Salud, Ciencia y Tecnología. 2025; 5:1971
doi: 10.56294/saludcyt20251971

ORIGINAL

SARIMA models for power evolution in photovoltaic systems

Modelos SARIMA para la evolución de la Potencia en sistemas fotovoltaicos

Christian Paul Reyes Orozco[1], Diana Carolina Campaña Días[1], Elsa Amalia Basantes Arias[1], Jesús Rodríguez[2], Juan Ennis Espinoza González[3], Chasiluisa Yanchatuña Sandra Marisol[3]

[1]Escuela Superior Politécnica de Chimborazo (ESPOCH)- Riobamba, Ecuador.
[2]Universidad UTE. Quito, Ecuador.
[3]Investigador Independiente. Ecuador.

Cite as: Reyes Orozco CP, Campaña Días DC, Basantes Arias EA, Rodríguez J, Espinoza González JE, Sandra Marisol CY. SARIMA models for power evolution in photovoltaic systems. Salud, Ciencia y Tecnología. 2025; 5:1971. https://doi.org/10.56294/saludcyt20251971

Submitted: 09-02-2025 **Revised:** 27-05-2025 **Accepted:** 10-08-2025 **Published:** 11-08-2025

Editor: Prof. Dr. William Castillo-González

Corresponding Author: Juan Ennis Espinoza González

ABSTRACT

Introduction: the increasing use of renewable energy in power generation systems has highlighted the need for efficient schemes to predict model parameters. In particular, photovoltaic systems require accurate tools to model and forecast solar energy generation behavior.
Objective: to formulate SARIMA models with high accuracy in fitting, explanation, and prediction of energy yields in solar photovoltaic systems, specifically focused on the plant located at Plaza del Duque de Béjar, Spain.
Method: a fitting strategy based on genetic algorithms was adopted to accelerate the estimation of the SARIMA model using hourly solar photovoltaic generation data. The auto.arima package in RStudio was employed as a methodological tool, enabling automatic selection and optimization of the best model parameters. **Results:** the selected model was SARIMA (5,0,0)(2,1,0)242424, characterized by a stationary stochastic process with a clear seasonal component. The model showed remarkable estimation accuracy, with low standard errors in the autoregressive coefficients. Additionally, the model residuals were well-adjusted, displaying independence and absence of serial autocorrelation.
Conclusions: the proposed model demonstrated excellent predictive performance, supported by training error metrics (ME (Mean Error)= -1,344268 and MASE (Mean Absolute Scaled Error)= 0,7048786). Its sound mathematical structure and strong fit make it a reliable tool for forecasting photovoltaic solar energy in systems with similar characteristics.

Keywords: Power; Generation Systems; Photovoltaic; SARIMA Models; Genetic Algorithms; Rstudio.

RESUMEN

Introducción: el creciente uso de energías renovables en los sistemas de generación eléctrica ha puesto de manifiesto la necesidad de contar con esquemas eficientes para la predicción de parámetros de modelos asociados a la producción energética. En particular, los sistemas fotovoltaicos requieren herramientas precisas que permitan modelar y anticipar el comportamiento de la generación de energía solar.
Objetivo: formular modelos SARIMA altamente precisos en cuanto a ajuste, explicación y predicción de los rendimientos energéticos en sistemas solares fotovoltaicos, centrando el estudio en la planta ubicada en la Plaza del Duque de Béjar, España.
Método: se adoptó una estrategia de ajuste basada en algoritmos genéticos para acelerar la estimación del modelo SARIMA a partir de datos horarios de generación solar fotovoltaica. Se utilizó el paquete auto.arima de RStudio como herramienta metodológica, lo que permitió realizar una búsqueda automática de los mejores parámetros del modelo y su correspondiente optimización.

Resultados: el modelo SARIMA seleccionado fue el (5,0,0)(2,1,0)242424, caracterizado por un proceso estocástico estacionario con una marcada componente estacional. La precisión del modelo fue notable, evidenciada por errores estándar reducidos en los coeficientes autorregresivos. Además, los residuos del modelo se ajustaron adecuadamente, mostrando independencia y ausencia de autocorrelación serial.
Conclusiones: **e**l modelo propuesto presentó un excelente desempeño predictivo, respaldado por errores cuantificados en la fase de entrenamiento (ME (Error Medio)= -1,344268 y MASE (Error Absoluto Medio Escalado) = 0,7048786). Su adecuada estructura matemática y buen ajuste lo convierten en una herramienta confiable para la predicción de energía solar fotovoltaica en sistemas con características similares.

Palabras clave: Potencia; Sistemas de Generación; Fotovoltaicos; Modelos SARIMA; Algoritmo Genéticos, RStudio.

INTRODUCTION

Solar energy is a viable resource for industrial, commercial, and domestic uses.[1] In recent years, photovoltaic solar energy (based on photovoltaic systems) has gained popularity because it is abundant, sufficient, clean, and environmentally friendly.[2] Since the weather can have a significant impact on electricity generation and reduce the integration of photovoltaic systems into the grid, accurate short-term forecasts are essential to optimize the power generated by solar photovoltaic power plants.[3]

Accurate forecasting of power in solar energy systems helps to effectively manage energy generation and storage on a daily or hourly basis.[4] It is a fundamental resource for photovoltaic systems to participate in the energy market, enabling efficient strategy planning.[5] At this point, a comprehensive review of the literature has described various methods for achieving this accuracy, including statistical approaches for time series forecasting based on the ARIMA (Autoregressive Integrated Moving Average) methodology,[6] machine learning techniques for building artificial neural network models,[7] adaptive prediction models of energy collected for solar energy harvesting sensor networks,[8] physical models based on numerical weather forecasts and satellite images,[9] and hybrid approaches combining the first three methods.[10] These are the methodologies that help photovoltaic plants participate more effectively in the energy market.

Specifically, estimates of the power generated in watts for solar energy systems are typically modeled by fitting various methodologies, which lead to typologies for specific prediction models. The most recent advances in this field are those focused on the contributions derived from the implementation of artificial intelligence in the field of neural networks, specifically support vector machines (SVM) or short-term and long-term memory recurrent neural networks (LSTM),[11,12] recurrent neural networks under the backpropagation paradigm,[13] autoregressive neural networks,[14,15] autoregressive neural network models with exogenous input,[16] swarm intelligence,[17,18] and uses of machine learning, including deep learning.

