







Solar Radiation Prediction Using Machine Learning Techniques

Luis Alejandro Caycedo Villalobos^(✉) ,
Richard Alexander Cortázar Forero^(✉) , Pedro Miguel Cano Perdomo^(✉) ,
and José John Fredy González Veloza^(✉) 

Fundación Universitaria Los Libertadores, Bogotá, Colombia
{lacaicedov,racortazarf,pmcanop,jjgonzalezv02}@libertadores.edu.co

Abstract. The proposal of a solar radiation estimation model using Machine Learning is submit, the processing of meteorological data measured by satellite and data measured on the land is made. The model uses two solutions using an artificial neural network and robust linear regression the climatic variables used as input to the model are solar radiation, temperature and clarity index, all get from satellite data. The main aim of this work is to propose a model that allows using the satellite data to get an estimate of the behavior of the solar resource on the ground, reducing the error between the satellite data and the data measured on the ground. The results of the model got by training an artificial neural network with hidden layers are submit, here the normal distributions of the data reported by the satellite and the data got by the proposed model are submit. In addition, the results of the daily average got by the model and the daily average values measured on land are submit. I conclude it by proposing a second estimation model using robust linear regression. A proposed model adjusted to the assumptions made during the regression process and acceptable results to those got by the satellite and reported by other works are got.

Keywords: Estimate · Solar radiation · Machine learning

1 Introduction

Solar radiation is the energy emitted by the sun. It propagates as electromagnetic waves through space and reaches the earth's surface after passing through the atmosphere; consider the most abundant and important source of energy for life on the Earth causes natural phenomena and chemical reactions for the growth of plants and animals. The energy captured by the earth from the sun is the primary source of renewable energy that we have available, the amount of energy that is captured annually is approximately 1.6 million KWh, which is much more than what it consumed worldwide [1]. In order to reduce dependence on electrical energy produced by non-renewable sources, the use of solar energy has gained

Supported by organization Fundación Universitaria Los Libertadores.

© Springer Nature Switzerland AG 2021

H. Florez and M. F. Pollo-Cattaneo (Eds.): ICAI 2021, CCIS 1455, pp. 68–81, 2021.

https://doi.org/10.1007/978-3-030-89654-6_6

great importance in recent years, proof of this is the increase in photovoltaic - PV solar systems as part of the generation of electrical energy, according to Enerdata in 2019, the generation of electrical energy from solar energy had an increase of 24% compared to 2018 [2]. With the current high demand for electricity, there has been great interest worldwide to integrate solar energy systems into the electricity grid, in order to improve the quality supplied in some places and reduce the costs associated with dependence on the grid conventional electric. However, implementing these systems presents an obstacle. The dimensioning of a PV system is closely related to the climatic conditions [3], for which it required reliable data sources that allow making a forecast of the solar resource.

Thanks to the Institute of Hydrology, Meteorology and Environmental Studies (IDEAM), there is irradiance data measured in different meteorological stations in Colombia, but not just any user has access to this data, so it is necessary to have an additional source that allows get reliable data that provide information for estimating the values on land [5]. An alternative is the access that users have to the data from NASA's PowerViewer database, but in order for it to be used, the error in the measurement of solar radiation taken by satellite and recorded in the database must be reduce data [6]. The present work proposes two linear regression models of the daily average radiation reported by the satellite, getting a normal distribution model that predicts the values of daily radiation with its maximums and minimums, taking as a reference the values recorded by IDEAM in the different stations weather of Colombia. For this reason, it proposed two models that can be use as a planning tool for the design of PV systems, for which it made use of the Python and R programming language.?

This paper is organized [14] as follows: Sect. 2 related work; Sect. 3 Estimation model using Machine learning; Sect. 4 Estimation model results using Machine learning. Conclusions are presented in the last section

2 Related Work

Solar radiation prediction provides the ability to optimize the operation of solar-powered systems, improve quality, optimize production costs, and estimate the amount of energy that the system can deliver. Several studies similar to those conducted in this work have conducted previously.

In [4] the author used meteorological data stored in GRIB and netCDF files to generate a predictive model in Python using the sklearn library, specifically the linear regression method between different irradiance data used, in which a coefficient of determination of 0.82 and a mean absolute error of 4.16.

In [5] a statistical model generated for the estimation of solar radiation applying alternate meteorological data. IDEAM provided the data used for the study, from a station that has the measurement of solar radiation, temperature, relative humidity and hour of sunshine. Angstrom-Preccott and Gueymard statistical regressions used, in which it evidenced little relationship between the variables, the best fit achieved was through the linear regression between relative humidity and solar radiation with a coefficient of determination of 11.14%.

