Solar Energy and Data Science: a prediction study for Manaus and the Amazon Basin

MARQUES, A. L.F.[*]; CORREA, P.L.P. [*] ; VALENCIA, F. [*]; TEIXEIRA, M.J [†];
[*] Departamento de Engenharia da Computação – Escola Politécnica - Universidade de São Paulo
[†]Instituto de Física – Universidade de São Paulo
4004andre3003@gmail.com

I. INTRODUCTION

The renewable energies have confirmed their potential to reshape the energy matrix of several countries, in the last two decades, reinforcing the set of actions to better protect the environment. For instance, the solar and wind sources of energy production can dim the production and release of pollutants associated to electricity production. Unfortunately, the atmospheric temperature rise has been linked to the increase of the gas emissions, such as Carbon Dioxide (CO_2) and Methane (CH_4), from manifold sources, such as the thermal power stations for electricity generation.

The Amazon basin has caught special importance for Brazil and abroad, and more recently, the use of satellites, fixed instrumented stations and airborne surveys has provided data focused on studies of the environment impact [1].

This work deals with the assessment of the solar irradiation on the Manaus city, the largest city in the Western Amazon region, using Data Science tools, in a way to help the evaluation of the renewable energy in that region.

II. MATERIAL & METHODS

The sun irradiation in Brazil has a non-homogeneous distribution, being measured as W.h/m^2 per day. Figure 1 shows a map of the sun incidence [2].

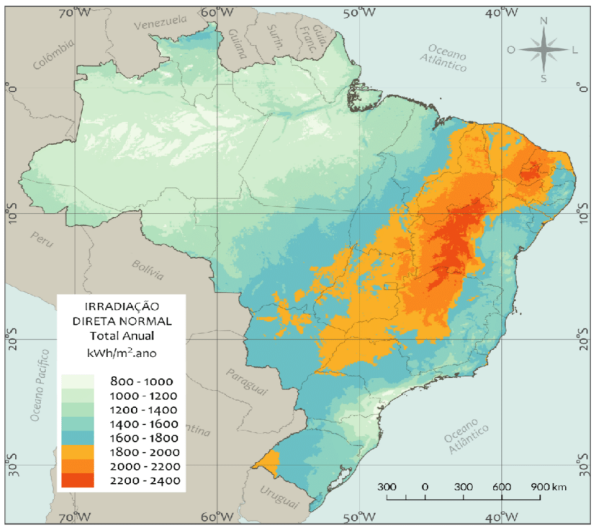


Figure 1: Sun incidence in Brazil [2]

The data measured came by land stations are the total solar incidence, as a function of the sky cloud coverage, the local temperature (maximum, minimum and mean values), the local atmospheric pressure, the local humidity, and other variables. The satellite can also gather the same data, with a different set of ways of measurement tough. These were the features for the Data Science (DS) methods too. Therefore, the total sun incidence can be set as the output or target variable, and all others variable as the input. As mentioned before, the season can outline another key factor, due to the atmospheric characteristics impacting into the total sun incidence measurement, which varies in time due other physical reasons.

This work focused on a forecast method for the sun incidence in the city of Manaus: latitude of -3.06 (S) and longitude of -60,0 (W), using Machine Learning models. The data came from local land stations and NASA [3,4], between 2013 to 2022, in a daily basis. As a first approach, the data from satellites was not concatenated with the data from the land station, due to the differences of the measuring procedures. Additionally, the data was aggregate in a month (113 data) and week (491 data) groups, to know whether there is any significant difference between them, when applying Time Series techniques.

Figure 2 shows the "All Sky Insolation Incident on a Horizontal Surface" in the vertical axel (kW.hr/m^2 per day), and the horizontal axel represents time (113 months) from the data exploratory analysis for the Manaus coordinates.

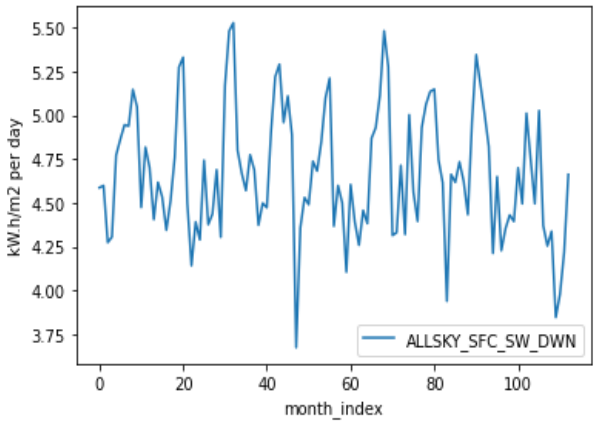


Figure 2: All sky insolation incident on a horizontal surface as a function of time.

The application of DS techniques considered Decision Tree (DT) models and Time Series (TS) as well. For the DT option, the data can be set with the measured total sun irradiance as the target variable, as a function of other variables. Prior to the use of the ML models, a feature importance analysis was carried out, as shown by Figure 3. The features linked to the water dispersion into the atmosphere explained with more relevance the output.

