Solar Generation Forecast from a Perspective of a DSO

Miguel Fonseca Monteiro

miguel.f.monteiro@ist.utl.pt

Instituto Superior Técnico
University of Lisbon, Portugal
November 2017

Abstract - The integration of renewable energy into the power system is getting more attention during the last years. The need to find an ecologic alternative to fossil fuels is increasing, and if one wants to preserve the planet Earth one should act as fast as possible. There are several ways of assisting the integration of renewables. The use of artificial intelligence (AI) techniques allows the accomplishment of forecasts, what aids the electrical companies to avoid unnecessary penalty payments in the market, for instance. In this work, it will be compared the performance of four AI techniques: artificial neural networks (ANN), adaptive neuro-fuzzy inference systems (ANFIS), random forests (RF) and k-nearest neighbors (kNN) in the task of predicting solar PV power output. The procedure will consist in the implementation of these four methods for three days with different meteorological conditions (a sunny day, a partly cloudy day and a cloudy day), in the RStudio software. After making a one-day-ahead solar PV power forecast, the goal is to conclude about the impact caused by cloudiness and to determine the algorithm with the best performance, including the computation of the mean average percentage error (MAPE) and measuring the simulation time. In the end, kNN clearly outperformed the other techniques for the three tests that were made. A global comparison between all the simulations is made.

Keywords: Solar PV power forecast, artificial intelligence, ANN, ANFIS, random forest, kNN

I. INTRODUCTION

Since the origin of mankind that there is one indispensable and infinite source of energy: the sun! Solar energy was the cornerstone for the primitive societies, providing, directly or not, ways to subsist and to also to evolve. These societies have been developing themselves and have always attempted to build new technologies that enable the profitable use of the solar energy. After the Industrial Revolution, the use of fossil fuels highly increased, in order to answer to the increasing energetic demand. Then, the limited nature of fossil fuels and its environmental impact changed the mentality of the people. It was necessary to replace them for sustainable sources of energy. Renewable energy sources are the most sustainable ones, due to all the advantages they can bring. Besides this, 75% of all the consumed energy still derives from fossil fuels [1]. Although hydro power energy being responsible for the biggest piece of renewable power capacity and generation, solar photovoltaic energy is the one that is growing at faster pace. The main focus of today is to take advantage of solar photovoltaics (PV), in which is being made the biggest investments.

This work intends to perform a comparison between four techniques for a one-day-ahead solar PV power forecast, for three different days with different meteorological conditions.

The structure for the paper is: Sector I – Introduction; Sector II – State of the art of the problem under study; Section III – Reference to the algorithms used; Section IV – Simulation conditions; Section V – Results; Section VI – Conclusions.

II. STATE OF THE ART

The development of efficient ways to integrate renewable energies into the power system is probably the major challenge when referring to the replacement of fossil fuels. Fossil fuels can be forced to generate a certain quantity of power, at any time, in opposite to renewable sources that cannot be adjusted to the actual power demand. The unpredictability and the variability of weather conditions truly constrains the efficiency of renewables. It is believed that, in 2050, 12% and 11% of the global electricity consumption will be provided by wind energy [2] and PV energy [3], respectively. This work is related to solar energy, which is the energy original from the sun and that is assumed as an endless non-polluting source of energy.

The construction of a PV power plant is highly affected by the location and whether the weather conditions are favorable or not. The power output of a PV power plant does not come in steady state. It has lots of fluctuations and it is crucial to deal with them. These fluctuations must be controlled through the adaptation of conventional power plants, in case of incapability of providing the committed power. It is important for power system operators to have an idea about how much power will flow at a certain time in order to schedule their actions and to avoid extra costs and unnecessary penalty payments.