With an emphasis on solar energy systems, deep learning models have been studied in depth for their potential application in time series prediction. While taking into account the limitations and particular factors of each architecture, such as training efficiency, interpretability, and data handling, these models focus on temporal dependencies, sequential data management, and improving prediction accuracy.[19]

The above arguments are subject to the type of data handled and the approach applied within the research paradigm to obtain adequate modeling. These methods for model construction are divided according to underlying interest, i.e., the analysis of the physical paradigm, which considers the use of predictor variables associated with meteorological records and solar irradiation impacts. Alternatively, a statistical approach can be used to predict the behavior of the phenomenon that guides the obtaining of solar power generation from a historical data matrix.[20,21,22]

On the other hand, the characterization of a deep-rooted seasonal condition in the temporal data, in line with the recommendation based on the literature review, the parameter estimates in seasonal autoregressive integrated moving average (SARIMA) models, have been successful in predicting a particular phenomenon.[5,21,23,24,25] Essentially, the extensive existing literature develops an analytical and comparative analysis of different models for predicting power generation by photovoltaic systems, focusing on the development of hybrid models with commendable strengths and implicit weaknesses.[10,26]

The significant advantage of using SARIMA modeling lies in its simplicity, which usually adjusts easily to stationary stochastic processes with a marked seasonal component.[6] Therefore, when a series exhibits a non-stationary process, transformations are performed to convert the series under analysis into a stationary one and to adjust the ARIMA model parameters appropriately. This model can be constructed by implementing sophisticated statistical techniques.[27] The latter approach involves optimizing the selection by using statistics and determining decision criteria, such as the Akaike Information Criterion (AIC) and the sum of squared errors (SSE).

In the current context of energy transition and the use of renewable sources, photovoltaic generation systems have taken on a leading role. In this sense, accurate prediction of the power generated is a strategic focus for operational planning and the sustainable development of these technologies. According to the specialized literature, seasonal autoregressive integrated moving average (SARIMA) models have proven to be highly effective in predicting phenomena with seasonal components, as evidenced by several successful empirical studies.

Despite the rise of hybrid models that integrate artificial intelligence and machine learning techniques, SARIMA approaches continue to offer significant advantages, particularly due to their simplicity, statistical robustness, and ability to adjust to stationary stochastic processes with strong seasonality.(6) This is particularly valuable when the time series exhibits non-stationary behavior, allowing it to be converted into a stationary series through appropriate transformations and the model parameters to be correctly adjusted. In addition, the construction of SARIMA models is enhanced by the use of advanced statistical tools and the implementation of decision criteria such as the Akaike Information Criterion (AIC) and the Sum of Squared Residuals (SSE), which allow for optimal model selection.(27)

Within this framework, the present study is justified by the need to model the electrical power generation series in a photovoltaic solar system located in the Plaza del Duque de Béjar, Villanueva del Duque, Spain. This series is characterized by its daily periodicity (every 24 hours) over a continuous period of four years (2019–2022), which reveals a clear seasonal component. Thus, the research is proposed as a first step towards the stochastic modeling of this series, establishing a solid basis for future energy forecasts that improve the management of renewable resources.

According to the background information provided, the development of this scientific article comprises a first stage defined by the description of the seasonal component present in the power series generated in the photovoltaic system located in the Plaza del Duque de Béjar, Villanueva del Duque, Spain, due to the nature of the information collection every 24 hours, 365 days a year, during the coverage period from 2019 to 2022. Secondly, the SARIMA model parameters were estimated, incorporating an optimization strategy based on genetic algorithms to improve the accuracy of the power generated by renewable energy systems. This approach enabled the acceleration of the model adjustment process using hourly solar photovoltaic generation data. Finally, the results obtained are discussed, and conclusions derived from the analysis are formulated, highlighting the effectiveness of the proposed model in similar contexts.

METHOD

This study adopts a quantitative, observational, and retrospective design based on historical hourly solar photovoltaic power generation data recorded between 2019 and 2022 at the plant located in Plaza del Duque de Béjar, Spain. Given the identified seasonal behavior in the time series, the SARIMA (Seasonally Adjusted Autoregressive Integrated Moving Average) model was selected as the modeling technique due to its suitability for capturing stochastic dynamics with a seasonal component.

To improve the parameter estimation process, an adjustment strategy based on genetic algorithms was implemented, which allowed for the optimization of model selection and accelerated convergence towards a better-fitting structure. The modeling process was carried out using the auto.arima function of the forecasting package in RStudio, which facilitates the automatic selection and refinement of model parameters based on criteria such as the Akaike Information Criterion (AIC) and the sum of squared errors (SSE).

The methodological procedure was developed in the following stages:

1. Preprocessing and transformation of the data to achieve stationarity of the series.
2. Application of the auto.arima algorithm to identify the optimal parameters of the SARIMA model.
3. Evaluation of model performance using statistical indicators such as ME (Mean Error) and MASE (Mean Absolute Scaled Error).
4. Residual diagnosis to verify independence and absence of serial autocorrelation.

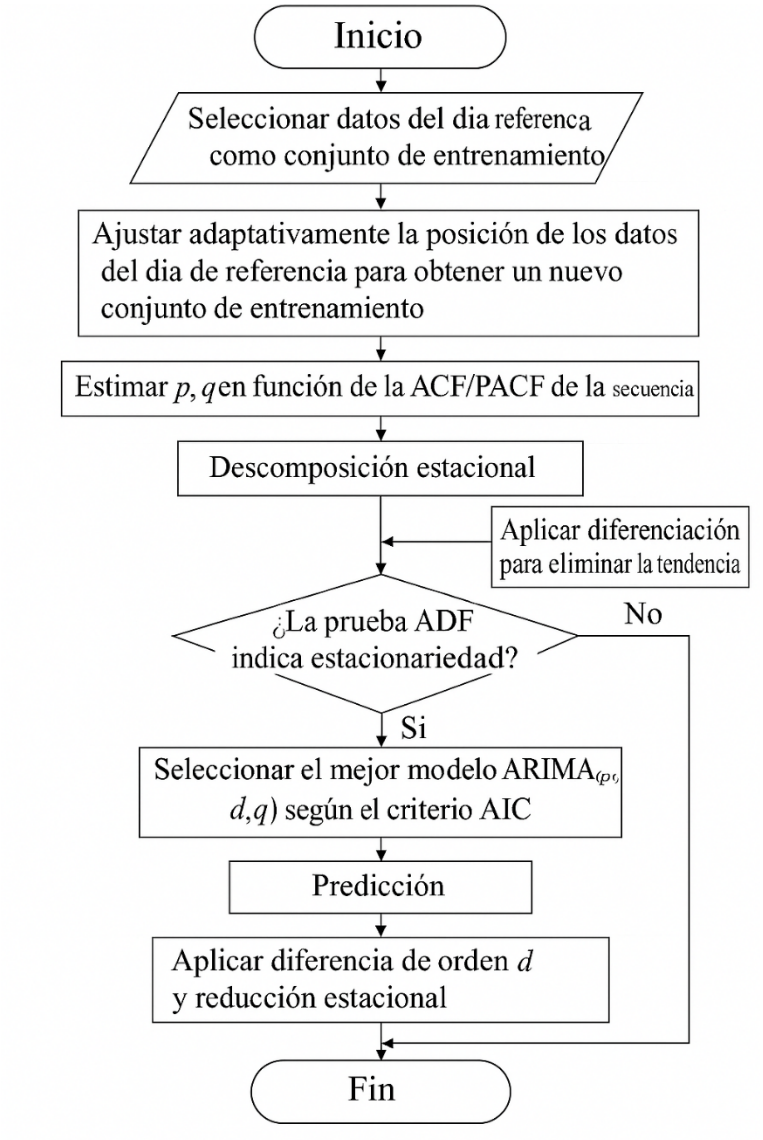
This methodology provided a robust and efficient framework for modeling and predicting solar power generation, in accordance with best practices in time series analysis.(4)

