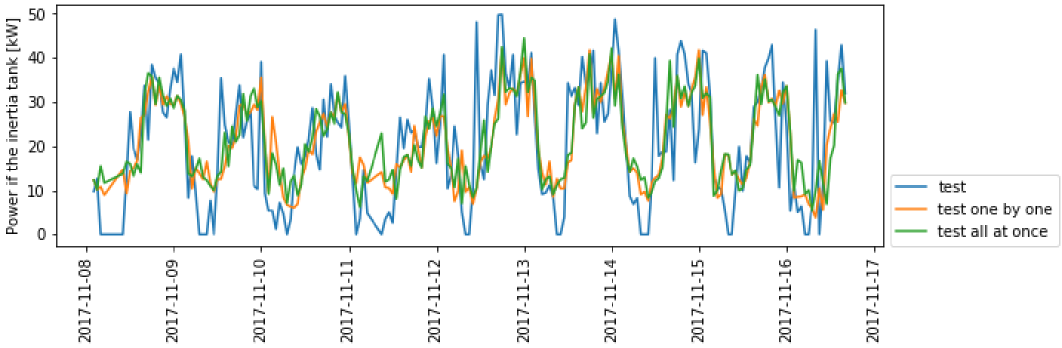


FIGURE 5.5: Models with the power on the previous time

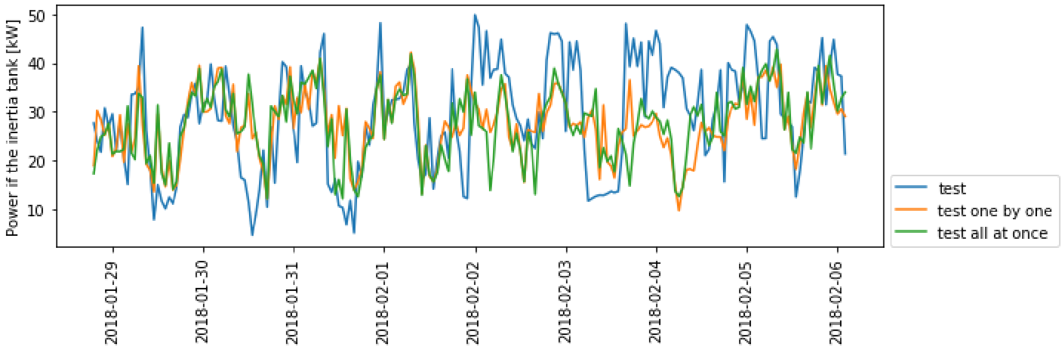


FIGURE 5.6: Models with the power on the previous time

the same than in the previous example because the tree has been trained with exactly the same data.

Some graph showing the results for both models are showed on figures 5.5, 5.6, 5.7, 5.8 and 5.9. The blue line corresponds to the real data of the installation. The green line represent the results found when all the power on the previous time was provided whereas the orange one represents the results obtained when this power is predicting on each time instant. It can be seen how the models follow the trend of the data but they are not able to predict accurately the extreme results, both zeros and peaks.