In [6] the author proposes a hybrid predictor model implemented in classification-grouping stages using Fuzzy Logic and for the estimation of Neural Networks and State Vector Machines, the main aim was to create a model that considered the conditions geographical areas of Colombia. The input variables to the model included temperature, wind speed, clarity index, total precipitation, relative humidity, and atmospheric pressure, all got from NASA's PowerViewer database. The correlation coefficients got by this model are close to 1 and the mean square error (RMSE) is between 0.04 and 0.09, showing that the model shows an acceptable performance in all the cities evaluated.

In [7] a model proposed to predict hourly solar radiation using linear regression and neural networks. Solar radiation got with mathematical models, temperature, atmospheric pressure and relative humidity used as input variables to the model. They compared the applied models to determine which one fitted the best with data from five meteorological stations in Tucumán, Argentina. The results got show that an average error of 11% achieved with linear regression and 7.84% with neural networks.

3 Estimation Model Using Machine Learning

The present work proposes a model that allows reducing the error in the measurement of solar radiation taken from the PowerViewer database, taking as a reference the data recorded by IDEAM in its meteorological stations near the geographic region of interest, getting from this proposal a normal distribution model for solar radiation and its standard deviation on a particular day and a linear regression model for estimating the daily mean value. The meteorological data used to generate the model proposed in this work were get from the database of the Institute of Hydrology, Meteorology and Environmental Studies (IDEAM) and from the NASA PowerViewer database, which are contained in CSV files and netCDF, respectively. The data from measurements of solar radiation, clarity index and temperature in the meteorological stations in the principal cities of Colombia is to have as a reference. The time window used in this job is between December 4, 2014 to August 31, 2020. It developed the proposed models from two different Machine Learning concepts and integrate their results to propose the estimation of solar radiation on the ground from the data recorded by the satellite and available in the PowerViewer database. The first model seeks to get through artificial neural networks - ANN the estimation of the normal distribution of solar radiation on the ground, which allows having average, maximum and minimum values during a day. The second model seeks to estimate through robust linear regression the daily average value of solar radiation on the ground from the data recorded by the satellite. The data of measurements on the ground provided by the meteorology station at the Jose Maria Cordova airport in the city of Medellin - Colombia by IDEAM, are contain in CSV files and through the PANDAS python library it is possible to visualize, analyze and process sets of data that are known as data frame. The data frame is reporte in hourly mode and when preprocessed it to have to daily average values to be compared with

the satellite data file, the records in which the equipment did not capture information are eliminate, the rows with records equal to 0, values that correspond to the hours where there is no incident solar radiation on the equipment; It filtered data using a moving average filter to smooth fluctuations produced by the data recording device. The data available in the PowerViewer database, the official page of NASA's Powerviewer and consulted from the page <http://power.larc.nasa.gov/data-Access-viewer/>, are contain in a netCDF file that contains the metadata with all the meteorological variables registered, in this case the python netCDF4 library is used to open and read this dataset. The database built with data from the two sources, IDEAM and PowerViewer, comprises 1638 records reported between 2014 and 2020. It described the fields for each record in Table 1. They performed the analysis of the behavior of each variable using the R program. The graphical results are presented in Figs. 1, 2 and 3. Figure 1 shows the extreme, atypical data with left bias affecting the measures of central tendency. In Fig. 2 the atypical data and left bias observed for the variables Sat and RH2M, the variables of precipitation (PRECTOT) and percentage of error between the measurement on the ground and the satellite, extreme values observed that skews the data to the right.

Table 1. Database record fields

Variable	Description	Units
Data	average daily radiation measured in surface	W/m ²
MM	moving average applied to data	w/m ²
Daily	daily average radiation measured on surface	KW/m ²
Data.filtered_mm	moving average of Data	KW/m ²
Sat	average radiation reported by the satellite	KW/m ²
error_sat	error percentage between Sat and Daily	
Prectot	precipitación	mm
RH2M	Relative humidity at 2 m height	
T2M_range	temperature range at 2 m height	°C
Ws50M_range	wind speed range at 50 m	m/s
KT	lightness or insolation index	

In Fig. 3 the data corresponding to the temperature range with the least amount of atypical data are to observe, the variable range of wind speeds presents the highest atypical data and for the variable of the clarity index the extreme, atypical data they are minimal. The correlations between the variables described in Table 1 are present in Fig. 4, there it can be observe very high correlations between several variables, which show a great tendency to linearity between them, without However, it must be clarified that the perfect correlations between variables are present because there are redundant data in the database that present the same variables but in different units. As of the variables

