Machine learning approach for next day energy production forecasting in grid connected photovoltaic plants

L.Mora-López[1,*], I.Martínez-Marchena[1], M.Piliougine[2], M.Sidrach-deCardona[2]

[1] *Dpto. Lenguajes y Ciencias de la Comunicación. Universidad de Málaga. Málaga, Spain*
[2] *Dpto. Física Aplicada II. Universidad de Málaga, Málaga, Spain*
** Corresponding author. Tel: +34 952250362, Fax: +34 952131397, E-mail: llanos@uma.es*

Abstract: This paper presents a model for predicting the next-day energy production of a photovoltaic solar plant. The model is capable of forecasting the next-day production profile of such a system, merely by using the information obtained from the plant itself and the solar global radiation values for the previous operation days. This prediction is key in many photovoltaic systems in order to interact with conventional electrical grids. For example, Spanish legislation requires this type of information for large photovoltaic plants. In fact, the deviations from the predicted values are financially penalized. A three-stage procedure is used to build the model, which is capable of learning specific information about each facility and of using this information to fit the prediction. This model binds the use of regression techniques and the use of a special type of probabilistic finite automata developed from machine learning. The energy prediction yearly error is less that 20 percent which is a significant improvement over previous proposed models, whose errors are around 25 percent.

Keywords: *short term forecasting, photovoltaic energy production, machine learning*

1. Introduction

Process forecasting has become a key tool in many areas, such as competitive electricity or economic markets. In the short term, forecasting the expected values of certain variables can be an important tool for optimal systems management and to decide on the best operation strategies. Forecasting energy production by large plants has thus become a requirement in competitive electricity markets. In the short term, expected produced energy can help producers to achieve optimal management and can also help to implement efficient operation strategies based on the best way of interacting with conventional grid. For example, since 1998, the Spanish electricity market has moved from a centralized operational approach to a competitive on. It encourages the deployment of solar plants with a financial penalty for incorrect prediction of solar yields for the next day. In this global market, energy generated by these systems for the grid needs to be predicted as accurately as possible in order to ensure that solar energy systems are truly penetrated in the electricity market. Forecast regarding energy production is necessary to manage and schedule electricity grids. This prediction will facilitate the use of these systems as distributed generators in grid connected photovoltaic systems.

Estimating the energy generated by solar plants is difficult mainly due to its dependence on meteorological variables, such as solar radiation and temperature. In fact, photovoltaic production prediction is mainly based on global solar irradiation forecasts. The behavior of this variable can change quite dramatically on different days, even on consecutive days. This is because global solar radiation is not a deterministic variable due to the climatic conditions. Although the extraterrestrial solar irradiation -defined as the solar irradiation that reaches the extra atmospheric zones of the earth- is deterministic, once this irradiation penetrates in the atmosphere, different variable phenomena come into play and only a fraction of the extraterrestrial solar irradiation therefore reaches the surface of the earth. This fraction is known as solar global radiation. These phenomena include the presence of clouds in the atmosphere that can significantly reduce the solar irradiation reaching the earth. Accurately

forecasting the energy generated by these systems is difficult as solar radiation is the energy source of solar systems.

In general, a wide range of statistical and artificial intelligence techniques have been developed for process forecasting. Statistical time series methods are based on the assumption that the data have an internal structure that can be identify by using simple and partial autocorrelation, [1], [2], [3], [4]. Time series forecasting methods detect and explore such a structure. In particular, ARMA (autoregressive moving average), ARIMA (autoregressive integrated moving average) models have been widely used. Artificial intelligence techniques and, in general, machine learning models have been also used for process forecasting, [5], [6], [7], [8], [9]. Different approaches have likewise been specifically developed for forecasting global solar irradiation, [10], [11], [12], [13], [14].