Data description

The construction of statistical models based on data referring to surface solar radiation (SARAH3) from the records of the Photovoltaic Geographic Information System (PVGIS). This dataset provides records at the location with Latitude: 38,393 and Longitude: -5,000, in decimal degrees, for the power generated (W) by a photovoltaic system centered in the Plaza del Duque de Béjar, Villanueva del Duque, Spain. This location allows for management in accordance with the European Union and other countries' regulations.(28) The window for photovoltaic generation is oriented to optimize incident solar radiation during the time coverage from 8:10 a.m. to 4:10 p.m., so that incident solar radiation is at its maximum. The database is recorded at 24-hour intervals, specifically after 10 minutes determined directly from each hour, from 2019 to 2021.

Protocol for the construction of SARIMA models

This section is necessary to evaluate the power (W) of solar energy generated by photovoltaic systems, with the aim of constructing SARIMA models, which essentially requires a systematic step-by-step data analysis process, highlighting the importance of using statistical techniques and tests for the proper modeling of the underlying structure in the data recorded for the power generated in photovoltaic systems (figure 1).


Inicio
Seleccionar datos del dia referenca como conjunto de entrenamiento
Ajustar adaptativamente la posición de los datos del dia de referencia para obtener un nuevo conjunto de entrenamiento
Estimar p, q en función de la ACF/PACF de la secuencia
Descomposición estacional
Aplicar diferenciación para eliminar la tendencia
¿La prueba ADF indica estacionariedad?
No
Si
Seleccionar el mejor modelo ARIMA(p, d,q) según el criterio AIC
Predicción
Aplicar diferencia de orden d y reducción estacional
Fin

Source: Kushwaha et al.[29]

Figure 1. SARIMA methodology flowchart

Figure 1 illustrates a diagram of how the solar energy prediction process should be carried out using the SARIMA method. This process begins with the identification of historical data, adjusting the descriptive and statistical measures required for an estimation of the parameters p and q, performing a statistical analysis, eliminating trends, and confirming the stability of the series. A key step is selecting the best ARIMA model for making predictions using the identified structure. Finally, the final predictions of power (W) in solar energy generation in photovoltaic systems are obtained.

Seasonal ARIMA (SARIMA) process methodology

If the series $\{Y_t\}$ has a marked seasonal component of period s, this can be eliminated by applying the seasonal difference operator with a lag of order s, equivalent to the data collection periods, thus obtaining a series $\{Z_t\}$ with an ARMA process structure.

In the context of a time series that exhibits a certain level of stationarity, when the mean and variance of the time series do not show significant fluctuations, the time series is said to be stationary.[23] This means that the curve fitted to the sample sequence can continue in the future depending on existing circumstances. In a different scenario, the regular difference operator must be applied to stabilize the mean value of the time series.[30] On the other hand, when the series exhibits seasonal characteristics, consideration should be given to correcting the seasonality inherent in the data collection with the seasonal difference operator to ensure that this component, which distorts the correct identification of the parameters in the model, is eliminated.[12] The key is to identify these periods of seasonality in time and to use a seasonal index to identify them.

Therefore, a time process that exhibits a regular trend and a marked seasonal component can define an integration parameter of the form d and D. In addition, considering d and D to be non-negative integers, then $\{Z_t\}$ is a multiplicative seasonal process ARIMA(p,d,q)×(P,D,Q)$_s$ with period s, if the series differentiated in its regular and seasonal components is denoted by $Z_t = (1 - B)^d \; (1 - B^s)^D \; Y_t$, it represents a causal ARMA process defined by:

$$\varphi_p(B)\Phi_P(B^s)Z_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \,, \{\varepsilon_t\} \sim \varepsilon_t \, N(0,\sigma^2) \qquad (1)$$

This type of model is called a seasonal ARIMA model. Essentially, according to [9], the seasonal ARIMA model includes autoregressive and moving average terms with lags of order s.

The seasonal ARIMA process methodology is used to forecast the future of a time series that exhibits a seasonal pattern (s= 24 hours). To do this, historical data from a time series is used to estimate the parameters of the seasonal ARIMA model. Once the parameters have been calculated and the model assumptions validated, the model can be used to forecast the future of the time series.[30]

In the verification stage to ensure that the time series is stationary, the Augmented Dickey Fuller (ADF) test is applied, which consists of determining whether a sequence contains a unit root; if so, the time series is considered non-stationary; if not, it is considered stationary.[6] This study uses ADF statistics on solar power generation collected from photovoltaic systems in the Plaza del Duque de Béjar, Spain. The absence of unit roots in the time series indicates that the collected time series is stable in terms of mean and variability, which is appropriate for constructing SARIMA models.

The models obtained from a seasonal ARIMA process have multiple applications for forecasting temporal variables, which must be implemented for processes that exhibit a variety of seasonal patterns, including additive and multiplicative seasonal patterns. They can also be used to predict time series that show a variety of non-stationary patterns, including linear and non-linear non-stationary patterns.[31] The robust advantage of using this type of model lies in the advanced accuracy of short-term prediction results.[32]

In terms of interpretation, autoregressive and moving average models are combined to create the ARMA model. The moving average (MA) parameters closely simulate the sample values to explain the disturbance effects in previous periods. In contrast, the autoregressive (AR) parameters assume that the sample values at current times for the time series are related to the values recorded at previous times. The weighted average of the earlier values taken by the time series represents the behavior of the sample values at current times.