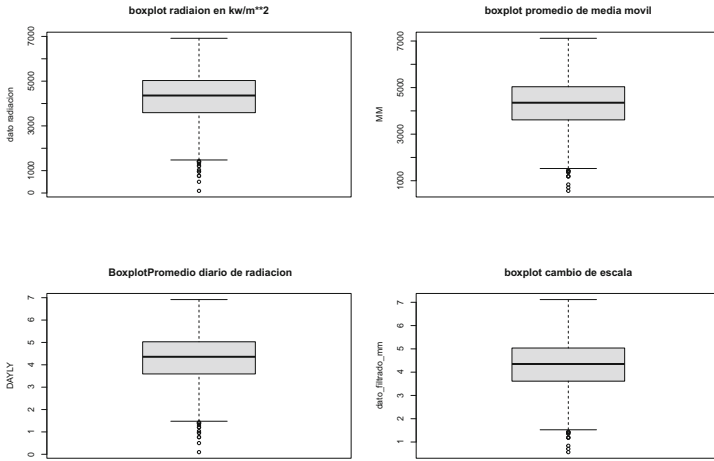


Fig. 1. Ground radiation analysis, moving average, daily ground radiation average, scale change

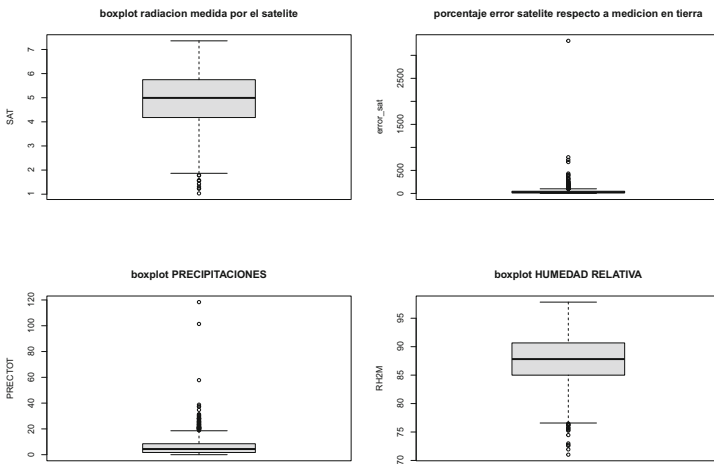


Fig. 2. Analysis of mean radiation by satellite, satellite error with respect to measurement on the ground, precipitation, relative humidity

DATA-DAILY, *MM-Filtereddata*, and there is a strong correlation between the variables *SAT-KT*, although in this case not It happens for the same reason as in the previous clarification. These variables are closely related and there is a strong tendency towards linearity between them. Also, there is a great correlation and therefore a linear trend with the variables *RH2M-T2MRANGE*, with the exception that it is an inverse linear trend. In [6] get evidenced that solar radiation presents a high correlation with other climatic variables, especially with temperature and clarity index, for which the input variables in the

