# Renewable Energy: a parallel between USA and Brazil using Data Science.

**Abstract.** One of the major challenges in the transition with cleaner energy sources will be the structural modification of the electrical power grids. Many countries have started the changes in their electric energy matrices. In Brazil, the wind and solar options sum up 67 TW.h, in a noteworthy swelling in the last decade, and the hydraulic source produces 960 TW.h nowadays. The potential use of renewable energies can be explored with Machine Learning and Deep Learning techniques, under the key hypothesis of similar economic, geographic and population profiles. Data of surface area, total electric power generated, population and the Gross National Product (GNP), from Brazil and USA, were arranged to create overall ratios: population/area, GNP/area, GPN/population, electric power energy/population and GNP/electric power energy. These indexes are linked to the potential use of renewable energy based on the US states. These data are used then in Multivariable Linear Regression, Artificial Neural Networks and Ensemble Gradient Boosting algorithms, to predict the potential use of wind, solar, biomass, hydraulic and geothermal energy sources for the Brazil states. The best results came from the XGBoost algorithm pointing out the potential use of wind, solar and biomass sources combined with the order of magnitude of 100TW.h.

**Keywords:** Machine Learning, Deep Learning, Renewable Energy.

## 1 Introduction

In the last decades, the environmental impacts have increased and demanded policy reviews and actions, with data processing in larger scales than previously, using Data Science. Among the atmosphere changes, $CO_2$ emissions and their side effects have induced temperature and ocean levels rise. Equally important, these changes harm most of the biodiversity and contribute to intensify the frequency of natural disasters, such as tropical storms, flooding, draught, and others [1]. Many countries have moved to change this trend for the future generations.

One of the alternatives focuses the electric power matrix reshape, concentrating in wind, solar, hydro, biomass, and geothermal models, for instance, and dimming the fossil fuel options (e.g., coal and oil). Nonetheless, not all countries have manifold options, such as Brazil and the United States have, based on their geographic and economical characteristics. These two countries have all options available practically, providing a stable power generation profile with the combination of the sources, including the nuclear technology.

Although the benefits from the sustainable/renewable energy sources, there are several technical limits that bound their use into the national electric power matrix. The crucial aspect remains in the transient power generation profile of most of the renewable sources, like the wind and solar options. In a few words, one must have a way to

forecast the power generation, from each source, to balance all the supplies to the consumption level timely.

Brazil has one of the cleanest energy matrices, mostly based on hydro power, and sums up to 40% of the total energy produced in Latin America. In 2020, Brazil generated: 57 TW.h (wind); 10.7 TW.h (solar); 168 TW.h (biomass) and 961 TW.h (hydro)[2]. Nowadays, Brazil has experimented some environmental burdens and needs to reshape its electric energy matrix, considering the $CO_2$ reduction at the same time. The USA have started a change in the energy policy towards a greener option [3].

In this work, focusing the challenge to change the energy matrix, data of potential use of renewable energy from each state of the USA are analyzed and used, with Machine Learning and Deep Learning, to generate predicting models for the renewable electric energy option of each Brazil´s state, based on economic and geographic data, per state, such as the contribution to the Gross National Product (GNP), the surface area, the population, and the electric power capacity.

The datasets and metadata are available as open data and accessible throughout Github [4] and Zenodo [5].

## 2 Methodology

The general strategies to manage the energy matrix evolution, in any country, remain as long-term work and policy. Therefore, data collection, interpretation and use represent a key challenge. In this work, one takes data from the US renewable energy potential, by generation type, to be processed and generate prediction models, with Machine Learning (ML) and Deep Learning (DL) techniques. The methods of ML were: multi variable linear regression and decision trees (Random Forest, Gradient Boosting, Light, Adaptative and Extreme Gradient Boosting). The DL tool considered a neural network, with 5 to 8 intermediate layers with ‘relu’ activation functions. Then, the models are applied in the Brazil´s case. The results from each model are discussed in the end [6].

### 2.1 Datasets

One of the most key tasks in the Machine Learning techniques considers data collection and analysis. Moreover, one may think to use them into multidimensional math matrices, which demands an exploratory data analysis. One takes open data sources, from government agencies [7,8], keeping in mind five renewable energy options: biomass, geothermal, hydro, solar and wind. Although Brazil and USA have different economic profiles, some characteristics seem to be close, such as surface area and types of natural resources, after checking each national data about their economic output, geographic, data and energy production data [2,7,8].