RESULTS

Stationarity tests in the series defining power in photovoltaic systems

Implementation of the Dickey Fuller statistic augmented according to the adf_test command (Table 1), applied in R-Studio to determine the existence of unit roots in the observed series consisting of the power in photovoltaic systems in the Plaza del Duque de Béjar, Spain. The formulation of the hypotheses defined focuses on a null hypothesis (H_0) that posits the existence of a unit root, indicating that the series is non-stationary, while the alternative hypothesis (H_1) suggests that there is no unit root, implying that the series is stationary.

Table 1. Augmented Dickey Fuller Statistic

Augmented Dickey-Fuller Test	Value
Dickey-Fuller	-15,808
Lag order	29
p-value	0

Based on the findings presented in table 1, it can be argued that the p-value=0,01 represents a value lower than the minimum permissible error (α=0,05), a determining factor in concluding that the time series constituted by the power recorded in photovoltaic systems is stationary.

Adjustment of the SARIMA model using genetic algorithms

The model used for visualization is obtained using genetic algorithms to speed up estimation. The main objective is not to obtain a perfect forecast, but to obtain a good visualization of the results.

The use of genetic algorithms for parameter estimation and selection in SARIMA models has proven to be an effective strategy for accelerating the adjustment process in high-dimensional and highly seasonal environments, such as the power generation series in solar systems. Table 2 shows a comparative evaluation between multiple combinations of parameters (p,d,q)(P,D,Q)[s](p,d,q)(P,D,Q)[s](p,d,q)(P,D,Q)[s], where the negative log-likelihood value or some related function (presumably the AIC or log-likelihood) is used as the selection metric.

During the initial process, which uses approximations to speed up calculations, various models are explored. Some models are immediately discarded because they generate infinite results ("Inf"), indicating convergence problems, overfitting, or numerical instability. On the other hand, viable models such as ARIMA(5,0,0)(2,1,0)[24] consistently show the lowest error values, positioning them as optimal candidates within the search space.

The ARIMA(5,0,0)(2,1,0)[24] model ultimately stands out as the best fit after a second "no approximation" refinement phase, confirming its robustness. This model adequately captures the daily seasonal structure (frequency 24) of the series and represents a stationary stochastic process in differences, with a dominant autoregressive structure at both the regular and seasonal levels.

It is important to note that, although the objective was not to obtain a perfect forecast, the quality of the fit achieved reinforces the validity of the model for visualization and exploratory analysis, as well as its possible future application in predictive scenarios. Furthermore, the use of genetic h l algorithms allows overcoming limitations inherent to traditional search methods such as *grid search*, by avoiding getting trapped in local minima.

Table 2. Results of the genetic algorithm for SARIMA estimation

```
Fitting models using approximations to speed things up...
 ARIMA(2,0,2)(1,1,1)[24] with drift        : Inf
 ARIMA(0,0,0)(0,1,0)[24] with drift        : 683269,9
 ARIMA(1,0,0)(1,1,0)[24] with drift        : 661046,7
 ARIMA(0,0,1)(0,1,1)[24] with drift        : Inf
 ARIMA(0,0,0)(0,1,0)[24]                   : 683267,9
 ARIMA(1,0,0)(0,1,0)[24] with drift        : 668133,7
 ARIMA(1,0,0)(2,1,0)[24] with drift        : 658551,4
 ARIMA(1,0,0)(2,1,1)[24] with drift        : Inf
 ARIMA(1,0,0)(1,1,1)[24] with drift        : Inf
 ARIMA(0,0,0)(2,1,0)[24] with drift        : 676584,4
 ARIMA(2,0,0)(2,1,0)[24] with drift        : 658430,5
 ARIMA(2,0,0)(1,1,0)[24] with drift        : 660897,2
 ARIMA(2,0,0)(2,1,1)[24] with drift        : Inf
 ARIMA(2,0,0)(1,1,1)[24] with drift        : Inf
 ARIMA(3,0,0)(2,1,0)[24] with drift        : 658400,5
 ARIMA(3,0,0)(1,1,0)[24] with drift        : 660882,5
 ARIMA(3,0,0)(2,1,1)[24] with drift        : Inf
 ARIMA(3,0,0)(1,1,1)[24] with drift        : Inf
 ARIMA(4,0,0)(2,1,0)[24] with drift        : 658361,5
 ARIMA(4,0,0)(1,1,0)[24] with drift        : 660851,7
 ARIMA(4,0,0)(2,1,1)[24] with drift        : Inf
 ARIMA(4,0,0)(1,1,1)[24] with drift        : Inf
 ARIMA(5,0,0)(2,1,0)[24] with drift        : 658333,6
 ARIMA(5,0,0)(1,1,0)[24] with drift        : 660820,9
 ARIMA(5,0,0)(2,1,1)[24] with drift        : Inf
 ARIMA(5,0,0)(1,1,1)[24] with drift        : Inf
 ARIMA(5,0,1)(2,1,0)[24] with drift        : 658340,1
 ARIMA(4,0,1)(2,1,0)[24] with drift        : 658337
 ARIMA(5,0,0)(2,1,0)[24]                   : 658331,6
 ARIMA(5,0,0)(1,1,0)[24]                   : 660818,9
 ARIMA(5,0,0)(2,1,1)[24]                   : Inf
 ARIMA(5,0,0)(1,1,1)[24]                   : Inf
 ARIMA(4,0,0)(2,1,0)[24]                   : 658359,5
 ARIMA(5,0,1)(2,1,0)[24]                   : 658338,1
 ARIMA(4,0,1)(2,1,0)[24]                   : 658335
 Now re-fitting the best model(s) without approximations...
 ARIMA(5,0,0)(2,1,0)[24]               : 658847,1
 Best model: ARIMA(5,0,0)(2,1,0)[24]
```

Estimation of SARIMA model parameters

In the photovoltaic system power series used to estimate the SARIMA model parameters, the auto.arima package in RStudio was used to automatically search for the best parameters and optimize the model.

Table 3 shows the values of the parameters estimated in the ARIMA(5,0,0)(2,1,0)$_{24}$model identified. In addition, the accuracy of the parameter estimates in the model is noteworthy, due to the small values for the standard error in the r autoregressive coefficients, both for the regular part of the series and for the seasonal part.

On the other hand, the variance of the model, according to the power records in the photovoltaic systems, determines that σ^2 = 4,519e+09, and the quality of the model fit (log likelihood = -329415,5) reveals that there is a very high quantification, which represents a good fit according to the parameters defined in the mathematical structure.