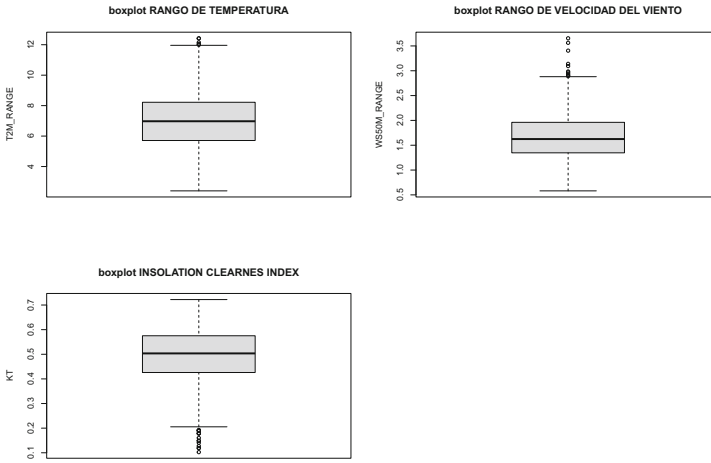


Fig. 3. Analysis of temperature range, wind speed range, clarity index

proposed models will be together with the solar radiation taken by the satellite. The prediction model presented in this work proposes the use of an artificial neural network - ANN, in problems where it is required making a prediction in time series using climatic variables as input to the model, it has been the ANN which provides a better response compared to other methods [12], in [13] it developed a model for forecasting the price of electrical energy by using ANN with a multilayer perceptron configuration. This architecture has proven to be one of the most useful to solve this type of problem. According to the information consulted in [8], the architecture most used in solar energy prediction applications is the multilayer perceptron, which is the most popular non-linear ANN architecture used to solve applied science problems. For this reason, the model selected for estimating solar radiation was an ANN with simple multilayer perceptron architecture [12,13]. The structure of the proposed ANN corresponds to: a. input layer: different climatic variables, b. hidden layer: contains the activation function, c. output layer: target variable. It represented the described structure in Fig. 5. According to the information consulted in [3], it is shown that solar radiation presents a high correlation in the data recorded by the satellite for the temperature variables, the clarity index, for which these climatic variables used together with the solar radiation taken by the satellite as the predictor characteristics represented as $[X]$ of the ANN. The target variable represented as $[Y]$ is the average solar radiation measured on the ground. The next part of the process comprises creating data sets, which correspond to a training set, equivalent to 70% and a test set equivalent to 30% of the selected data.

To estimate the average of the solar radiation on the surface, a Machine Learning model proposed with the robust linear regression technique (rlm), because with the multiple regression technique (lm), when doing the analysis for the residuals, the model did not comply with the assumptions of normality,

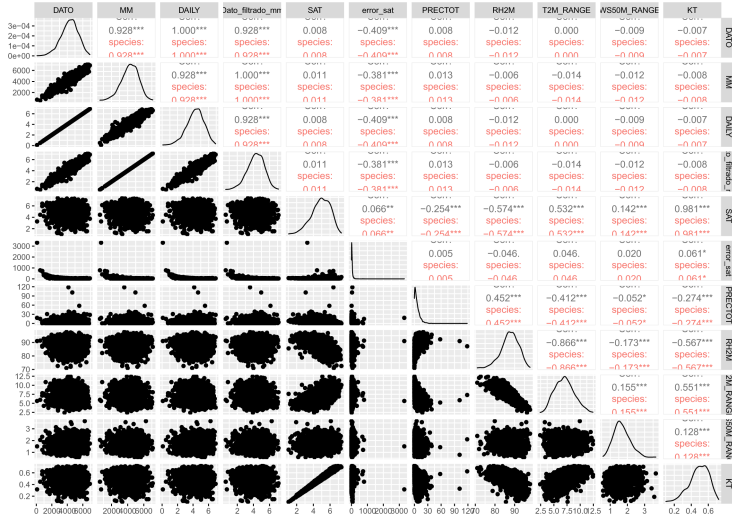


Fig. 4. Scatter diagram and correlation of the variables used

homoscedasticity, independence of residuals and multicollinearity. For the model (*lm*) it got that the significant variables for the regression were five: satellite radiation (SAT), percentage error of the solar radiation measurement between the station and satellite measurement (*errorsat*), relative humidity (RH2M), temperature range (*T2MRANGE*) and lightness index (KT). Although the (*lm*) towards a good forecast for the estimated radiation values with mean square errors (RSME) that oscillated between 1.1 KW/m² and 0.87 KW/m² as Outliers were being treated, with the particularity that each time the RSME improved,