There are several ways to perform a forecast. In this work, two topics were approached, Numerical Weather Prediction (NWP) methods and other forecasting methods that implement mathematical processes into time series. The lack of ground resources able to measure solar radiation in an efficient way is, let us say, the cause for the creation of NWP...
models. Vilhelm Bjerknes exposed one fact of extreme importance and that is considered as the heart of NWP methodology. He revealed that it is imperative to analyze how the atmosphere changes from one initial state to the next one, knowing the intrinsic physical laws under that transition. NWP models use a diverse range of mathematical and meteorological contents, including the understanding of baroclinic instability, the general circulation modelling, spectral and transform methods, the semi-implicit time differenting and the atmospheric initialization. The actual NWP models take advantage of the computational power of today, what result in higher resolutions and the possibility to cover larger domains. NWP models are divided into two, the regional and the global models. The main difference between them is the covered area, where the regional models are specific for a single region and the global models can cover plural regions.

On the other side, there are techniques that are more indicated to the analysis of time series. Starting from regressive methods to artificial intelligence ones, there was a lot of attention to this topic, what led to great improvements. Some initial forecasting processes were simply regressive methods, such as the Moving Averages (MA) [4] and the Auto-regression (AR) [5]. Then, emerged the Auto-Regressive Moving Averages (ARMA) for stationary time series and the Auto-Regressive Integrated Moving Averages (ARIMA), focused on non-stationary processes [6].

The appearance of artificial intelligence techniques was the answer researchers found against non-linearity. AI techniques have been used for a long time and for a wide variety of areas and applications. There are numerous methodologies, each one with its strengths and weaknesses. According to the type of problem to solve, one will try to choose the best method for its resolution. More recently, hybrid systems have arisen. The possibility to combine more than one method, rapidly became one of the many choices for solar forecasting. Adding strengths from different techniques really changed the bounds, as can be seen in Fig. 1, leading to more efficient prediction.

The main topic of this work was to make a comparison of performances of some AI techniques. The four methods that have been chosen were the artificial neural network (ANN), the adaptive neuro-fuzzy inference system (ANFIS), the random forests (RF) and the k-nearest neighbors (kNN). It will not be a full description of them in this report, one decided only to make a short approach of each one.

The ANN consists in a computational model based on the structure and functions of a biological neural network. Its structure is affected according to the inputs and it has a learning capability, which depends on its ways of being built and the information that flows through the network. ANN is recognized as a nonlinear statistical data modelling tool, which purpose is to model complex relationships between inputs and outputs and to find possible patterns. ANFIS is a combination between adaptive neural networks and the fuzzy set theory. It is attempted to approximate to human knowledge by efficiently using linguistic information. The introduction of fuzzy if-then rules [7] into the system led to the modeling of qualitative aspects of human knowledge.

The next two methods enter in the classification field. The random forest uses the idea of growing an ensemble of classification trees in order to perform a weighted voting among them. The creation of independent classifiers is the way to reach a solution, which consists in the global aggregation of the predictors, as referred by Breiman in [8]. Lastly, the kNN method is known for being one of the simplest machine learning algorithm. kNN is a pattern recognition algorithm for classifying patterns or features. It is a nonparametric technique used to estimate a density function from a sample pattern. When referring to weather forecasting, it is known as analog method and has the aim of searching over the time series history in order to find previous timestamps as close as possible to the actual situation.

Each algorithm required a specific way to be implemented and its implementation varied according to the type of test performed, i.e., for the sunny day, the parameters of each technique were not the same as for the cloudy day, for instance. As it was said previously, this work intended to compare the results of these four methods, for three different situations. The entire process of implemented each technique, for each test, will not be included in this report.

IV. SIMULATION CONDITIONS

This section’s purpose is to refer how the data was handled and processed in order to achieve the results. All the work was driven in RStudio, which is an integrated development environment for R, a statistical language used for data analysis. Using some R packages, it was possible to perform a one-day-ahead solar PV power output forecast, for three distinct days.