Table 3. SARIMA parameter estimation

Parameters	**ar1**	**ar2**	**ar3**	**ar4**	**ar5**	**sar1**	**sar2**
Coefficients	0,6608	0,0955	-0,0057	-0,0173	-0,0343	-0,6414	-0,3023
Standard error (s.e.)	0,0062	0,0074	0,0074	0,0074	0,0062	0,0059	0,0059

The RStudio software adapts series to the models ARIMA(p,d,q)x(P,D,Q)$_s$, i.e., the autoregressive and moving average parts for the regular and seasonal components in the series under study. In this regard, the parameter estimates in the model determined using the package are shown (table 2), in this case a multiplicative model:

$$ARIMA\ (5,0,0)x\ ARIMA(2,1,0)_{24} \rightarrow \phi_5(L).\Phi_2(L^{24}).Y_t = \varepsilon_t \qquad (2)$$

The mathematical representation is formulated specifically using the following expression:

$$(1 - 0,6608L - 0,0955L^2 + 0,0057L^3 + 0,0173L^4 + 0,0343L^5)(1 + 0,6414L^{24} + 0,3023L^{24}).Y_t = (1 + 0,6414L^{24} + 0,3023L^{24}).\varepsilon_t \qquad (3)$$

Simplifying the above structure, we obtain:

$$Y_t = \mu + 0,6608.Y_{t-1} + 0,0955.Y_{t-2} - 0,0057.Y_{t-3} - 0,0173.Y_{t-4} - 0,0343.Y_{t-5} - 0,6414.Y_{t-24} - 0,3023.Y_{t-24} + \varepsilon_t \qquad (4)$$

This model is used to analyze stationary stochastic processes with a marked seasonal component of 24 hours in daily records for power in photovoltaic systems. In this context, this graphical representation can be highlighted in figure 1, which represents a deterministic process defined by autoregressive parameters within the five (previous 05 hours) time lags. About the random component, it is characterized in the seasonal structure of the series by a marked component defined by two previous autoregressive parameters, in a time frame consisting of daily records every 24 hours.

Once the mathematical equation has been obtained, errors in the training data set are determined. This refers to the determination of discrepancies between the output values predicted by a model and the actual target values in the training data set. Among the measures used to evaluate the model's performance in the training set are the Mean Error (ME), the Root Mean Square Error (RMSE), and the Mean Absolute Error (MAE).

Table 4 summarizes the main error statistics used to evaluate the performance of the SARIMA(5,0,0)(2,1,0)[24] model during the training phase. These metrics allow us to quantify the accuracy and robustness of the model's fit to the historical data, providing quantitative evidence of its ability to model hourly photovoltaic solar energy generation adequately.

- ME (Mean Error = -1,344268): indicates a slight negative bias in the predictions, suggesting that the model tends to slightly underestimate the actual values. However, since the value is close to zero, the bias is minimal and acceptable.
- RMSE (Root Mean Square Error = 67185,63): provides a measure of the average magnitude of the prediction error, heavily penalizing large errors. Although the value is relatively high in absolute terms, it should be interpreted in relation to the scale of the power generated and is considered consistent with the inherent variability of the series.
- MAE (Mean Absolute Error = 29122,39): reflects the average absolute value of the errors, which gives a clear idea of the average magnitude of the error without considering its direction. Its value is consistent with the RMSE, indicating stability in the model without the presence of large outliers.
- MASE (Scaled Mean Absolute Error = 0,7048786): this is a scaled metric that allows the accuracy of the model to be compared to a naïve model (without predictive capacity). A value less than 1 indicates that the SARIMA model offers better predictive performance than a base model, validating its suitability for the analysis and visualization of the series.
- ACF1 (Autocorrelation of the residual at lag 1 = -0,00019): evaluates whether the model residuals

show immediate serial correlation. The value is practically zero, confirming that the residuals are independent and random, fulfilling a key validity assumption for SARIMA models.

Taken together, these indicators support the statistical consistency, low bias, generalizability, and good fit of the proposed model, fully justifying its selection as the base structure for future analysis and predictions in similar energy contexts.

Table 4. Evaluation metrics for training data set

Error statistics	Measures on training set
ME (Mean Error)	-1,344268
RMSE (Root Mean Square Error)	67185,6
MAE (Mean Absolute Error)	29122,39
MASE (Scaled Mean Absolute Error)	0,7048786
ACF1 (Autocorrelation of the residual at lag 1)	-0,0001895767

A low training error indicates that the model has learned well from the data during the training process, but a very low error can also be a sign of overfitting. A negative value for the mean error (ME=-1,344268) indicates that, on average, the predictions are lower than the observed values. Another statistic, represented by the root mean square error (RMSE=67185,63), expresses the magnitude of the error in the predictions; a lower value indicates a better fit. The mean absolute error (MAE=29122,39) provides a clear measure of the accuracy of the model, and the value close to zero for ACF1 suggests that there is no significant correlation in the prediction errors (table 4).

Validation of the SARIMA model

The validation of an identified and estimated model is an essential task to check whether the SARIMA(5,0,0)(2,1,0)$_{24}$model fits the data adequately, reviewing parameters such as stability and seasonality. The first premise must be to assess whether the residuals represent Gaussian white noise processes. Therefore, a residual plot must be constructed to verify that the expected mean value for the residuals is zero for all values over time (figure 2).

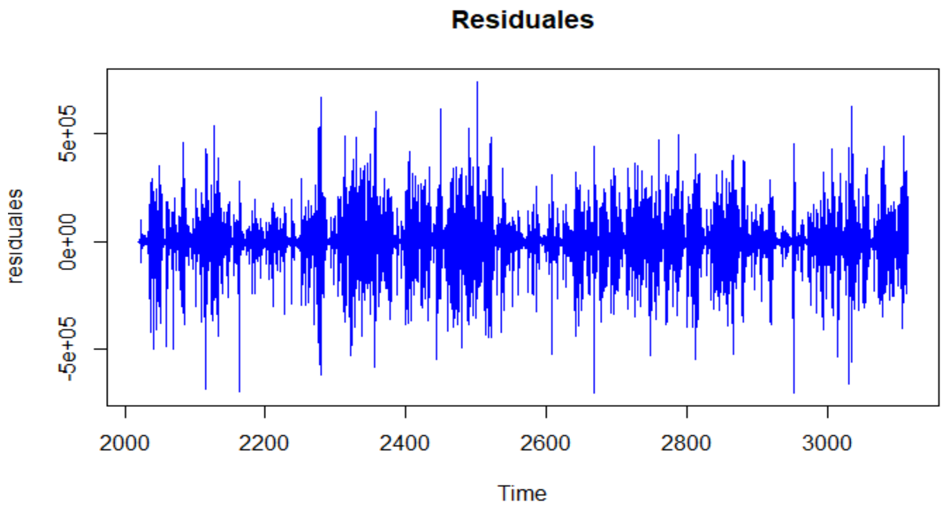


Figure 2. Residual plot in the SARIMA(5,0,0)(2,1,0)[24] model

The conjecture regarding figure 2 is to establish, by calculating the augmented Dickey-Fuller statistic = -26,123, lag order = 29, p-value = 0,01, whether the residuals are stationary. Therefore, this SARIMA(5,0,0)(2,1,0)$_{24}$ model represents a good structure that provides a good fit. Combined with the result of the Box-Ljung-Box test = 0,00094546, with a probability value (p-value) = 0,9755, this suggests that the residuals do not show significant serial autocorrelation and are therefore independent. Given that the p-value is much greater than the standard significance level (α=0,05), it can be concluded that the residuals obtained by the SARIMA model are appropriate for defining an optimal fit in terms of independence. These assertions are consolidated by a visual examination of the graphs shown in figure 3.