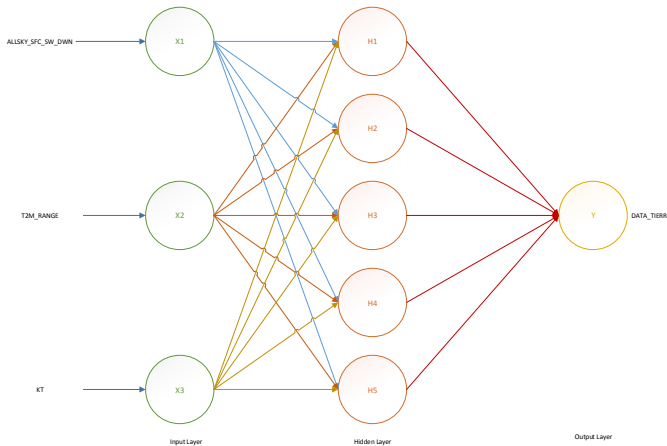


Fig. 5. Proposed multilayer perceptron neural network structure

the problem of non-compliance with assumptions became more acute, especially the assumption of normality. To correct these problems, it made a modification to the variable (*errorsat*), applying the logarithm function, since this variable is the one with the greatest bias because of the maximum extreme data it has. It also made a transformation to the response variable (DAILY), with the boxcox function, raising the variable to the exponent (1, 4). These transformations to the variables evidently improved the fulfillment of the assumptions, although the normality test continued to fail. As an additional procedure trying to make adjustments to the model, the variable (*errorsat*) eliminated, and the variable (DAILY) left with the exponent suggested by the boxcox function. This change guaranteed the fulfillment of the assumptions, but seriously affected the predictions made by the model, and increased my RSME from 0.87 KW/m² to 1.7 KW/m², this increase in RSME It attributed to the fact that for the multiple regression model, the discarded variable had a significance greater than 95%. After doing all these tests, we opted for the development of a robust regression model, because this type of regression is less affected by extreme values. For the creation of the robust regression model, the same variables of the multiple linear regression used and besides these variables included: average moving average of radiation (MM), precipitation (PRECTOT) and range of wind speeds (*WS50MRANGE*). considering the T tests, a change made to the intercept because it is not significant, I also eliminate the (rlm), some variables that can affect the model because of multicollinearity problems, because they present the same information but in different units, This is the case of the variable: solar radiation (DATA), which contains the same information as the response variable (DAILY), except that they have different units. A similar situation occurs with the variable moving average of radiation (MM) and filtered data of moving average of radiation (*Data_filtered_mm*). For this reason, the variables MM, SAT, *errorsat*, PRECTOT, RH2M, *T2MRANGE*, *WS50MRANGE*, KT taken. In the residuals histogram of (rlm), Fig. 6, a normal trend observed, which shows signs of normality of the residuals, Fig. 7, the same observed in Fig. 8 normal QQ with confidence intervals. According to the tests of: Kolmogorov-Smirnov, shapiro-Wilk, Ks.test the p-value = 0.2655 with a 95% confidence there is not enough statistical evidence to reject the null hypothesis in favor of the alternative hypothesis.

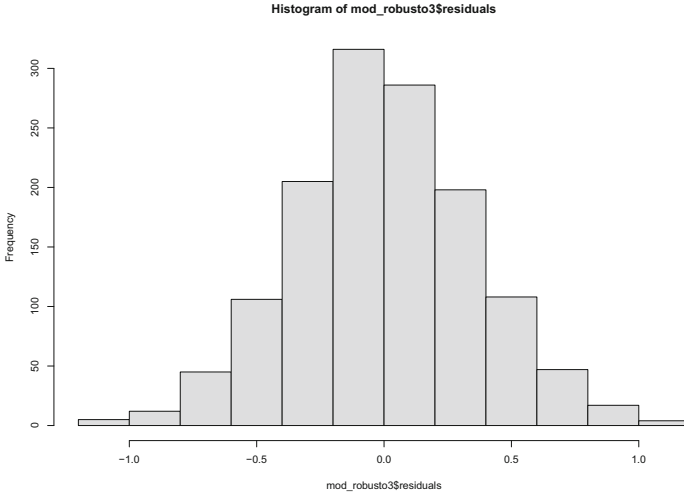


Fig. 6. Histogram of residuals for the robust linear regression model

The proposed model (rlm) complies with the assumption of normality, if the robust regression model (rlm) fulfills the assumption of normality, it shows that linear regression is a good choice as a model to predict solar radiation on the surface using environmental variables recorded by meteorological stations. To validate the assumption of homoscedasticity, the `ncvTest` (robust mod3) of the model performed, getting a $p\text{-value} = 0.14579$, with a 95% confidence, there is not enough statistical evidence to reject the null hypothesis in favor of the alter-

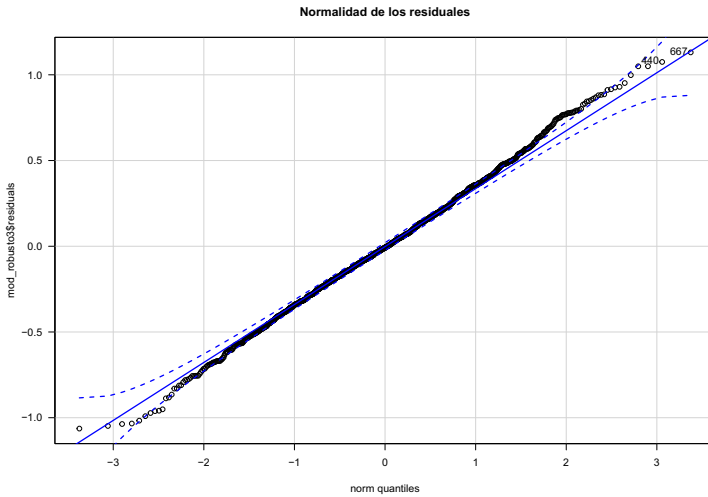


Fig. 7. Normality of the residuals of the proposed robust linear regression model

