Scientific Paper / Artículo Científico

UNIVERSIDAD POLITÉCNICA SALESIANA ECUADOR
INGENIUS
Revista de Ciencia y Tecnología
https://doi.org/10.17163/ings.n30.2023.02
pISSN: 1390-650X / eISSN: 1390-860X

Methodology based on data science for the development of a forecast of the ower generation of a photovoltaic solar plant

Metodología basada en ciencia de datos para el desarrollo de pronóstico de la generación de energía de una planta solar fotovoltaica

César A. Yajure-Ramírez[1,*]

Received: 03-03-2023, Received after review: 21-04-2023, Accepted: 26-04-2023, Published: 01-07-2023

Abstract

The use of photovoltaic solar plants for the generation of electrical energy has been constantly increasing in recent years, and many of these plants are connected to the external electrical network, which makes it necessary to forecast the electrical energy generated by the solar plants to assist in the management of the network operator. This research presents a methodology based on data science to develop the forecast of electrical energy generated from photovoltaic solar plants, using three different techniques for comparison purposes: time series analysis, multiple linear regression, and artificial neural network. Historical data of peak power, solar irradiance, ambient temperature, wind speed, and soiling rate from an experimental NREL photovoltaic solar plant were used. To evaluate the performance of the models, the RMSE, MAE, and MAPE metrics are used, resulting in the ARIMA model of the time series analysis having the best performance with a MAE of 1.38 kWh, RMSE of 1.40 kWh, and MAPE of 6.35%. In the correlation analysis, it was determined that power generation was independent of the soiling rate, so this variable was discarded in the regression models.

Keywords: Machine learning, solar irradiance, artificial neural network, linear regression, time series, ambient temperature.

Resumen

El uso de plantas solares fotovoltaicas para la generación de energía eléctrica ha ido en constante aumento en los últimos años. Muchas de estas se conectan a la red eléctrica externa, por lo que se hace necesario el pronóstico de la energía eléctrica generada por las plantas solares para coadyuvar en la gestión del operador de la red. En esta investigación se presenta una metodología basada en la ciencia de datos para desarrollar el pronóstico de energía eléctrica generada de plantas solares fotovoltaicas, utilizando, para efectos de comparación, tres técnicas diferentes: análisis de series de tiempo, regresión lineal múltiple, y red neuronal artificial. Se trabajó con los datos históricos de la potencia pico, la irradiancia solar, la temperatura ambiente, la velocidad del viento, y la tasa de suciedad, de una planta solar fotovoltaica experimental del NREL. Para evaluar el desempeño de los modelos se utilizan las métricas RMSE, MAE, y MAPE, resultando que el modelo ARIMA del análisis de series de tiempo fue el que mejor desempeño tuvo con un MAE de 1.38 kWh, RMSE de 1.40 kWh, y MAPE de 6.35%. En el análisis de correlación se determinó que la generación de energía era independiente de la tasa de suciedad, por lo que se descartó esta variable en los modelos de regresión.

Palabras clave: aprendizaje automático, irradiancia solar, red neuronal artificial, regresión lineal, serie de tiempo, temperatura ambiente

[1,*]Posgrado en Investigación de Operaciones, Universidad Central de Venezuela, Venezuela.
Corresponding author ✉: cyajure@gmail.com.

Suggested citation: Yajure-Ramírez, C. A. "Methodology based on data science for the development of a forecast of the ower generation of a photovoltaic solar plant," *Ingenius, Revista de Ciencia y Tecnología*, N.° 30, pp. 19-28, 2023, DOI: https://doi.org/10.17163/ings.n30.2023.02.

1. Introduction

The use of renewable energy sources for electricity production has increased in recent years due to public policies in some countries aimed at reducing environmental pollution caused by fossil fuel sources and bringing electricity to remote places where the traditional power grid does not reach. According to the 2022 Global Renewable Energy Status Report, in 2011, 20.4% of electricity came from renewable sources, mainly hydro, solar, wind, bioenergy, and geothermal. In 2021 this percentage increased to 28.3% (15% hydro, 10% solar and wind, and 3% bioenergy and geothermal). As for solar photovoltaic energy, in 2021, there were 942 GW of installed capacity for electricity generation worldwide, showing an increase of 23% compared to 2020 [1].

The use of solar energy for electricity production has been evolving technologically, so the use of solar photovoltaic plants connected to the external power grid has been increasing, reporting an increase of 20% worldwide by 2021 [1]. The energy coming from solar photovoltaic plants is subject to climatic variations, specifically solar irradiance, and temperature. To contribute to the stability and reliability of the electrical system, it is necessary to develop forecasts of the energy generated considering the historical data of these climatic variables. This forecast also contributes to improving the management of the operation and maintenance of these solar photovoltaic plants.