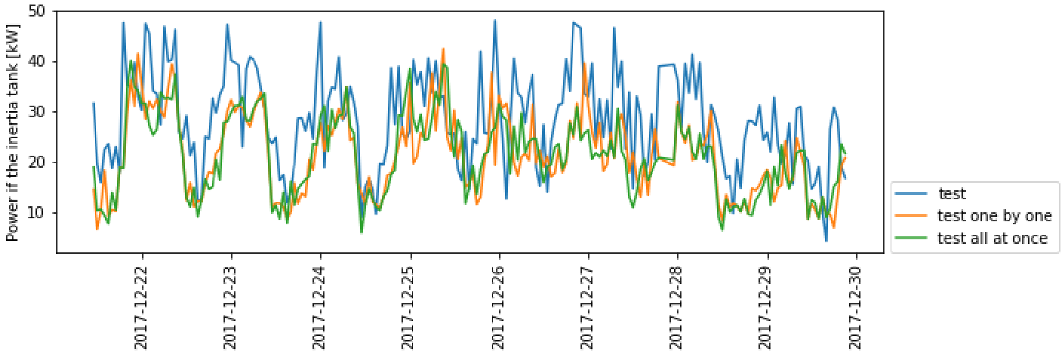


FIGURE 5.7: Models with the power on the previous time

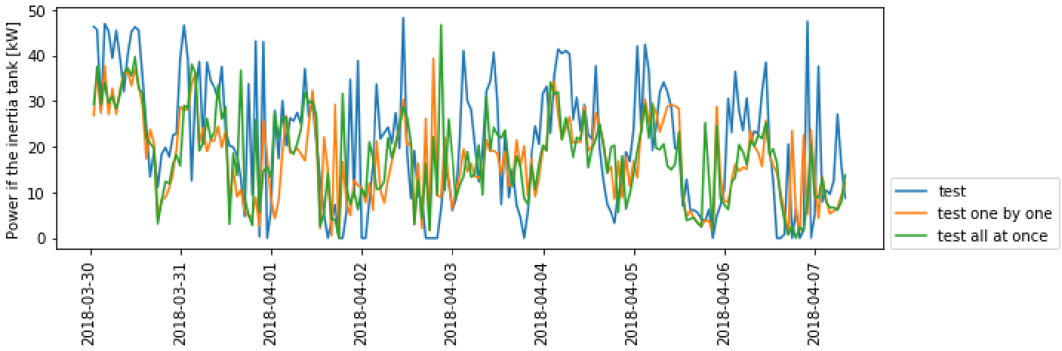


FIGURE 5.8: Models with the power on the previous time

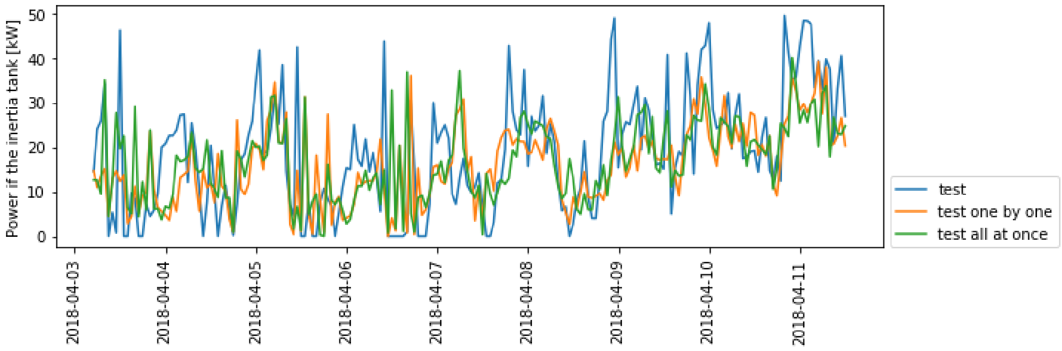


FIGURE 5.9: Models with the power on the previous time

Chapter 6

Conclusions and future work

This work provides a model to predict the power consumption of a building with a geothermal installation from meteorological predictions. Among all the machine learning techniques, the Random Forest Regressor algorithm has been used as it provides some information about its internal functioning. The correct set of chosen variables to be provided to the algorithm plays a key role in the elaboration of the model, as they contain the information required for the tree to train the model.

In this work, adding the power consumption on the previous hour improves the accuracy of the predictions. Shuffling the data also improves its performance in relation to ordered data, but the predictive power is lost. That may indicate that there are some type of bias on the data.

The best results have values of 5 for the MAE and 7 for the RMSE. As the Random Forest algorithm is only able to predict values inside the range of the provided data, the maximum possible values are +50kW and -50kW. This errors correspond to around a 5%, which may be acceptable for some applications. For this particular case, the results are not completely satisfactory as the model is going to be used for optimizing the power consumption of the heat pump by optimizing its control system. It has to be noticed that the sign of the prediction is not an useful information for the heat pump, as its working mode is changed manually. Having that in mind, the relative error duplicates as the extreme values are cut in half. A prediction with an error of 10% could be used optimize a little the heat pump, but a better model is needed to make an actual difference with the optimization.