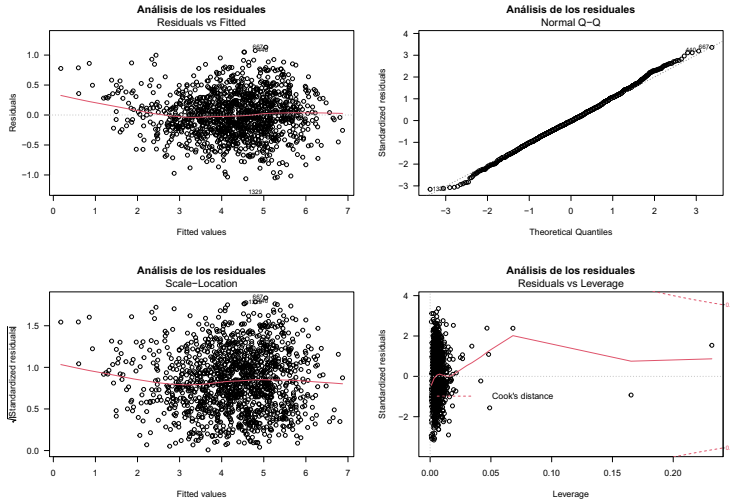


Fig. 8. Analysis of residuals for the proposed robust linear regression model. to. Residual Vs Fitted, b. normal Q-Q, c. Scale Vs Location, d. Residual Vs Leverage.

native hypothesis, the proposed model complies with the assumption of constant variance. The model is homoelastic, which guarantees that the regression is adequate for forecasting. It is important that someone distribute constantly the variance in the errors throughout all the observations. This characteristic in a regression increases the efficiency and reliability of the model. To validate the independence assumption, the test performed with the Durbin Watson statistic, the DW statistic is close to two ($DW = 2.365$) which is a good indicator and there is a $p\text{-value} = 1$, with a confidence of 95%, there is not enough statistical evidence to reject the null hypothesis in favor of the alternative hypothesis. The proposed model fulfills the assumption of independence. Failure to comply with the autocorrelation error test for the model can explained because of the phenomenon of solar radiation and the other associated variables, it has a space-time domain, which was not used in the analysis and subsequent estimation of the values of daily solar radiation. Using the temporal variable recommended for later work, in order to improve the estimation and reliability of the model. Because of the randomness of this natural phenomenon, there is a spatiotemporal correlation, which gives indications that from data analytics without considering physical linear models, it can become more efficient to forecast a model in time series.

4 Estimation Model Results Using Machine Learning

The results got from the estimation model using ANN with multilayer perceptron configuration between the target variable and the predictor variables carried out for one year, divided into sub-periods of four months, gave the following results,

which provide the information to optimal interpretation the results delivered by the model. The first result delivered by the model corresponds to the estimated values of daily solar radiation for the selected period, in the Fig. 9 are the target values compared with the values predicted by the model. The comparison of the error between the satellite data and ground data with the error between the predicted data and the ground data, Fig. 10. Table 2 shows the results got for each analysis period together with the reference values measured by satellite and land measurements. Here, it shown that the model has an acceptable behavior, since it replicates the variability of solar radiation over time and that it maintains its performance when evaluated at different times of the year. Between the data got by the model and the reference data, an average difference of 0.1916 kW/m^2 got, compared to the difference with the satellite data, which presents an average difference of 0.5879 kW/m^2 . It showed the average error measurement for each period in Table 3. The percentage of error between the satellite-measured and land-based values shown along with the estimated values and the land-measured values.

It presented the results of the model proposed by robust linear regression in Fig. 11. The RSME value for the estimates made by the model is 0.3927409 KW/m^2 , for this case, this value shows that the predicted value is around more or less 0.4 KW/m^2 of the real value measured by surface stations. Based on the results presented, it was possible to show that the robust regression complies with the assumptions of normality, homoscedasticity and independence, and cannot comply with the assumption of autocorrelation errors. This explained because of the phenomenon that has a space-time domain, which was not used in the anal-