Therefore, this research aims to present a methodology based on data science to develop the forecast of electric power generation from solar photovoltaic plants and to present a comparative study of three different techniques to obtain the forecast models: ARIMA (Autoregressive Integrated Moving Average) model of time series analysis, multiple linear regression, and artificial neural network. For the evaluation of the models, the metrics mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), and the coefficient of determination R^2 were used.

Several research studies on the proposed objectives were reviewed, and various publications were found. Mittal *et al.* [2] reviewed the use of machine learning for photovoltaic power forecasting, reaffirming that solar irradiance and temperature are essential for this forecast. They concluded that hybrid models are the best option for better predicting solar photovoltaic energy.

Sharkawy *et al.* [3] developed a study using a neural network to create a short-term solar plant power forecasting model. They considered five days of data to train the model and the remaining day of data to evaluate the model. The input variables were temperature and radiation. They concluded that the model obtained is adequate since, in training, the RMSE was 0.187 MWh, and in the forecasting phase, the absolute error was 0.08 MWh. Kasagani y Manickam [4] conducted a daily power forecasting study using artificial neural networks and the historical data of power, operating hours, daily global solar radiation, and ambient temperature of the solar photovoltaic plant. As a performance metric, they used the relative RMSE. They concluded that the forecast using an artificial neural network with three neurons in the hidden layer was the best performing, with a MAPE of 4.18% and a relative RMSE of 5.74%. Pattanaik *et al.* [5]performed a comparative analysis of different methods for power forecasting of a solar photovoltaic plant. They concluded that forecasting using genetic algorithms is more convenient and accurate than statistical analysis.

Akhter *et al.* [6]reviewed the methods for forecasting electric power generated by solar photovoltaic plants based on machine learning and metaheuristic techniques. They showed the advantages and disadvantages of each method and compared heuristic methods with machine learning methods. They concluded that hybrid techniques (composed of at least two methods) are the most accurate for all forecast horizons, with a reduction of about 15% in MAPE and RMSE. Alaraj *et al.* [7] developed a decision tree ensemble-based model for power forecasting of a solar photovoltaic plant, using meteorological data from Qassim in Saudi Arabia and comparing their results with other models. They considered the metrics RMSE, MAE, MAPE, and training time to evaluate the model. They concluded that the ENBG model is the best-performing model, with an MAE of 8.89 W in the training phase and 12.05 W in the test phase.

Anuradha *et al.* [8] analyzed the power forecasting of a solar photovoltaic plant by applying different machine-learning techniques and using historical data of climatic variables and generated power. The techniques used were support vector machine, random forests, and linear regression. They concluded that the random forest regression model was the most accurate in its results, with 94.01%. Borunda *et al.* [9] presented a fast methodology to evaluate the best location of a solar photovoltaic plant and to forecast the electric power it will generate, using historical data of climatic variables and machine learning algorithms. They validated the methodology by comparing it with real solar photovoltaic plants in Mexico.

This research consists of several sections; Section 2 explains the methodology and data used in the study; Section 3 shows the results obtained; and Section 4 shows the research conclusions. The bibliographic references used are also shown.

2. Materials and methods

The methodology consists of applying the stages of a data science project, applying its methodology in each stage. According to VanderPlas [10], data science is an interdisciplinary area that includes, in turn, three different areas: statistical skills to model and summarize data; computer skills to design and use algorithms to store, process and visualize these data efficiently; and expertise in the specific field or business of research, which in this work, is the generation of electric power from solar photovoltaic plants.

Cielen's work [11] presents the stages of a data science project. The first stage consists of establishing the objectives to be achieved, which requires knowledge of the field, i.e., generation from solar photovoltaic plants and meeting needs. The second stage consists of obtaining the data of interest, which, in this case, correspond to the regular measurements of the variables in a solar photovoltaic plant from its data acquisition system. The variables required to form the data set depend on the project objective(s). Once the dataset is available, the next step is to process the data, which consists of reviewing, cleaning, transforming, and combining these data to have the appropriate structure. Then, exploratory data analysis is performed using statistical and graphical techniques that can be univariate, bivariate, or multivariate. At this stage, knowledge of interest for the study may already be found, so some projects only reach this stage. If the knowledge from the previous step is insufficient, or if the idea is to move on, the data modeling stage is implemented, which consists of applying mathematical algorithms to obtain models that deepen the knowledge acquired. The number and type of algorithms depend on the objectives set in the first stage. Finally, the decision-making stage is reached, considering the results obtained.

One might think that the stages of a data science project are applied sequentially; however, there may be cases where this does not occur. Depending on the results obtained in the exploratory analysis stage and/or the modeling stage, it may be necessary to return to the data processing stage to improve the structure of the data, to the data collection stage to obtain some other variable, or even to the first stage to reformulate the project objectives.