The dataset1 consisted in measurements from a group of photovoltaic panels and it is constituted by values of PV module temperature, solar irradiance and solar PV power output, per

---

1 The data was kindly provided by EDP-Inovação.
minute, for the first five days of January and the first five of August. From the dataset, one selected three days to perform the forecast. The 4th of January consisted in a sunny day, the 5th of January was a partly cloudy day and the 3rd of August a cloudy day, as can be seen in Fig. 2, Fig. 3 and Fig. 4, respectively.

Therefore, one decided to choose an interval for the prediction, as an attempt to avoid low values of solar irradiance, and so the power, which would difficult the forecasting process. Saying this, for the days of January it was chosen the interval between 9:00 and 16:00, while, for August, one worked with the time interval between 10:00 and 17:00, to realize the prediction.

![Figure 2. Time evolution of the irradiance for the 4th of January](image1)

![Figure 3. Time evolution of the irradiance for the 5th of January](image2)

![Figure 4. Time evolution of the irradiance for the 3rd of August](image3)

The data was splitted into training set and testing set. The testing set had the values of all the variables for the interval specified for each day, whereas all the other values belonged to the training set.

Each method required a specific package in RStudio for its implementation. The detailed specification of the parameters used to perform each experiment is included in the thesis report. Therefore, four methods have been applied for a one-day-ahead solar PV power output forecast for three days with different meteorological conditions. Let us explore the results next section. One aspect that must be referred is that, for neural network and for random forest, one decided to create a cycle in order to obtain more precise results, since both of the methods originate different forecasts for simulations with the same parameters.

V. RESULTS

After implementing all the algorithms, the results became available and a comparison between them can be made. In this section, one chose to divide it into four main parts. These sectors are regarded to the sunny day results, the partly cloudy day results, the cloudy day results and the results for another test, where one decide to change the time scale of the process.

In the thesis report, it was illustrated, for each experiment, the time evolution of the prediction power and the real power in a single graphic and the time evolution of the error of each minute, which is computed by

\[
error = \left| \frac{p'_i - p_i}{p_i} \right| \times 100\%
\]

where \(p'_i\) represents the predicted power and \(p_i\) is the real power.

Since one computed the error according to (1), the forecast quality was assessed calculating the mean average percentage error (MAPE), given by

\[
MAPE = \frac{100\%}{N} \sum_{i=1}^{N} \left| \frac{\hat{p}_i - p_i}{p_i} \right|
\]

where \(\hat{p}_i\) is the predicted value of the power, \(p_i\) the real value and \(N\) is the number of samples.

For each day to be predicted, four simulations were run. Therefore, in the end the total number of graphics obtained was quite high. For this reason, in this paper only a few ones will be exposed, the simulations with the best and with the worst results, for each day. However, some tables will be included to show the achieved results. All the techniques were also compared with the persistence.

A. Sunny Day

The 4th of January was a sunny day and the goal was to make a prediction for the underlying day, between 9:00 and 16:00. For this test, one chose to illustrate the time evolutions of the solar PV power output for the kNN and the random forest, the methods with the best and the worst results, respectively.
These two graphics enable to notice the proximity between the real values of the power and the forecast. There are some deviations in the random forest predictions, but, in general, all the methods achieved good results and the time evolution had a similar shape as the previous one. This fact can be seen in Table 1.

<table>
<thead>
<tr>
<th>Models</th>
<th>MAPE (%)</th>
<th>Simulation Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistence</td>
<td>0.678</td>
<td>-----</td>
</tr>
<tr>
<td>Neural Network</td>
<td>1.993</td>
<td>981.24</td>
</tr>
<tr>
<td>ANFIS</td>
<td>1.707</td>
<td>60.64</td>
</tr>
<tr>
<td>Random Forest</td>
<td>2.158</td>
<td>444.96</td>
</tr>
<tr>
<td>kNN</td>
<td>0.677</td>
<td>0.03</td>
</tr>
</tbody>
</table>

The MAPE for the persistence was 0.678%, which is a very low value. This shows the low variability of the sunny day, practically with no clouds. kNN performed better than the persistence and all the other methods had a MAPE around 2%. kNN was also the fastest method, while neural network was the most time-consuming.