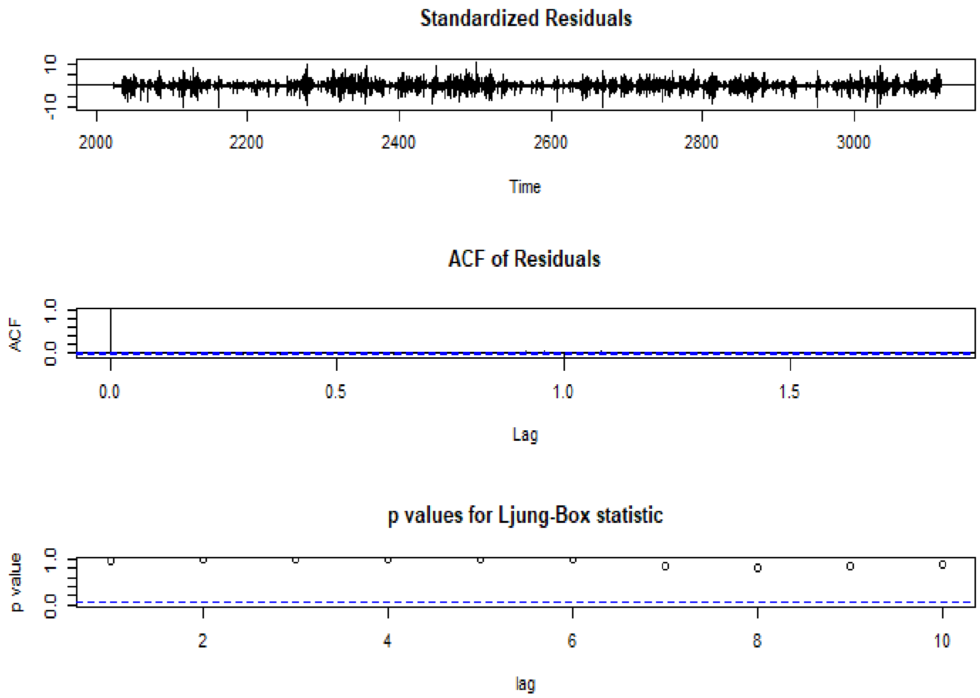


Figure 3. Diagnostic graphs for the residuals in the SARIMA(5,0,0)(2,1,0)[24] model

Figure 3 shows three graphical representations for the residual variable. First, the standardized residuals of the model are represented as a function of time centered around the mean value of zero and should be distributed randomly, with no apparent patterns. Second, the ACF correlogram of the residuals maintains lags within the confidence limits, indicating that the residuals are independent and do not exhibit statistically significant serial autocorrelation. Finally, the p-values of the Ljung-Box statistic for different lags are shown in a graphical representation, whose interpretation implies that all p-values are greater than the standard significance level (suggesting that there is no evidence of serial autocorrelation in the residuals produced by the SARIMA(5,0,0)(2,1,0)(24) model. In short, these findings on the general diagnosis indicate that the SARIMA model adjusted from the photovoltaic system power data appears to be adequate, since the residuals meet the assumptions of independence and absence of autocorrelation. This is a good indication that the model is well specified and can be used to make forecasts in future periods.

DISCUSSION

The contribution of this study has been remarkable in terms of findings, particularly in the comparative analysis of results obtained through numerical modeling, with the data recorded in the photovoltaic system in the Plaza del Duque de Béjar, Spain. It also provides a comparative efficiency analysis with the possible models obtained with a genetic algorithm implemented in specific packages within the RStudio 2024.09.1 program. The SARIMA model parameters were adjusted using a set of 26 304 data points related to the photovoltaic power records registered in the system from January 1, 2019, to December 31, 2021, considering the optimization of the parameters in the photovoltaic system to ensure increased energy efficiency. They improved overall efficiency.[33,34,35]

Therefore, an adequate estimation of SARIMA models allows the underlying mathematical structure to be used to explain the phenomenon under study and predict possible future scenarios for the power (W) generated by surface solar radiation (SSR) using a set of data collected from photovoltaic systems in the Plaza del Duque de Béjar, Spain. At this point, the most significant statistics are related to the mathematical structure of the SARIMA (5, 0, 0)(2, 1, 0)(24) model, which is why this variant was chosen after performing 35 iterations to form this integrated moving average autoregressive model with a marked regular and seasonal component. In a general context, for the final selection of the best structure, the adequacy of the identification and estimation of the aforementioned SARIMA mathematical equation must be evaluated by applying the Ljung-Box statistical test used to identify the most significant model[36] in terms of the statistical insignificance of the residual autocorrelations in the different time lags.

The SARIMA(5,0,0)(2,1,0)(24) model, identified using genetic algorithms, has demonstrated statistically robust performance in the training set, as evidenced by a MASE value of 0,7048786, indicating that its predictive capacity significantly exceeds that of a naïve model. This result validates the efficiency of the model in capturing the seasonal and stochastic structure of the solar power generation series. Furthermore, as it has the lowest error value among all the configurations evaluated by the genetic algorithm, it is established as the optimal modeling option for the study context.

This finding is consistent with previous research that has supported the use of SARIMA models in similar contexts. For example, [39] demonstrated that SARIMA models allow accurate forecasts of daily solar radiation, facilitating energy planning. Similarly, [40] applied SARIMA to hourly solar power data in Japan, achieving robust predictions in the presence of seasonality. Likewise, [41] highlighted that, although hybrid models offer marginal improvements in some cases, SARIMA models remain reliable tools, mainly due to their interpretable structure and low computational complexity.

Under this approach, it is worth reaffirming the usefulness of applying SARIMA models optimized using evolutionary algorithms in photovoltaic systems with extensive hourly records, such as the plant in Plaza del Duque de Béjar, Spain. This methodology not only improves the efficiency of parameter estimation but also provides a solid method on which to base strengthening forecasting, operational management, and energy planning processes in the context of the transition to renewable sources.