The methodology is applied to data from a particular solar photovoltaic plant. This section presents the data collection and data processing stages. The exploratory data analysis and modeling stages are shown in the next section.

2.1. Data collection

The data used in this research come from the data acquisition system of a solar photovoltaic plant at the U.S. National Renewable Energy Laboratory (NREL) in Golden, Colorado. They were collected from the public dataset web page of the NREL data acquisition system [12].

The plant comprises five Sanyo mono-silicon solar panels with 200 watts of peak power each [13]. These panels are installed in a fixed mounting, with a 40° tilt angle and an azimuth angle 180°. The data correspond to measurements taken and stored every minute of the plant's peak power output ("ac_power") in watts, ambient temperature ("ambient_temp") in degrees Celsius, irradiance ("poa_irradiance") in watts per square meter, wind speed in meters per second ("wind_speed"), and soiling rate ("soiling"). Data collection began on February 25, 2010, and ended on December 13, 2016, with 1.558.875 rows (records or instances).

2.2. Data processing

Data processing techniques are applied at this stage, such as detecting possible missing data or duplicate rows, outlier detection, data transformation, data column combination, and verifying the appropriate format for the different variables. The way to apply these and other techniques using the Python programming language is described in [14].

After an initial review, seven missing data were detected in the ambient temperature variable, and 17 362 missing data in the wind speed variable. The rows with missing temperature data correspond to less than 0.001% of the total rows, and the rows with missing wind speed data correspond to approximately 1.11% of the total rows. Although these percentages are low, it was decided to attribute them to the mean value of the three data closest to the missing data. It is worth mentioning that no duplicate rows were detected.

Considering the data in the original date column, columns corresponding to the data reading's year, month, week, day, and hour were created. Likewise, considering the column of peak electric power, the column of generated electric energy ("ac_energy") in kilowatt hours (kWh) was created, which is used as the target variable in the forecast models. As for the ambient temperature, it was rescaled from Celsius to Kelvin units.

It was detected that there were no records for January, September, and October 2010, which could alter the results of the exploratory analysis. Consequently, this analysis was performed with existing data between 2011 and 2016.

3. Analysis and results

This section presents the stages of exploratory analysis, data modeling, and the discussion of results.

3.1. Exploratory data analysis

After processing the data in the previous stage, 1,429,678 rows or records were obtained corresponding to the minute values of the measurements of the variables and the other variables that were generated. Data sets with daily, weekly, monthly, and annual resolutions were obtained for analysis. This was achieved by grouping the original data with minute resolution in the corresponding period.

Table 1. shows the results of the descriptive analysis of the daily data using univariate statistics. It can be observed that, except for the (soiling) rate, the mean value of the variables is close to the median value. The range of the solar irradiance and wind speed variables is high, with the mean value closer to the minimum value than the maximum value.

Table 1. Descriptive statistical summary of the data.

Statistics	Variables				
	ac_energy	ambient_temp	poa_irradiance	soiling	wind_speed
Mean	4.19	286.96	465.02	95.87	1.76
Standard deviation	1.66	9.91	165.34	4.28	0.71
Minimum	0.00	252.21	33.34	75.89	0.00
First quartile	3.22	279.67	360.27	94.12	1.32
Median	4.57	287.30	490.45	97.40	1.60
Third quartile	5.48	295.28	595.11	99.00	1.99
Maximum	6.98	306.29	1,237.92	100.00	6.15

Next, a correlation analysis was performed considering the climatic variables, the soiling rate, and the AC electric generation, with data on a daily scale. According to Navlani *et al.* [15], Pearson's method is used when the data are symmetrically distributed (normal) to calculate the correlation coefficient. However, Spearman's method is recommended when the data has asymmetry and/or outliers. Kendall's method is used when the data is not required to follow some distribution. Because of this, all three methods were used to calculate the coefficients of all variables concerning AC electric power. The results are shown in Table 2.

Table 2. Correlation coefficients.

Variable	Method		
	Pearson	Spearman	Kendall
ac_energy	1.00	1.00	1.00
poa_irradiance	0.78	0.83	0.71
ambient_temp	0.43	0.32	0.21
wind_speed	0.27	0.30	0.20
soiling	0.01	0.02	0.02

To interpret the values shown in Table 2, it should be remembered that the correlation coefficient varies between "–1" and "1". When the value is positive, the direction of increase or decrease of the pair of variables is the same, and when the value is negative, the direction is the reverse. On the other hand, the absolute value "1" means that the magnitude of growth or decrease is equal for both variables, while the value "0" means that the pair of variables is not related at all. For the values between "0" and "1", we consider Ratner [16] , who states that "values between 0 and 0.3 (0 and –0.3) indicate a weak positive (negative) relationship. Values between 0.3 and 0.7 (–0.3 and –0.7) indicate a moderate positive (negative) relationship. Values between 0.7 and 1.0 (–0.7 and –1.0) indicate a strong positive (negative) relationship".