We propose a model that is capable of learning the important facts in the prediction of photovoltaic plants energy production. A new approach based on the use of probabilistic finite automata and multivariate regression analysis is proposed here for short-term forecasting of the production of solar plants. The forecasting model is built in three stages and has been previously used for short-term forecasting of hourly global solar radiation, [15], [16]. The first and second stages of the procedure are used to identify and capture the significant information for predicting the production of a photovoltaic plant and to build the model using this information. In the first stage, the most significant independent variables are selected by using a multivariate regression analysis. In the second stage, probabilistic finite automata are built using the significant variables obtained in the first stage. The next values of the dependent variable are predicted using an algorithm for short term forecasting which is based on the information stored in the built model. In the third stage, the next-day solar energy production forecasting is calculated using the estimates values in the second step and the parameters of each solar photovoltaic plant. The methodology and the proposed model are described in the second section. In the third section, the results obtained when the model is used for next-day energy production forecasting in photovoltaic plants are presented. The conclusions of the paper are presented in the last section.

2. Methodology

This paper seeks to propose a model for forecasting next-day energy production in grid connected photovoltaic plants. The model is based on the model developed for short-term forecasting of hourly global solar radiation described in [15]. We propose the use of several independent variables to build the model; these variables are usually available in large photovoltaic solar plants: irradiation values and temperature. Moreover, specific parameters of the plant, such as power installed, orientation and tilt of the panel arrays, have been included in the final model. The model is built in three stages.

In the first stage, statistical techniques are used to determine the most significant information among the independent variables used. Using this information, the data are divided into different groups and for each group the new significant variables are determined. In the second stage, a special type of probabilistic finite automata is built for each group taking into account the significant variables of the group. In the third stage, the model of prediction is used for forecasting the energy produced by the photovoltaic solar plants for the next day.

The mathematical model proposed to store the information contained in solar irradiance is based on the use of a special type of probabilistic finite automata (PFA). The use of this mathematical model is envisaged to select both the most meaningful information included in a

stationary continuous time series and the information obtained from other sources. A detailed description of this model can be found in [15].
The power generated at the output of the inverter, P_{AC} can be estimated using the expression:

$$P_{AC} = \eta_{inv} * P_m^{STC} * \frac{G_\beta}{1000} * (1 + \gamma * (T_{\mathrm{mod}} - 25)) \quad (1)$$

where, η_{inv} is the efficiency of the inverter, $P_m{}^{STC}$ is the power generated by the photovoltaic generator in standard conditions of radiation and temperature (1000W/m^2, 25ºC), $G_{\beta t}$ is the global irradiance on the surface of the modules (W/m^2) – β is the inclination of the modules, γ is the temperature coefficient of P_m, and $T_{mod,t}$ is the module temperature. In the case of monocrystalline silicon, the value of the coefficient γ is 0.48%/ºC (these type of modules are used in all the facilities analyzed).

The irradiance on the surface of the modules is the most difficult parameter to estimate using Eq.(1). Moreover, this parameter presents a seasonal trend due to the changes in the relative sun-earth position. Using the values of clearness index is proposed to remove this seasonal trend. This parameter is estimated using the following expression:

$$k_t = \frac{G_t}{G_{0,t}} \quad (2)$$

where G_t is the global irradiance (Wh/m2) at time t and $G_{0,t}$ is the extraterrestrial solar irradiance at time t (Wh/m2); the expression for estimating $G_{0,t}$ can be found in [17].

2.1. First stage

In the first stage, the following linear regression model is estimated using ordinary least squares :

$$k_{t,d} = \beta_0 + \beta_1 k_{t,d-1} + \beta_2 k_{t,d-2} + \beta_3 k_{t,d-3} + \beta_4 S_{1,t,d} + \beta_5 S_{2,t,d} + \beta_6 ST_{3,t,d} + Error \quad (3)$$

where t means time, d means day and S_i , $i=1,2,3$, are three dummy variables to represent the season to which the observation $k_{t,d}$ belong (only three dummy variables are used to avoid multicollinearity problems). Among these independent variables, the most significant variable for predicting the next value of clearness index is used for splitting the observation into G groups. For each one of these groups, the Eq.(3) is again estimated to determine the significant variables of the group.