The requirement for an accurate forecast of photovoltaic solar energy generation within a short period is essential to ensure safe and economical operation,[37] as well as for energy management in the grid. This forecast should be made using a Seasonally Autoregressive Integrated Moving Average (SARIMA) model for forecasting various periods (resolution of 1 hour and 10 minutes) of solar photovoltaic generation.[29] While long- and medium-term forecasts maximize the benefits of photovoltaic power generation and its market penetration, short-term forecast horizons estimate the rates of increase in photovoltaic power.[38] According to Husein et al.[4], their contributions were decisive in demonstrating that forecasts of solar radiation and its impact on generation power in photovoltaic systems one day in advance can reduce the annual energy costs of microgrid operations in commercial buildings.

A viable alternative for future developments in this line of research is represented by the growing preference for deep learning approaches in time series prediction for solar power generation. This assertion has been supported by the findings of Kim et al.[3], who conducted a study on prediction accuracy for photovoltaic power generation using seven models in the cities of Ansan and Suwon. This study discovered a decline in seasonal and meteorological forecasts. Alternatively, hybrid deep learning models could be used for solar energy time series forecasting.[10]

CONCLUSIONS

SARIMA model estimation has proven to be an effective tool for short-term prediction of the power generated by photovoltaic systems, allowing their performance to be evaluated using statistical metrics applied to the training data set. In this study, the SARIMA(5,0,0)(2,1,0) model was identified as the most appropriate mathematical structure after multiple iterations using genetic algorithms, which facilitated efficient optimization of the adjustment process.

The model was validated using the Ljung-Box statistic, which is used to analyze the independence of the residuals. The result (p-value > 0,05) confirmed the absence of significant serial autocorrelation, thus supporting the consistency of the model and its ability to capture the seasonal and stochastic dynamics of the series correctly.

As a future projection, we propose exploring hybrid and advanced approaches based on deep learning, which allow addressing the nonlinear complexity inherent in photovoltaic generation time series. In particular, we suggest developing a polynomial neural network based on the GMDH (Group Method of Data Handling) method, capable of inductively optimizing the prediction of the power generated. This type of architecture is promising for scenarios where solar irradiation undergoes sudden and unpredictable changes, affecting the reliability and stability of the energy system.

BIBLIOGRAPHIC REFERENCES

1. Sobri S, Koohi-Kamali S, Rahim NA. Solar photovoltaic generation forecasting methods: A review. Energy Convers Manag. 2018; 156:459–97. http://dx.doi.org/10.1016/j.enconman.2017.11.019

2. Dada M, Popoola P. Recent advances in solar photovoltaic materials and systems for energy storage applications: a review. Beni-Suef Univ J Basic Appl Sci. 2023;12(1). http://dx.doi.org/10.1186/s43088-023-00405-5

3. Kim E, Akhtar MS, Yang O-B. Designing solar power generation output forecasting methods using time series algorithms: Global warming affected weather conditions. SSRN Electron J. 2022; http://dx.doi.org/10.2139/ssrn.4170513

4. Husein M, Chung I-Y. Day-ahead solar irradiance forecasting for microgrids using a long short-term memory recurrent neural network: A deep learning approach. Energies. 2019;12(10):1856. https://www.mdpi.com/1996-1073/12/10/1856

5. Aliberti A, Fucini D, Bottaccioli L, Macii E, Acquaviva A, Patti E. Comparative analysis of neural networks techniques to forecast global horizontal irradiance. IEEE Access. 2021;9:122829–46. http://dx.doi.org/10.1109/access.2021.3110167

6. Atique S, Noureen S, Roy V, Subburaj V, Bayne S, Macfie J. Forecasting of total daily solar energy generation using ARIMA: A case study. En: 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC). IEEE; 2019. p. 0114–9.

7. Runge J, Zmeureanu R. Forecasting energy use in buildings using artificial neural networks: A review. Energies. 2019;12(17):3254. http://dx.doi.org/10.3390/en12173254

8. Li L, Han C. ASARIMA: An adaptive harvested power prediction model for solar energy harvesting sensor networks. Electronics (Basel). 2022;11(18):2934. http://dx.doi.org/10.3390/electronics11182934

9. Qing X, Niu Y. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. Energy (Oxf). 2018;148:461–8. http://dx.doi.org/10.1016/j.energy.2018.01.177

10. Salman D, Direkoglu C, Kusaf M, Fahrioglu M. Hybrid deep learning models for time series forecasting of solar power. Neural Comput Appl. 2024;36(16):9095–112. http://dx.doi.org/10.1007/s00521-024-09558-5

11. Junhuathon N, Chayakulkheeree K. Comparative study of short-term photovoltaic power generation forecasting methods. En: 2021 International Conference on Power, Energy and Innovations (ICPEI). IEEE; 2021. p. 159–62.

12. Jiang Y, Zheng L, Ding X. Ultra-short-term prediction of photovoltaic output based on an LSTM-ARMA combined model driven by EEMD. J Renew Sustain Energy. 2021;13(4). http://dx.doi.org/10.1063/5.0056980

13. Aghmadi A, El Hani S, Mediouni H, Naseri N, El Issaoui F. Hybrid solar forecasting method based on empirical mode decomposition and Back Propagation Neural Network. E3S Web Conf. 2021;231:02001. http://dx.doi.org/10.1051/e3sconf/202123102001

14. Mughal SN, Sood YR, Jarial RK. Design and optimization of photovoltaic system with a week ahead power forecast using autoregressive artificial neural networks. Mater Today. 2022;52:834–41. http://dx.doi.org/10.1016/j.matpr.2021.10.223

15. Rogier JK, Mohamudally N. Forecasting photovoltaic power generation via an IoT network using nonlinear autoregressive neural network. Procedia Comput Sci. 2019;151:643–50. http://dx.doi.org/10.1016/j.procs.2019.04.086

16. Sultan Mohd MR, Johari J, Ruslan FA, Abdul Razak N, Ahmad S, Mohd Shah AS. Analysis on parameter effect for solar radiation prediction modeling using NNARX. En: 2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS). IEEE; 2021. p. 69–74.

17. Boubaker S, Kamel S, Kolsi L, Kahouli O. Forecasting of one-day-ahead global horizontal irradiation using block-oriented models combined with a swarm intelligence approach. Nat Resour Res. 2021;30(1):1–26. http://dx.doi.org/10.1007/s11053-020-09761-w

18. Zou L, Munir MS, Kim K, Hong CS. Day-ahead energy sharing schedule for the P2P prosumer community using LSTM and swarm intelligence. En: 2020 International Conference on Information Networking (ICOIN). IEEE; 2020. p. 396–401.