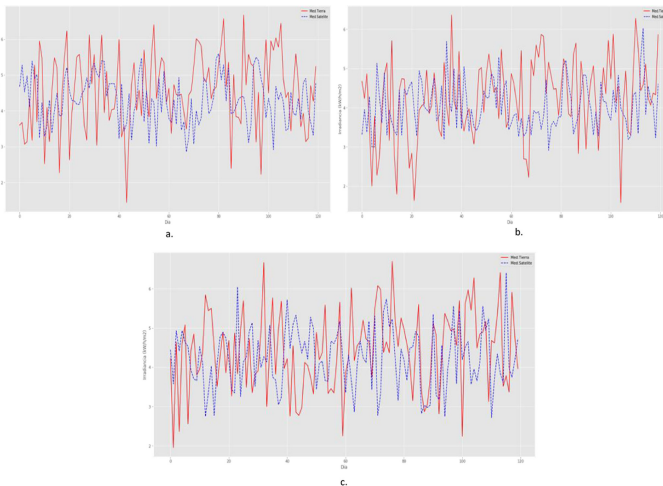


Fig. 9. ANN Model, target values compared to model predicted values separated in 4-month periods over a year. a) Radiation data for four-month period 1- Julian day 1–120. b) Radiation data four-month period 2- Julian day 120–240. c) Radiation data four-month period 3- Julian day 240–360

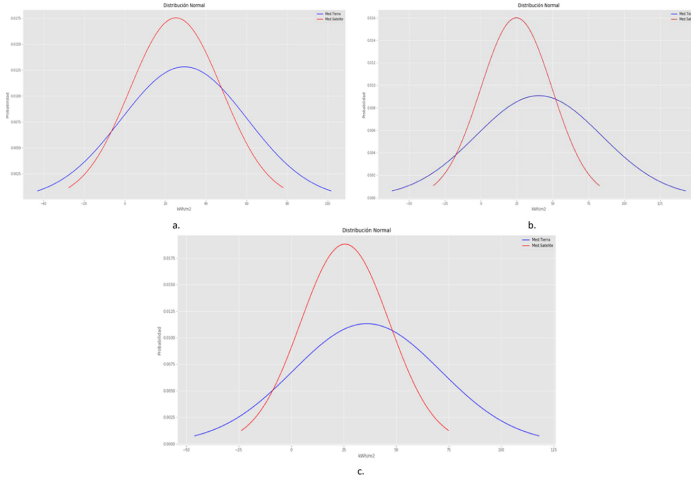


Fig. 10. ANN Model, comparison of the normal distribution of the estimated data, separated into 4-month periods over a year. a) Comparison of the normal distribution for four months 1- Julian day 1–120. b) Comparison of the normal distribution four-month period 2- Julian day 120–240. c) Comparison of the normal distribution four-month period 3- Julian day 240–360.

Table 2. Results of the estimation model - ANN in the evaluated time periods

	First four months	Second four months	Third four months
Measurement	Average in kW/m ²	Average in KW/m ²	Average in KW/mm ²
Land	4.572559	4.247772	4.434122
Satellite	4.640138	5.305578	5.063395
Estimation - RNA	4.297978	4.035805	4.255731
Difference (earth-satellite)	0.067579	1.057806	0.629273
Difference (earth-estimate RNA)	0.274581	0.211967	0.178391

Table 3. Comparative average error of the ANN model results in the evaluated time periods

	First four months	Second four months	Third four months
Comparative	Average error (%)	Average error (%)	Average error (%)
Satellite - land	29.268876	40.742404	35.88887
estimate - land	25.389562	25.141444	25.619838
Decrease	3.879314	15.60096	10.269049

ysis and subsequent estimation of the values of daily solar radiation. Besides this peculiarity, the condition of apparent correlation between the variables clarity index (KT) and satellite radiation (SAT) can be associated. Using the temporal variable recommended for later studies, in order to improve the estimation and reliability of the model, since the randomness of this phenomenon causes a spatio-temporal correlation to exist.

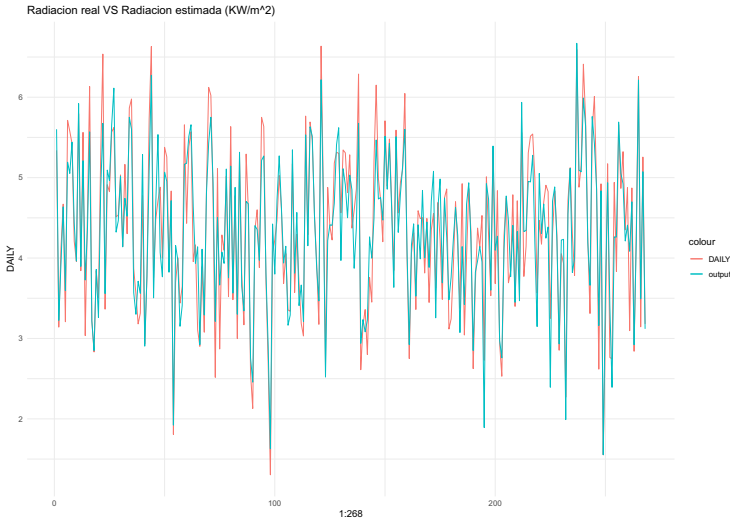


Fig. 11. Result of the robust linear regression model Vs real values of solar radiation for one year. In red the actual data of the average radiation measured by the meteorological station are represented. The data estimated by the proposed robust linear regression model are represented in Blue (Color figure online)