Considering the results in Table 2, solar irradiance has a strong and positive relationship with electric power. Ambient temperature has a positive and moderate relationship with electric power. The relationship of wind speed with electric power is positive, weak to moderate. While for this case, the relationship between soiling rate and electric power is practically independent.

Then, time curves were generated for the main variables of the data set. Figure 1 shows the behavior of the average solar irradiance (bars) vs. the electrical energy generated (line) for each of the years of the study period.

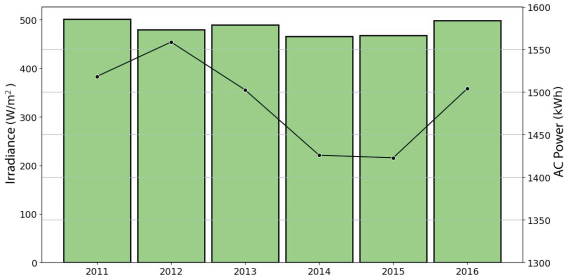

Figure 1. Solar irradiance vs. AC power

The average solar irradiance remained approximately constant during the study period. The energy generated reached its maximum value in 2012, decreased to minimum values during 2014 and 2015, and increased again in 2016.

Figure 2 shows the average monthly values of solar irradiance, electric power, and AC energy generated. The monthly energy production remained relatively constant during the whole period, and the behavior of the power almost perfectly followed the behavior of the solar irradiance.

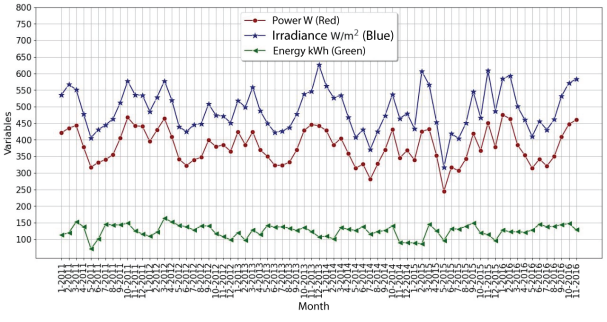

Figure 2. Monthly behavior of the variables

Figure 3 shows the weekly average values of solar irradiance and electric power, and the AC power generated per week of the year. The behavior of electric power and solar irradiance is almost identical. As for electric power, it reached its minimum value in the fifth week of the year and its maximum value in the thirteenth week. The energy generated decreased in the last five weeks of the year.

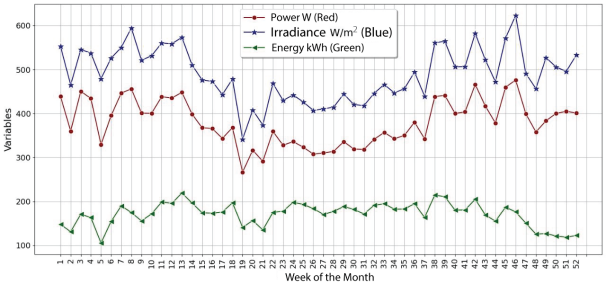

Figure 3. Weekly behavior of the variables

Figure 4 shows the daily average values of solar irradiance and electrical power, and the energy generated on each day of the month. Unlike the previous curves, in this case, the shapes of the three curves are approximately the same, with minimum values at the beginning and middle of the month. The curves do not have any defined trend (upward or downward).

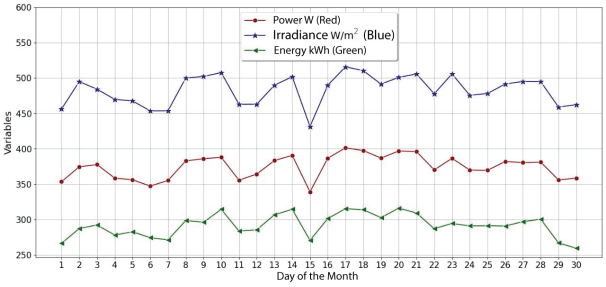

Figure 4. Daily behavior of the variables

To visualize the symmetry and dispersion of the data, Figure 5 shows the Box-Plot diagrams of each variable on a weekly scale. Previously, the values of each variable were scaled between zero and one for comparison.