The reference [9] presents the potential use of renewable energies in each USA state. In terms of broad evaluation, the solar, hydro and biomass options require larger surfaces areas than the wind and geothermal cases. In addition, the energy production is strongly related to the GNP and the population, considering the living standard. For instance, the solar, wind and biomass options demand normally large surface areas. The electricity production has links to the size of consumption market, which is related to

the size of the population. The distribution of the electrical energy generated can be associated to the demographic distribution or the ration between population and surface area.

To develop predicting models, within the renewable energy field of the two countries, one has chosen to work with ratio between overall data, derived from social, geographic, and economic areas: GNP per capita (GNP/pop); GNP/surface area (GNP/area); population/surface area (pop/area); GNP/electric power generated (GNP/epw); and electric power generated per capita (epw/pop). These indexes will be the parameters associated to the renewable energy potential of each type: wind, solar, biomass, geothermal and hydro. Figure 1 shows these indexes (normalized) based on the states data of the USA and Figure 2 for Brazil.

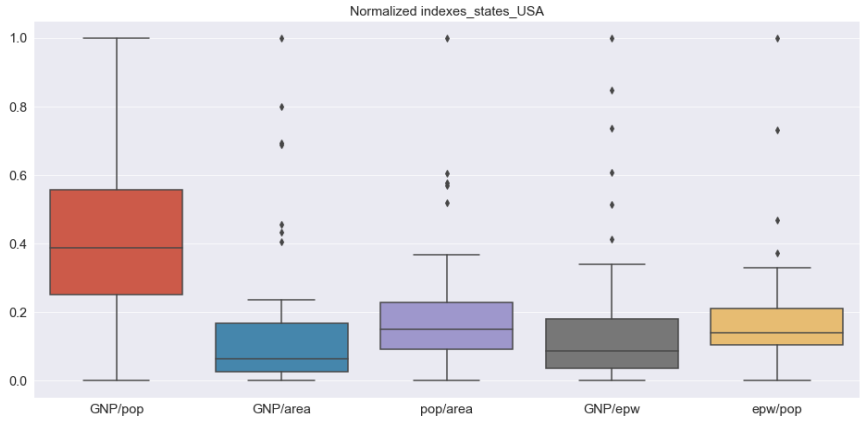


Figure 1: normalized overall indexes from USA states (boxplot)

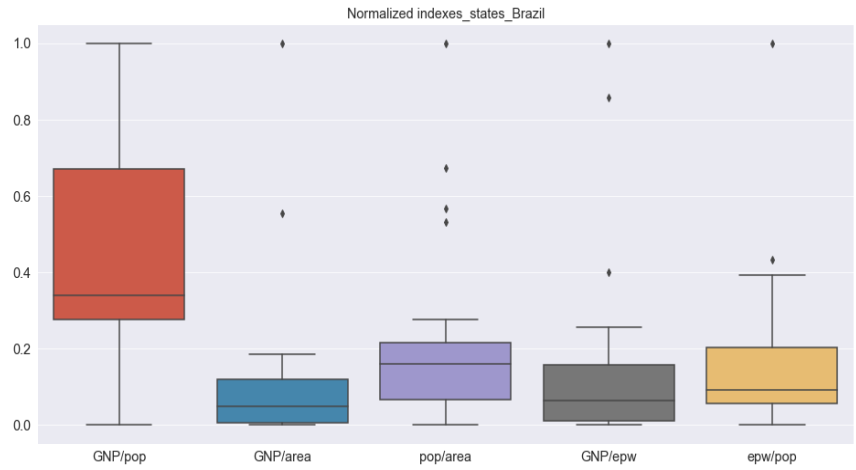


Figure 2: normalized overall indexes from Brazil states (boxplot)

Comparing the two figures, one sees a similar distribution of the overall indexes between the two countries. In addition, there are some ‘outliers’ in both cases, more numerous in the USA set. Reviewing some details about these outliers, one has found the states with a higher technology standard output, high demographic concentration and smaller surface area, such as the American states of Massachusetts and Connecticut. About the Brazilian outliers, one has seen a quite similar situation, mostly with the states of São Paulo, Paraná and Rio de Janeiro. For the sake of this work, all outliers were considered, except the data concerning the two capital cities (Brasilia – DF and Washington – DC) because they present a trend for higher values (kind of ‘bias’), although there is no significant local contribution (e.g. power plants, large rivers or dams etc.) for the energy production and distribution in each region.