5 Conclusions

The work presented, proposed a model for estimating direct solar radiation using data consulted from the NASA PowerViewer database, getting an error rate between the proposed model and the values measured on the ground of 25.61%. Which is less than the 40.74% taken by the satellite. The data used for the development of the model selected according to the correlation they present with solar radiation, these data correspond to the clarity index and temperature, provided by the PowerViewer database, the best fit achieved allowed to have an Average error between the aim variable and the estimated data of 25.38%, regarding the error of the satellite data corresponding to 40.74%, which shows that an average decrease of 9.91% achieved. The model captures the variability of the solar resource and maintains a trend towards the average value of the target data, getting an average error in the estimate of $\pm 216.6 \text{ W/m}^2$ that is lower than that got by the satellite corresponding to 581.3 W/m^2 , which is considered

an improvement in the value of solar radiation calculated using the proposed model. The RSME for the estimates made by the robust regression model is 0.3632, which is approximately equivalent to an error value in the estimated radiation of 0.6 KW/m², compared to other works that do the validation of the model with the same metric [11]. Therefore, this confirms that the choice of the robust regression model to estimate the solar radiation on the surface was a wise decision.

References

1. Secretaria de Energía Argentina (2008) - Coordinación de Energías Renovables - Dirección Nacional de Promoción - Subsecretaría de Energía Eléctrica. Energías Renovables 2008 - Energía Solar
2. Enerdata - Estadísticas Energéticas Mundiales. (11 de Abril del 2021). Anuario estadístico mundial 2020. <https://datos.enerdata.net/energias-renovables/eolica-solar-produccion.html>
3. Obando-Paredes, E., Vargas-Cañas, R.: Desempeño de un sistema fotovoltaico autónomo frente a condiciones medioambientales de una región en particular, Rev. la Acad. Colombia. Ciencias Exactas, Físicas y Nat. **40**(154), 27–33 (2016)
4. Bella-Santos, J.: Herramientas python para la predicción de energías renovables (trabajo de pregrado). Universidad Autónoma de Madrid, Madrid, España (2018)
5. Vélez-Pereira, A., Vergara, E., Barraza, W., Agudelo, D.: Determinación de un modelo paramétrico para estimar la radiación solar. Ingenium **7**(18), 11–17 (2013)
6. Obando Paredes, E. Modelo de pronóstico de radiación solar basado en Machine Learning. 2018
7. Jimenez, V., Will, A., Rodriguez, S. Estimación de Radiación Solar Horaria Utilizando Modelos Empíricos y Redes Neuronales Artificiales. Ciencia y Tecnología. 1. 10.18682/cyt.v1i17.608. 2017
8. Tymvios, Filippou S, Michaelides. Silas, Chr. Skouteli, Chara S. Estimation of Surface Solar Radiation with Artificial Neural Networks. Modeling Solar Radiation at the Earth's Surface. Springer - Verlag Berlin Heidelberg. 2008
9. Boland, J.: Time Series Modelling of Solar Radiation. Springer - Verlag Berlin Heidelberg, Modeling Solar Radiation at the Earth's Surface (2008)
10. Mora-López, L.: A new Procedure to Generate Solar RAdiation Time Series from Machine Learning Theory. Modeling Solar Radiation at the Earth's Surface. Springer, Heidelberg (2008)
11. Ordoñez-Palacios, L.-E., León-Vargas, D.-A., Bucheli-Guerrero, V.-A., Ordoñez-Eraso, H.-A.: Solar Radiation Prediction on Photovoltaic Systems Using Machine Learning Techniques. Revista Facultad de Ingeniería, vol. 29 (54), e11751 (2020). <https://doi.org/10.19053/01211129.v29.n54.2020.11751>
12. Brentan, B., et al.: Hourly water demand forecasting using nonlinear autoregressive with exogenous artificial neural networks -narx (2015)
13. Villada, F., Cadavid, D.R., Molina, J.D.: Electricity price forecasting using artificial neural networks. Revista Facultad De Ingeniería Universidad De Antioquia **44**, 111–118 (2014)
14. Misra, S.: A step by step guide for choosing project topics and writing research papers in ICT Related Disciplines. In: Information and Communication Technology and Applications, Third International Conference, ICTA 2020, Minna, Nigeria, 24–27 November 2020