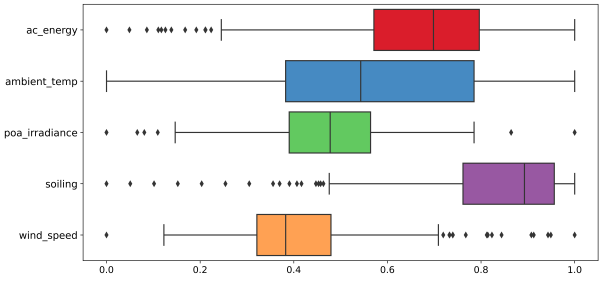

Figure 5. Box-Plot diagrams of the variables

According to Figure 5, it can be said that except ambient temperature, all the other variables present outliers. It is worth mentioning that they are slight outliers, according to Tukey's test [17], so they are not imputed. Solar irradiance is the most symmetrical variable, besides having few outliers. The soiling rate is the variable with the most outliers and the highest asymmetry. Ambient temperature is the variable with the highest dispersion in its data, and solar irradiance is the variable with the lowest dispersion.

3.2. Data modeling

Mathematical algorithms were applied to obtain forecasting models of the generated electric power. Specifically, a multiple linear regression model, an artificial neural network regression model, and a time series analysis model were obtained using weekly data. The data correspond to 310 weeks, from week 41, 2010, to week 47, 2016. The data from week 48 to week 50, 2016, were used to compare the forecast obtained from the three models mentioned above.

3.2.1. Multiple Linear Regression Algorithm

The multiple linear regression algorithm (MLR) is a supervised machine learning algorithm. The model obtained from this algorithm is linear in the parameters (coefficients) and not necessarily in the explanatory or predictor variables. The target variable is AC electric energy in kWh, while the predictor variables are solar irradiance ("poa_irradiance"), ambient temperature ("ambient_temp"), and wind speed ("wind_speed"). The soiling rate was not considered due to its null correlation with the target variable; moreover, in the first regression model, its coefficient in the regression equation was not statistically significant.

It was verified that there are no significant correlations between the predictor variables, as shown in Figure 6. All the absolute values of the correlation coefficient are less than 0.3, indicating weak relationships between the variables.

The data set, consisting of the objective and predictor variables, was randomly divided into two parts. The first part, composed of 80% of the data (256 records), was used to create and train the regression model. The second part, consisting of 20% (64 records), evaluated the model obtained in the training phase. The metrics used to assess the model were MAE and RMSE since, according to [18], they are statistical measures used to evaluate models. The R^2 was used, which according to Hair *et al.* [19], is a "measure of the proportion of the variance of the dependent variable concerning its mean that is explained by the independent or predictor variables". Alaraj *et al.* [7] use the same metrics except for the R^2.

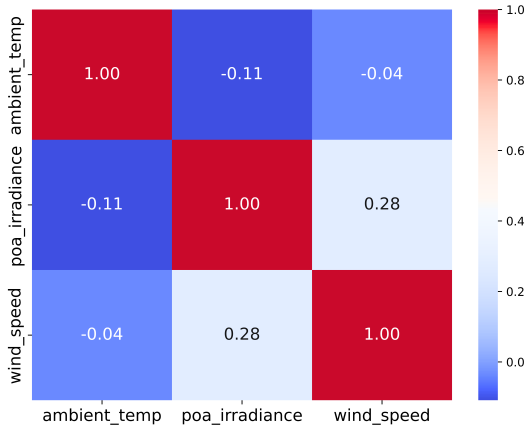

Figure 6. Correlation matrix of the predictor variables

After applying the algorithm, the coefficients 0.446, 0.043, and 4.002 were obtained for the predictor variables ambient temperature, solar irradiance, and wind speed, respectively. This indicates that a unit increase in the weekly average ambient temperature means an increase of 0.446 kWh; a unit increase in solar irradiance implies a rise of 0.043 kWh in power generation, and a unit increase in the weekly average of wind speed means an increase of about 4 kWh in weekly power generation. The value of the intercept is -125.98.

Table 3 shows the results of the performance metrics. The predictor variables explain about 81% of the variance of the objective variable, indicating that the model has a good fit quality. Since the average weekly generated electric power is 29 kWh, the obtained RMSE (2.87) corresponds to almost 10% of the mean, and the obtained MAE (2.30) is nearly 8%.

Table 3. RLM model indicators.

Indicator	Value Obtained
R^2	0.81
RMSE (kWh)	2.87
MAE (kWh)	2.30
Shapiro-Wilk test to the residuals	
Statistic	0.995
p-value	0.998

The Shapiro-Wilk test statistic was used to verify the statistical assumption of normality of the residuals required by this model, which has as its null hypothesis that the data are normally distributed. The test statistic varies between 0 and 1, indicating that the data are normally distributed when it is close to 1. To verify the rejection or not of the null hypothesis, the p-value is considered. Table 3 also shows that the value of 0.995 for the statistic and a p-value of 0.998 (greater than 5% statistical significance) suggest insufficient evidence to reject the null hypothesis that the residuals are normally distributed [20].