2.2. Second stage

For the observations of each group, a special type of probabilistic finite automata is built using the significant variables of the group. The continuous variables need to be discretized to use this model. A static discretization method has been used. The range of each continuous variable has been divided into q equals intervals. Several values of q have been proved for each group in order to select the best discretization, taking into account the performance of the probabilistic finite automata in the short-term forecasting of clearness index. The proportional mean prediction error (*PMPE*) has therefore been estimated, i.e.

$$PMPE = \sum_{t=1}^{N} \frac{| k_{t,d} - k_{t,d}^{*} |}{k_{t,d}} \quad (4)$$

where N are the number of observations in each group and $k_{t,d}^{*}$ is the predicted value of clearness index.

2.3. Third stage

In the third stage, the values of solar irradiance G_t are estimated from the values of clearness index predicted using the PFAs built in the second stage. With these values, the power generated at the output of the inverter is estimated using the Eq. (1). For evaluating the model, the mean prediction error for these values has been estimated, i.e.:

$$MPE = \frac{\sum_{t=1}^{N} | P_{AC,t} - P_{AC,t}^{*} |}{\sum_{t=1}^{N} P_{AC,t}} \quad (5)$$

3. Data

The data used have been recorded from four photovoltaic plants installed in different Spanish locations. The data used for these facilities are the following: power generated at the inverter output, irradiance on the surface of the modules and modules temperature. Moreover, the season to which each observation belongs has been included. Table 1 sets out a summary of the characteristics of each facility.

Table 1. Description of the data used.

Location	Latitude/Longitude	Peak power (kW)	Inclination of modules	Data
Location 1	43.30/-1.95	14.08 kW	20	01/10/2009-10/12/2010
Location 2	43.18/3.00	13.86 kW	20	01/10/2009-10/12/2010
Location 3	43.37/-1.85	20.16 kW	30	01/11/2009-10/12/2010
Location 4	43.37/-1.85	20.16 kW	30	01/11/2009-10/12/2010

4. Results

In the first stage, the linear regression model, Eq. (3), has been estimated using the ordinary least square (OLS) for the data of each location. In all cases, the most significant variable proves to be $k_{t,d-1}$, that is the clearness index for the same hour at the previous day (significance level=0.05).

Using this variable, the observations of each location have been split into 5 different groups depending on the value of this variable. The model, Eq. (3), has been estimated by OLS for each group.

Table 2 summarizes the significant variables for each group, taking into account the values of the t-statistic for a significance level of 0.05, for Location 3. As can be observed, these variables differ depending on the group. This result is similar for all locations.

Table 2. Significant variables for each group of observations (Eq.1, significance level=0.05) for Location 3

Interval	Significant variables
[0.0-0.2[	$K_{t,d-2}$, $K_{t,d3}$, $S_{3,t,d}$
[0.2-0.4[	$K_{t,d-2}$, $K_{t,d-1}$, $S_{3,t,d}$
[0.4-0.6[	$K_{t,d-2}$, $S_{3,t,d}$
[0.6-0.8[	$K_{t,d-2}$, $S_{1,t,d}$
[0.8-1.0]	$K_{t,d-1}$, $K_{t,d-2}$, $S_{1,t,d}$

A probabilistic finite automata (PFA) has been built for each location and group of observations using the significant variables and the procedure described in [15]. Using these PFAs, the values of clearness index have been estimated. The values of irradiance at the surface of the modules are also obtained using these estimates and the Eq.(2) . Finally, Eq.(1) is used to calculate the power generated at the inverter output for each instant and the daily profiles are also obtained. The mean prediction error has been estimated using Eq.(5) for the power generated at the output of the inverter for each location. These values are reported in Table 3.

Table 3. Mean prediction error of the proposed model.