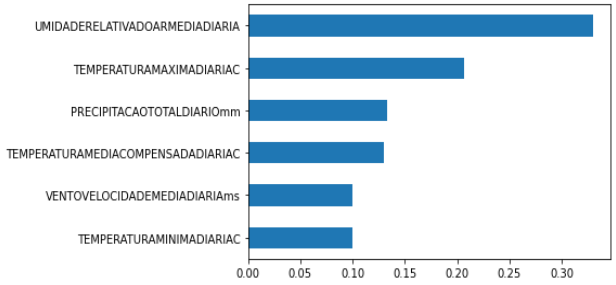


Figure 3: feature importance of the land station variables

The following methods were applied: Gradient Boost (GB), eXtreme Gradient Boost (XGBoost), Light Gradient Boost (LGB) and Adaptative Gradient Boost (AdaBoost). DT tools were selected because their simplicity to understand, to interpret and to visualize. The hyperparameters were optimized using a range of values (e.g.,4 or more, as the number of estimators etc.), and the cross-validation technique was used in this task too, with a fold of 5. It is worth noting other DS tools may also be used in a future work, such as neural networks.

The TS tool selected was the Vector Autoregression (VAR), once the work considered the hypothesis the variables may interfere among themselves. The influence between the time series can be checked whether they are bidirectional. In this model, each variable is taken as a linear combination of all past values, of itself and of the other variables. The causality test of Granger was used to check how strong this situation might be. Another test to check whether the connection between the variables was statistically significant was the Cointegration test.

The Data Science experiment was set in a matrix 2x2: (DT, TS) x (sat data, land data) and the four outcomes were evaluated by the mean absolute error (MAE) metric. The computing means were from the AWS cloud services (SageMaker), dealing with a 2CPU and 4GB for around 2 hours. Equally important, the target variable was the four more recent measurements, for both week and month groups.

III. RESULTS & CONCLUSIONS

The results can be seen in Tables 1 (DT) and 2 (TS). For the land measurements, the smallest MAE came from the week group. However, for the satellite group the best MAE were from the month group. More specifically, for the land station/week group, the lowest MAE (1.45) came from the LGB and, on the other hand, the GB method had the best results (0.38) for the satellite /week group. Focusing the month group, the best MAE came from the XGB (2.23), for the land measurements, and from the Random Forest (0.31) for the satellite data. Table 1 summarizes the best MAE. For the TS, the satellite groups had the least MAE.

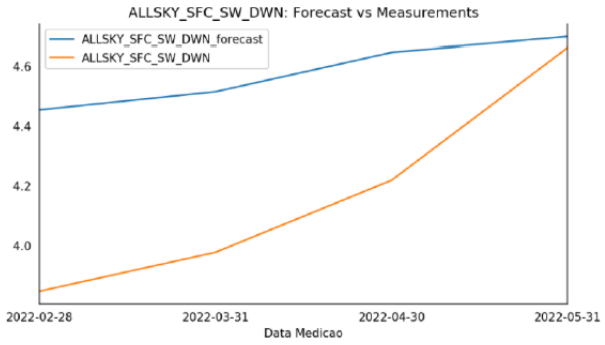
Table 1 – Best MAE from Decision Tree (DT) methods

ML model	Land/ wk data	Sat/ mo data
Random Forest	1.46	0.31
LGB	1.45	0.49
XGB	1.89	0.32
GrB	2.47	0.51
AdaB	2.67	0.41

Table 2 – MAE from the Time Series (TS) application

MAE	Week group	Month group
Land station	5.82	3.08
Satellite	1.42	0.40

Figure 4 shows the comparison between the forecast values and measurements for the TS satellite month group.



As a general conclusion, the ML models had smaller MAEs than ther TS VAR model, The order of magnitude of the MAE found in this work is close to other similar cases [5].

REFERENCES

[1] Almeida, R. M., Fleischmann, A. S., Brêda, J. P. F., Cardoso, D. S., Angarita, H., Collischonn, W., Forsberg, B., García-Villacorta, R., Hamilton, S. K., Hannam, P. M., Paiva, R., Poff, N. L. R., Sethi, S. A., Shi, Q., Gomes, C. P., & Flecker, A. S. (2021). Climate change may impair electricity generation and economic viability of future Amazon hydropower. *Global Environmental Change*, *71*. https://doi.org/10.1016/j.gloenvcha.2021.102383

[2] Pereira, Enio & Martins, Fernando & Costa, Rodrigo & Gonçalves, André & Lima, Francisco & Rüther, Ricardo & Abreu, Samuel & Tiepolo, Gerson & Pereira, Silvia & Souza, Jefferson. (2017). Atlas Brasileiro de Energia Solar – 2ª Edição. 10.34024/978851700089.

[3] https://power.larc.nasa.gov/data-access-viewer/

[4] https://tempo.inmet.gov.br/TabelaEstacoes/82331

[5] Ahmed, R., Sreeram, V., Mishra, Y., & Arif, M. D. (2020). A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. In Renewable and Sustainable Energy Reviews (Vol. 124). https://doi.org/10.1016/j.rser.2020.109792