3.2.2. Artificial neural network algorithm

When applying the artificial neural network (ANN), the same data used for the multiple linear regression model, the same objective variable, and the same predictor variables were used. For cross-validation of the model, 80% of the data (256 records) were used for training, and 20% of the data (64 records) were used for evaluation.

According to Kapoor *et al.* [21], a "multilayer perceptron" model was used, composed of the input layer, the output layer, and a group of hidden layers between the input and output. Three layers were used for this study: an input layer, an output layer, and a hidden layer. All the layers are dense because, according to Moolayil [22], "a dense layer is a regular layer that connects all its neurons with all the neurons of the previous layer".

The activation functions were defined for each of the network layers. The rectified linear activation function (ReLU) was applied for the input and hidden layers, allowing only positive values to pass through. These two layers have a total of 256 neurons each. As for the output layer, this has a linear activation function so as not to limit the forecast values, and it has only one neuron, which is needed to forecast the electrical energy. According to Chollet [23], a loss function is required, which is used to control the deviation of the forecast from its expected value; so, for this study, MAE and MSE were used as loss functions. If the deviation is not adequate, its value is fed back to the input through an optimization function, which according to Chollet [23], updates the input weights and repeats the cycle. In this research, he used the root-mean-square propagation optimizer (RMSProp). Sharkawy *et al.* [3] also use an ANN with three layers but with a hyperbolic activation function.

Table 4 shows the results obtained by applying the ANN algorithm. The quality of the fit is around 88%, which is better than that obtained with the RLM model. Both RMSE and MAE are lower than those obtained with the MLR model. As for the analysis of the residuals, they are normally distributed since the test statistic is close to 1 and the p-value is higher than 5% of statistical significance.

Table 4. ANN model indicators.

Indicator	Value obtained
R^2	0.88
RMSE (kWh)	2.35
MAE (kWh)	1.85
Shapiro-Wilk test to the residuals	
Estadístico	0.967
p-valor	0.422

3.2.3. Time series analysis

When applying the analysis to the time series of the generated AC electric power, an ARIMA model is obtained, which requires three parameters: the order of the autoregressive part p, the order of integration d, and the order of the moving average q. If the series is seasonal, the three parameters for the seasonal part *(P, D, Q)* must also be considered. The model is obtained by applying the Box-Jenkins methodology, presented in [24] and mentioned in more detail in [25].

The methodology starts with data preparation, including transformation to stabilize variance and/or differencing to make the series stationary (parameter d is defined). The potential initial models are selected using the autocorrelation and partial autocorrelation functions (parameters p and q are defined). The parameters of the possible models are estimated, and the best of them is selected using a performance criterion. This criterion is usually the AIC (Akaike Information Criteria), which, according to [26] , is the most popular for selecting the best model. This is followed by the diagnostic stage, in which the residuals are analyzed to verify that they are equal or approximately equal to white noise. Finally, the model is used to forecast the time series.

Following the methodology, the extended Dickey-Fuller test is applied to verify the stationarity of the AC power series. According to Gujarati and Porter [27], this test is also known as the unit root test and is popular in determining the stationarity or non-stationarity of a time series. The test statistic was less than the three critical values (1%, 5%, 10%). The p-value is approximately equal to zero, so the null hypothesis of the existence of a unit root is rejected. Therefore, it can be said that the series in level is stationary. The latter implies that the parameter d is zero.

Figure 7 shows the graphs of the autocorrelation function (upper) and the partial autocorrelation function (lower) of the AC power series, considering up to 106 lags because the data present annual seasonality (52 weeks). There are at least two significant autocorrelation values. The series is confirmed to be seasonal, with the first seasonal value (week 52) significant for both graphs, which should be considered in the proposed model.

After performing the corresponding iterations minimizing the value of the AIC metric and checking the characteristics of the residuals obtained with each model, the model selected for forecasting is ARIMA(0,0,2)(1,1,1,1)52. The results obtained from forecasting with the ARIMA model and the other models are presented below.