19. Benti NE, Chaka MD, Semie AG. Forecasting renewable energy generation with machine learning and deep learning: Current advances and future prospects. Sustainability. 2023;15(9):7087. http://dx.doi.org/10.3390/su15097087

20. Basmadjian R, Shaafieyoun A, Julka S. Day-ahead forecasting of the percentage of renewables based on time-series statistical methods. Energies. 2021;14(21):7443. http://dx.doi.org/10.3390/en14217443

21. Natarajan VA, Karatampati P. Survey on renewable energy forecasting using different techniques. En: 2019 2nd International Conference on Power and Embedded Drive Control (ICPEDC). IEEE; 2019. p. 349–54.

22. Singh B, Pozo D. A guide to solar power forecasting using ARMA models. En: 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe). IEEE; 2019. p. 1–4.

23. Fara L, Diaconu A, Craciunescu D, Fara S. Forecasting of energy production for photovoltaic systems based on ARIMA and ANN advanced models. Int J Photoenergy. 2021;2021:1–19. http://dx.doi.org/10.1155/2021/6777488

24. Marikkar U, Hassan ASJ, Maithripala MS, Godaliyadda RI, Ekanayake PB, Ekanayake JB. Modified Auto Regressive technique for univariate time series prediction of solar irradiance [Internet]. arXiv [eess.SP]. 2020. Disponible en: http://arxiv.org/abs/2012.03215

25. Reikard G, Hansen C. Forecasting solar irradiance at short horizons: Frequency and time domain models. Renew Energy. 2019;135:1270–90. http://dx.doi.org/10.1016/j.renene.2018.08.081

26. Seyedmahmoudian M, Jamei E, Thirunavukkarasu G, Soon T, Mortimer M, Horan B, et al. Short-term forecasting of the output power of a building-integrated photovoltaic system using a metaheuristic approach. Energies. 2018;11(5):1260. http://dx.doi.org/10.3390/en11051260

27. Das UK, Tey KS, Seyedmahmoudian M, Mekhilef S, Idris MYI, Van Deventer W, et al. Forecasting of photovoltaic power generation and model optimization: A review. Renew Sustain Energy Rev. 2018;81:912–28. http://dx.doi.org/10.1016/j.rser.2017.08.017

28. European Commission. PVGIS-SARAH3: Photovoltaic Geographical Information System. Obtenido de EU Science Hub https://ec.europa.eu/jrc/en/pvgis. 2024 may.

29. Kushwaha V, Pindoriya NM. Very short-term solar PV generation forecast using SARIMA model: A case study. En: 2017 7th International Conference on Power Systems (ICPS). IEEE; 2017. p. 430–5.

30. Kushwaha V, Pindoriya NM. A SARIMA-RVFL hybrid model assisted by wavelet decomposition for very short-term solar PV power generation forecast. Renew Energy. 2019;140:124–39. http://dx.doi.org/10.1016/j.renene.2019.03.020

31. Rajagukguk RA, Ramadhan RAA, Lee H-J. A review on deep learning models for forecasting time series data of solar irradiance and photovoltaic power. Energies. 2020;13(24):6623. http://dx.doi.org/10.3390/en13246623

32. Zhang X, Wu X, Zhu G, Lu X, Wang K. A seasonal ARIMA model based on the gravitational search algorithm (GSA) for runoff prediction. Water Sci Technol Water Supply. 2022;22(8):6959–77. http://dx.doi.org/10.2166/ws.2022.263

33. Moeeni H, Bonakdari H, Ebtehaj I. Monthly reservoir inflow forecasting using a new hybrid SARIMA genetic programming approach. J Earth Syst Sci. 2017;126(2). http://dx.doi.org/10.1007/s12040-017-0798-y

34. Fashae OA, Olusola AO, Ndubuisi I, Udomboso CG. Comparing ANN and ARIMA model in predicting the discharge of River Opeki from 2010 to 2020. River Res Appl. 2019;35(2):169–77. http://dx.doi.org/10.1002/rra.3391

35. Harrou F, Taghezouit B, Sun Y. Robust and flexible strategy for fault detection in grid-connected photovoltaic systems. Energy Convers Manag. 2019;180:1153–66. http://dx.doi.org/10.1016/j.enconman.2018.11.022

36. Taghezouit B, Harrou F, Larbes C, Sun Y, Semaoui S, Arab A, et al. Intelligent monitoring of photovoltaic systems via simplicial empirical models and performance loss rate evaluation under LabVIEW: A case study. Energies. 2022;15(21):7955. http://dx.doi.org/10.3390/en15217955

37. Yesildal F, Ozakin AN, Yakut K. Optimization of operational parameters for a photovoltaic panel cooled by spray cooling. Eng Sci Technol Int J. 2022;25(100983):100983. http://dx.doi.org/10.1016/j.jestch.2021.04.002

38. Belghiti H, Kandoussi K, Chellakhi A, Mchaouar Y, El Otmani R, Sadek EM. Performance optimization of photovoltaic system under real climatic conditions using a novel MPPT approach. Energy Sources Recovery Util Environ Eff. 2024;46(1):2474–92.

39. Kumar, M., & Kumar, Y. Solar radiation forecasting using SARIMA model for Patiala city, Punjab, India. Materials Today: Proceedings. 2020; 33, 3739-3744. https://doi.org/10.1016/j.matpr.2020.08.413

40. Yona, A., Senjyu, T., Saber, A. Y., Urasaki, N., & Funabashi, T. Application of recurrent neural network to short-term-ahead generating power forecasting for photovoltaic system. Energy. 2013; 30(11-12), 2191-2204. https://doi.org/10.1016/j.energy.2004.03.001

41. Benmouiza, K., & Cheknane, A. Forecasting hourly global solar radiation using hybrid k-means and nonlinear autoregressive neural network models. Energy Conversion and Management. 2013; 75, 561-569. https://doi.org/10.1016/j.enconman.2013.08.027

FINANCING

None.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Jesús Rodríguez.
Data curation: Christian Reyes.
Formal analysis: Diana Campaña.
Research: Elsa Basantes.
Methodology: Juan Espinoza.
Project management: Sandra Chasiluisa.
Resources: Jesús Rodríguez.
Software: Christian Reyes.
Supervision: Diana Campaña.
Validation: Elsa Basantes.
Visualization: Juan Espinoza.
Writing - original draft: Sandra Chasiluisa.
Writing - review and editing: Jesús Rodríguez.