Location 1	Location 2	Location 3	Location 4
0.18	0.14	0.17	0.16

5. Conclusions

We have developed a model to predict the energy that a photovoltaic solar plant will produce for the next day. This model only uses the information obtained in the own plant and the values of solar global radiation for the previous operation days. A three stage procedure was used to build the model. The model is estimated using the data from each facility and is capable of learning specific information about each facility and of using this information to fit the predictions.

This model binds the use of regression techniques and the use of a special type of probabilistic finite automata developed from machine learning. The mean prediction error of the energy predictions is less that 20 percent which is a significant improvement over previous proposed models, whose errors are about 25 percent.

Further research would lead to further information that is usually available at large grid-connected photovoltaic plants being included in the model

References

[1] G.E.P. Box, G,M. Jenkins, Time Series Analysis forecasting and control. USA. Prentice Hall, 1976.

[2] J.G. Gooijer, R.J. Hyndman, 25 Years of IIF Time Series Forecasting: A Selective Review," Tinbergen Institute Discussion Papers 05-068/4, Tinbergen Institute, 2005.

[3] P.J. Brockwell, A.D. Richard, Introduction to Time Series and Forecasting, Springer Texts in Statistics, 2002.

[4] J. Hwang, S.M. Chen, C.H. Lee, Handling forecasting problems using fuzzy time series. Fuzzy sets and Systems, 100, 1998.

[5] J.J. Guo, P.B. Luh, Selecting input factors for clusters of Gaussian radial basis function networks to improve market clearing price prediction, IEE Trans Power Syst, 18 (2), 2003, pp. 665-672.

[6] C.H.Wang, L.C. Hsu,. Constructing and applying an improved fuzzy time series model: Taking the tourism industry for example. Expert Systems with Applications 34, 2008.

[7] Q. Song, B.S. Chisson, Forecasting enrollments with fuzzy time series. Part I. Fuzzy Sets and Systems, 54, 1993.

[8] Q. Song, B.S. Chisson, Forecasting enrollments with fuzzy time series. Part II. Fuzzy Sets and Systems, 54, 1993.

[9] J. Hwang, S.M. Chen, C.H. Lee, Handling forecasting problems using fuzzy time series. Fuzzy sets and Systems, 100, 1998.

[10] R. Perez, K. Moore, P. Stackhouse, Forecasting solar radiation Preliminary evaluation of an approach based upon the national forecast database, Solar Energy, vol. 81, no. 6, 2007, pp. 809-812.

[11] L. Mora-Lopez, J. Mora, M. Sidrach-de-Cardona, R. Morales-Bueno, Modelling time series of climatic parameters with probabilistic finite automata. Environmental modelling and software, 20(6), 2005, pp. 753-760.

[12] B. Viorel (Ed). Modeling Solar Radiation at the Earths Surface. Recent Advances. Springer, 2008.

[13] R.A. Guarnieri, E.B. Pereira, S.C. Chou, 2006, Solar radiation forecast using articial neural networks in South Brazil, in Proc. 8 ICSHMO, Foz do Iguau, Brazil, Apr. 2428, 2006, 17771785, INPE.

[14] D. Heinemann, E. Lorenz, M. Girodo, Forecasting of solar radiation, Solar Energy Resource Management for Electricity Generation From Local Level to Global Scale, Hauppauge, NY: Nova, 2005.

[15] L. Mora-Lopez, J. Mora, M. Piliougine, M. Sidrach-de-Cardona. "An Intelligent Memory Model for Short-Term Prediction: An Application to Global Solar Radiation Data". LNAI 6098, 2010, pp. 596–605.

[16] L. Mora-López, M. Piliougine, J.E. Carretero, M. Sidrach-de-Cardona. Integration of Statistical and Machine Learning Models for Short-term Forecasting of the Atmospheric Clearness Index. International Congress on Environmental Modelling and Software Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada, Julio, 2010.

[17] M. Iqbal,An introduction to solar radiation. Academic Press Inc. New York – London, 1984.