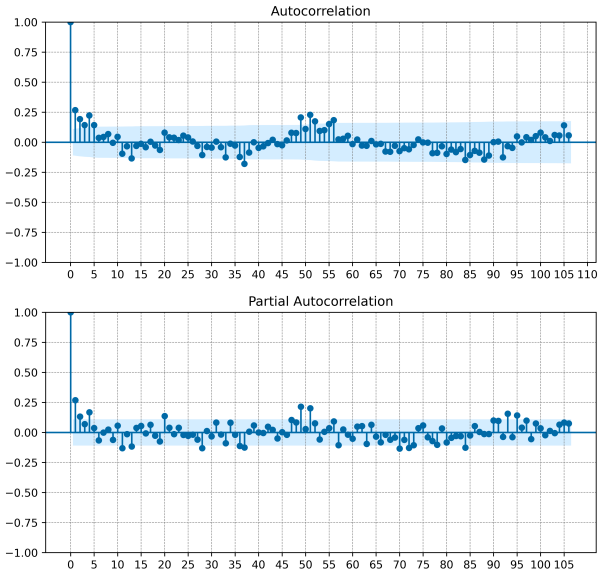

Figure 7. Autocorrelation and partial autocorrelation functions

3.2.4. Comparison of forecasts

Forecasts of electric power generated for weeks 48, 49, and 50 of 2016 were performed using each of the three models and were evaluated using RMSE, MAE, and MAPE metrics. Table 5, shows the results of the energy forecast in kWh, indicating that the ARIMA model forecast is the closest to the actual values of energy generated.

Table 5. Pronósticos de energía AC.

Week	Real Energy	MRL Forecast	ANN Forecast	ARIMA Forecast
48	23.63	28.15	30.13	25.04
49	20.20	23.68	22.48	18.54
50	21.92	27.36	26.02	22.97

Table 6 shows the performance metrics of the three models for the forecasts presented in Table 5. The ARIMA model is the best performer, with a MAPE of about 6% versus almost 20% for the other two models. The MAE and RSME are much lower for the ARIMA model.

Table 6. Performance of the models.

Metrics	Models		
	MLR	ANN	ARIMA
MAE (kWh)	4.48	4.29	1.38
RMSE (kWh)	4.55	4.63	1.40
MAPE (%)	20.39	19.16	6.35

The results in Table 6 agree with those reported by [25] since these authors state that moving average

methods are suitable for the short term and that regression methods are more appropriate for the medium and long term. For these authors, the "short term" is associated with periods of up to three months, while the "long term" refers to more than two years.

4. Conclusions

The behavior of the electrical energy generated over time is similar to the behavior of the solar irradiance for data with a resolution close to the minute resolution of the measurements, i.e., daily resolution. This result agrees with the correlation analysis, which showed that solar irradiance correlates 0.78 with the electrical energy generated. Regarding ambient temperature and wind speed, the correlation coefficient with electric power is between moderate and weak, with 0.43 and 0.27, respectively.

The predictor variables of the multiple linear regression model explain 81% of the variability of the target variable. The analysis of the residuals derived from this model indicates that they follow a normal distribution. As for the artificial neural network model, the coefficient of determination was 88%; the MAE and RMSE indicators were lower compared to the regression model, and the residuals were normally distributed.

While finding the appropriate ARIMA model, it was determined that the AC electric power level series is stationary and has annual stationarity. The model obtained minimizes the AIC criterion; the residuals are independently distributed and are not serially correlated.

When forecasting with the models obtained, the ARIMA model performed best, with the lowest values of the three error indicators: MAE, RMSE, and MAPE, with 1.38 kWh, 1.40 kWh, and 6.35%, respectively. The neural network model showed lower MAPE and MAE indicators than those obtained with the multiple linear regression model, but its RMSE metric was the highest of the three models.

References

[1] REN21, *Renewables 2022 - Global Status Report*. Renewables Now - Paris 2022, 2022. [Online]. Available: https://bit.ly/3I09MhE

[2] A. Kumar Mittal, K. Mathur, and S. Mittal, "A review on forecasting the photovoltaic power using machine learning," *Journal of Physics: Conference Series*, vol. 2286, no. 1, p. 012010, jul 2022. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/2286/1/012010

[3] A.-N. Sharkawy, M. Ali, H. Mousa, A. Ali, and G. Abdel-Jaber, "Machine learning method for solar PV output power prediction," *SVU-International Journal of Engineering Sciences and Applications*, vol. 3, no. 2, pp. 123–130, 2022. [Online]. Available: https://doi.org/10.21608/svusrc.2022.157039.1066

[4] D. V. S. Krishna Rao Kasagani and P. Manickam, "Modeling of solar photovoltaic power using a two-stage forecasting system with operation and weather parameters," *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, vol. 0, no. 0, pp. 1–19, 2022. [Online]. Available: https://doi.org/10.1080/15567036.2022.2032880

[5] D. Pattanaik, S. Mishra, G. P. Khuntia, R. Dash, and S. C. Swain, "An innovative learning approach for solar power forecasting using genetic algorithm and artificial neural network," *Open Engineering*, vol. 10, no. 1, pp. 630–641, 2020. [Online]. Available: https://doi.org/10.1515/eng-2020-0073