The low accuracy obtained may have been caused by different factors. Firstly, the election of the variables and the interpolating procedure used to calculate the power may have not been optimal. Only the data around the hour on the hour has been used, while all the intermediate points have been neglected. Besides, only the external temperature and the radiation have been taken into account. Using more weather related variables could improve the accuracy of the model, as well as variables with information on previous time instants.

In addition, as the data to be predicted obey the logic of the control system, they may be harder to predict, as their relationship to the external variables is not direct. Taking a look at the power evolution, it does not follow any periodic trend and it is not easy to find any relationships with the external variables. There are some points where there is no energy consumption that are not predicted correctly by the algorithm. The Random Forest Regressor, due to the way on which it works, is not able to predict zero values in this type of problems, always predicting low values.

6.1 Future work

As the results found in this work are not accurate enough to be used in the optimization of the heat pump, some future work must be performed. Regarding the evolution of the model, the first thing to do could be trying new variables containing information on previous time instants. For example, the temperature on the previous hour, or the power consumption on the previous 3 hours could be used.

In addition, some other machine learning algorithms could be used trying to avoid the problems of the Random Forest Regressor. Some authors have already used Artificial Neural Networks for similar works, obtaining good results predicting the energy consumption. Predicting this energy consumption, instead of the power consumption, could be another thing to do as work, as it does not require to interpolate the energy.

Another possible improvement could be achieved by treating the discrete variables as categories. To do that, some specific library should be used such as CatBoost. This gives more importance to categorical variables over the continuous ones, training the model in a different way and obtaining usually better results.

Bibliography

[1] C. Deb, L. Eang, J. Yang, and M. Santamouris. Forecasting diurnal cooling energy load for institutional buildings using artificial neural networks. *Energy and Buildings*, 121:284–297, 2016.

[2] S. Ferlito, M. Atrigna, G. Graditi, S. De Vito, M. Salvato, A. Buonanno, and G. Di Francia. Predictive models for building's energy consumption: an artificial neural network (ann) approach. In *Proceedings of the 2015 XVIII AISEM Annual Conference*, 2015.

[3] A. Hernández and F. Sanzono. Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. *Energy and Buildings*, 40:2169–2176, 2008.

[4] S. Kalogirou and M. Bojic. Artificial neural networks for the prediction of the energy consumption of a passive solar building. *Energy and Buildings*, 25:479–491, 2000.

[5] H. Khosravani, M. Castilla, M. Berenguel, H. Ruano, and P. Ferreira. A comparison of energy consumption prediction models based on neural networks of a bioclimatic building. *Energies*, 9, 2016.

[6] C. Li, Z. Ding, D. Zhao, J. Yi, and G. Zhang. Building energy consumption prediction: An extreme deep learning approach. *Energies*, 10, 2017.

[7] C. Roldán-Blay, G. Escrivá-Escrivá, C. Álvarez-Bel, C. Roldán-Porta, and J. Rodríguez-García. Upgrade of an artificial neural network prediction method for electrical consumption forecasting using an hourly temperature curve model. *Energy and Buildings*, 60:38–46, 2013.

[8] A. Shabani and O. Zavalani. Hourly prediction of building energy consumption: An incremental ann approach. *European Journal of Engineering Research and Science*, 2(7):27–32, 2017.

[9] "`http://www.enesco.es/geotermia/`".

[10] "http://qjegh.lyellcollection.org/content/50/2/187/tab-figures-data".

[11] "http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html".

[12] Z. Wang, Y. Wang, R. Zeng, R. S. Srinivasan, and S. Ahrentzen. Random forest based hourly building energy prediction. *Energy and Buildings*, 171:11–25, July 2018.

[13] M. Waseem Ahmad, M. Mourshed, and Y. Rezgui. Trees vs neurons: Comparison between random forest and ann for high-resolution prediction of building energy consumption. *Energy and Buildings*, 147:77–89, 2017.

[14] J. Yang, H. Rivard, and R. Zmeureanu. On-line building energy prediction using adaptive artificial neural networks. *Energy and Buildings*, 37:1250–1259, 2005.

[15] H. Zhao and F. Magoulès. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, 16:3586–3592, 2012.