B. Partly Cloudy Day

The day with this characteristic was the 5th of January. In agreement to what was said, the chosen time period was the same as the previous test. Like the sunny day, for the partly cloudy day, the best and the worst methods were the kNN and the random forest, respectively.

For the second test, the partly cloudy day, it was noticed an increase in the variability, caused by cloud cover in the first hours of the day. This variability highly conditioned the forecasting quality, leading to greater errors for all the algorithms. The time evolution of the other two methods revealed to be similar to the one illustrated in Fig. 8. After 11:30, approximately, it behaves as a sunny day and so the forecast is more similar to the real values. Let us take a look in the results, in Table 2.

<table>
<thead>
<tr>
<th>Models</th>
<th>MAPE (%)</th>
<th>Simulation Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistence</td>
<td>4.523</td>
<td>-----</td>
</tr>
<tr>
<td>Neural Network</td>
<td>5.238</td>
<td>564.41</td>
</tr>
<tr>
<td>ANFIS</td>
<td>5.057</td>
<td>23.25</td>
</tr>
<tr>
<td>Random Forest</td>
<td>5.312</td>
<td>453.25</td>
</tr>
<tr>
<td>kNN</td>
<td>3.688</td>
<td>0.03</td>
</tr>
</tbody>
</table>

The MAPE value of 4.523% for the persistence indicates the existence of cloud cover. The biggest errors derived from the first half of the day, where there are more relevant oscillations of the power. Besides kNN, which outperformed the persistence, none of the other techniques was able to obtain a MAPE lower than 5%. As expected, the simulation time of the neural network and the random forest was significantly higher. ANFIS and kNN were the fastest and also achieved better values.
C. Cloudy Day

The 3rd of August revealed to be a cloudy day and one chose to make the prediction for the time interval from 10:00 to 17:00. The comparisons that are going to be illustrated are the one for the kNN and for the neural network, the methods with the best and the worst performances, respectively.

The cloudy day revealed a huge increase in the cloudiness. There are numerous peaks in the time evolution of the real power, and of course, in the time evolution of the forecast. For this test, the high variability led to biggest deviations in relation to the real values of the power. There is a more relevant discrepancy between predicted and real values, which derived from abrupt changes of the power. For instance, one noticed that for close inputs, i.e. similar values of the temperature and the irradiance, the output power had quite different values. This unpredictability is function of cloud cover and it significantly restrained the accuracy of a forecast. Let us look to Table 3, where it was made a global comparison of the results.

D. Time Scale Change

This last test was led in order to prove the importance carried by the chosen time scale. The main experiments worked with a time scale of minutes, what really influenced the accuracy of the tests. This happened because of the proximity of the measurements, i.e. the variations per minute could be very small. Therefore, one agreed to convert the data into hourly averages. In this step, it was performed a mean of some of the measurements so that one could handle the data as hourly values. Applying it to the sunny day, one had the hourly values for each day of January, between 7:00 and 17:00. Following the procedure for the sunny day, it was made a forecast for the interval between 9:00 and 16:00. One applied it only for the neural network and the kNN. The time evolutions for these techniques are now illustrated in Fig. 11 and Fig. 12.

The cloudy day had a MAPE of 11.2%, what represents a significant rise compared to the previous tests. This fact proves that the appearance of clouds really compromise the results. In this case, there was an even higher variability. The existence of several peaks in the time evolution led to huge errors, around 100%. Even more for this case, kNN stood out as the algorithm with the best performance, with a MAPE of only 4.2% and an extremely reduced simulation time. ANFIS was able to reach a MAPE lower than the persistence and ran in less than two minutes. Then, neural network and random forest obtained a MAPE around 14% and took more time to perform the simulation, especially neural network, which demanded a tremendously high simulation time.