[6] M. N. Akhter, S. Mekhilef, H. Mokhlis, and N. Mohamed Shah, "Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques," *IET Renewable Power Generation*, vol. 13, no. 7, pp. 1009–1023, 2019. [Online]. Available: https://doi.org/10.1049/iet-rpg.2018.5649

[7] M. Alaraj, A. Kumar, I. Alsaidan, M. Rizwan, and M. Jamil, "Energy production forecasting from solar photovoltaic plants based on meteorological parameters for qassim region, Saudi Arabia," *IEEE Access*, vol. 9, pp. 83 241–83 251, 2021. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3087345

[8] K. Anuradha, D. Erlapally, G. Karuna, V. Srilakshmi, and K. Adilakshmi, "Analysis of solar power generation forecasting using machine learning techniques," *E3S Web Conf.*, vol. 309, p. 01163, 2021. [Online]. Available: https://doi.org/10.1051/e3sconf/202130901163

[9] M. Borunda, A. Ramírez, R. Garduno, G. Ruiz, S. Hernández, and O. A. Jaramillo, "Photovoltaic power generation forecasting for regional assessment using machine learning," *Energies*, vol. 15, no. 23, p. 8895, 2022. [Online]. Available: https://doi.org/10.3390/en15238895

[10] J. VanderPlas, *Python data science handbook: Essential tools for working with data*. O'Reilly Media, Inc., 2016. [Online]. Available: https://bit.ly/3BkwSeM

[11] D. Cielen, A. Meysman, and M. Ali, *Introducing Data Science: Big Data, Machine Learning, and more, using Python tools*. Manning Publication, 2016. [Online]. Available: https://bit.ly/42wWD80

[12] DuraMAT. (2023) PVDAQ time-series with soiling signal - Data and Resources. Durable Module Materials Consortium. [Online]. Available: https://bit.ly/42NKc7t

[13] SolarDesignTool, *Sanyo HIP200BA3 (200W) Solar Panel*. SolarDesignTool, 2023. [Online]. Available: https://bit.ly/3pu1dFk

[14] W. McKinney, *Python for Data AnalysisOreilly and Associate Series*. "O'Reilly Media, Inc.", 2013. [Online]. Available: https://bit.ly/3HZnfGr

[15] A. Navlani, A. Fandango, and I. Idris, *Python Data Analysis: Perform data collection, data processing, wrangling, visualization, and model building using Python*. Packt Publishing Ltd, 2021. [Online]. Available: https://bit.ly/42voHsb

[16] B. Ratner, *Statistical and Machine-Learning Data Mining:: Techniques for Better Predictive Modeling and Analysis of Big Data*. CRC Press, 2017. [Online]. Available: https://bit.ly/3VPx933

[17] I. A. Uribe, "Guía metodológica para la selección de técnicas de depuración de datos," Master's thesis, Universidad Nacional de Colombia, Medellín, Colombia, 2010. [Online]. Available: https://bit.ly/3VQ5n6t

[18] D. C. Montgomery, C. L. Jennings, and M. Kulahci, *Introduction to Time Series Analysis and Forecasting*. Wiley Series in Probability and Statistics, 2015. [Online]. Available: https://bit.ly/3LTZiRS

[19] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*. Pearson Education Limited, 2013. [Online]. Available: https://bit.ly/3LWEHMN

[20] V. Platas García, *Contrastes de normalidad*. Universidade de Santiago de Compostela. Facultade de Matemáticas, 2021. [Online]. Available: https://bit.ly/3MfxZ5Z

[21] A. Gulli, A. Kapoor, and S. Pal, *Deep Learning with TensorFlow 2 and Keras*. Packt Publishing, 2019. [Online]. Available: https://bit.ly/42MPT5r

[22] J. Moolayil, *Learn Keras for Deep Neural Networks: A Fast-Track Approach to Modern Deep Learning with Python*. Apress, 2018. [Online]. Available: https://bit.ly/3nMtrL4

[23] F. Chollet, *Deep Learning with Python*. Manning Publications Company, 2017. [Online]. Available: https://bit.ly/3LV4a9w

[24] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics, 2008. [Online]. Available: https://bit.ly/44OEALU

[25] S. Makridakis, S. Wheelright, and R. Hyndman, *Manual of Forecasting: Methods and Applications*. Wiley-Interscience, 1998. [Online]. Available: http://dx.doi.org/10.13140/RG.2.1.2528.4880

[26] T. C. Mills, *Applied Time Series Analysis: A Practical Guide to Modeling and Forecasting*. Elsevier, 2019. [Online]. Available: https://bit.ly/42sM5Xd

[27] D. N. Gujarati and D. C. Porter, *Econometría*. McGraw-Hill Interamericana, 2010. [Online]. Available: https://bit.ly/44Tq0mc