<table>
<thead>
<tr>
<th>Models</th>
<th>MAPE (%)</th>
<th>Simulation Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistence</td>
<td>11.196</td>
<td>-----</td>
</tr>
<tr>
<td>Neural Network</td>
<td>13.956</td>
<td>12650.18</td>
</tr>
<tr>
<td>ANFIS</td>
<td>10.558</td>
<td>96.19</td>
</tr>
<tr>
<td>Random Forest</td>
<td>13.631</td>
<td>581.12</td>
</tr>
<tr>
<td>kNN</td>
<td>4.182</td>
<td>0.03</td>
</tr>
</tbody>
</table>

The cloudy day had a MAPE of 11.2%, what represents a significant rise compared to the previous tests. This fact proves that the appearance of clouds really compromise the results. In this case, there was an even higher variability. The existence of several peaks in the time evolution led to huge errors, around 100%. Even more for this case, kNN stood out as the algorithm with the best performance, with a MAPE of only 4.2% and an extremely reduced simulation time. ANFIS was able to reach a MAPE lower than the persistence and ran in less than two minutes. Then, neural network and random forest obtained a MAPE around 14% and took more time to perform the simulation, especially neural network, which demanded a tremendously high simulation time.
As can be seen, there is a strong correlation between the forecast of both methods and the real values, in opposite to the persistence, which reveals to be very displaced. In order to show the difference, let us have a look in Table 4.

Table 4. MAPE after the time scale change

<table>
<thead>
<tr>
<th></th>
<th>MAPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistence</td>
<td>49.81</td>
</tr>
<tr>
<td>Neural Network</td>
<td>1.33</td>
</tr>
<tr>
<td>kNN</td>
<td>1.41</td>
</tr>
</tbody>
</table>

The MAPE of the persistence reached a value of 49.8%, whereas for the AI techniques, it was close to 1.5%. This fact confirmed the importance of using the persistence as a benchmark only for forecasts with small time scale, otherwise it reveals to be practically useless.

VI. CONCLUSIONS

From a DSO perspective, it is essential to have accurate forecasts in order to schedule important actions in the electricity market. So, there is a strong relation between forecasting and the operation of the power grid. This is a central part of the process of integrating renewable energies into the power grid and an attempt to take advantage from what they can provide. Therefore, the objective of this thesis was to make a global comparison between four artificial intelligence techniques, in the matter of a one-day-ahead PV power output forecast. Three main tests were performed, for three days with different conditions. The different weather conditions were mainly because of cloud cover, which brings variability for the forecasting process.

For the sunny day, the forecasts were good, showing small deviations from the real power. The MAPE for the implemented techniques was around 2%, except for kNN that obtained a MAPE of 0.68%, similarly to the MAPE of the persistence. Overall, it were achieved good values.

During the partly cloudy day, there was two main parts. The first half of the day had some cloudiness and so the results were affected by that fact. The rest of the day behaved as a sunny day and therefore the predictions were reasonably in agreement with the real power. There was an increase in the MAPE, due to the variability introduced by cloud cover. At the end, kNN reached the best values, with a MAPE of 3.67%, better than the persistence’s (4.52%). The other methods had similar values with MAPE values slightly above 5%.

The forecasts for the cloudy day significantly suffered the impact of the cloudiness. For this experiment, there was a huge variability of the PV power due to the big variations of the solar irradiance during the day. The MAPE of the persistence ascended to a value of 11.2%, what proves those variations. One more, kNN performed better, with a MAPE of 4.2%, meaningfully lower than the persistence. ANFIS obtained a MAPE of 10.6%, slightly lower than the persistence one. Neural network and random forest had a MAPE close the 14% and were the two most time-consuming techniques.

In the end, one noticed the increase in the struggle of performing a reliable forecast. Deal with cloud cover requires extra efforts among forecasting communities. The comparison was made and the kNN revealed to be the algorithm that achieved the best results. In order to improve the overall performance, one would recommend the introduction more meteorological variables, variables correlated with cloudiness, if possible such as air pressure, air density, among others. Another action that could lead to better results is possibility of hybridization through the combination of AI techniques with other forecasting methodologies.

REFERENCES