Other comparisons can be made with the overall indexes of the two countries. Figure 3 presents the distribution of the ratio between the GNP and area, with normalized data, where the USA is related to the red/pink colors and Brazil to the green/blue ones. Note that the overall distribution keeps quite the same shape. In this work, the present comparison was taken as practical indication of some similarity between the two economies. For future work, one will use some statistical method to check the ‘likenesses’ numerically.

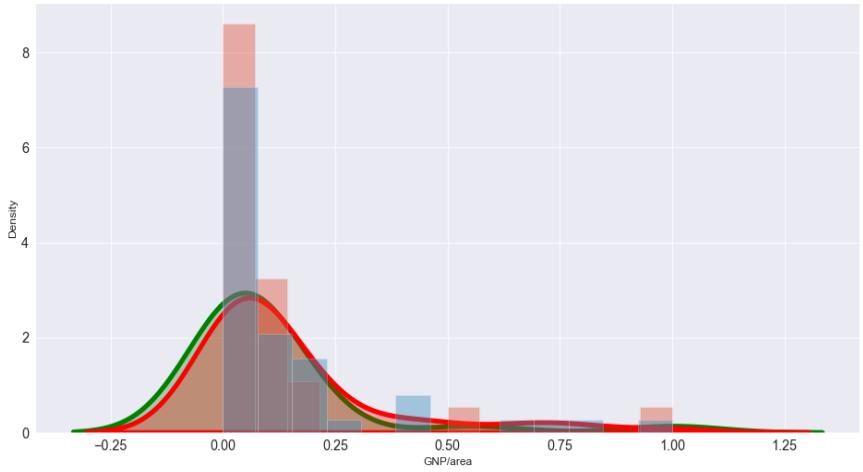


Figure 3: GNP/area ratio, with normalized data, from Brazil and USA states.

## 2.2 Numeric Models

Since the 80´s, Machine Learning models have been used to predict behaviors, using methods like Support Vector Machines (SVM), Decision Trees, Random Forest, and

others. In the 90´s, neural networks had taken a revamp, with better computer hardware and software [10].

In this work, the methods of Linear Regression (multivariable), Decision Trees (e.g., Gradient Boosting, Random Forests etc.) and artificial neural network (ANN) were used. From each state of the USA, the input data were GNP/pop, GNP/area, pop/area, GNP/epw, and epw/pop. The output data were the total electric energy generated, in GWh, by each renewable option: biomass, geothermal, hydro, solar and wind.

The Decision Tree methods are recommended to be used, mostly, when data have broad variability. A single decision tree may be not enough to explain all the data variability. Thus, one can combine several of them for the model to learn the features from the data. In this case, it can be referred as a 'Forest', with a specific way to create the decision trees, in this case 'randomly'. In short words, the 'boost' approach looks to increase the prediction performance by the combination of several simple decision trees, weighted by each accuracy. This work considered the split between the test and train data the threshold of 25%. The 'ensemble' methods used in this work were: Random Forest (RF), Gradient Boost (GBoost), Light Gradient Boost (LGBoost); Adaptative Boost (AdaBoost) and Extreme Gradient Boost (XGBoost).

Equally important, the use of 'Ensemble' methods increase the overall prediction performance and allow to understand the 'feature importance', or how a specific input data feature can explain the overall result, which is more relevant when dealing with manifold features.

In terms of ML parameters, one has searched for the best result, varying: the number of estimators; the maximum number of features; the maximum depth of the trees; the maximum of leaf nodes; the minimum samples at each leaf; the learning rate and others. Moreover, one has taken the option to use the 'bootstrap' method. For the boosting options, one has tested the options: 'dart', 'gbtree' and 'gblinear'. The overall optimal parameters found were learning rate of 1%; maximum number of estimators equals to 100; maximum tree depth of 5; maximum leaf nodes of 4; minimum samples at each leaf of 40. The key parameter was the mean square error (MSE).

# 3 Results and Discussion

## 3.1 Linear Regression

The multivariable linear regression took the input data as mentioned above, predicting the electric power generated by each renewable option for the US case. Table 1 presents the related math indicators of this method, after the use of the 'scikitlearn' tool.

**Table 1**: indexes from the multivariable linear regression (USA case)

| | R2 | R2(adj) | Prob(F) | P (value) | Durbin-Watson coef. | Jarque-Bera coef. |
|---|---|---|---|---|---|---|
| Solar | 0.98 | 0.97 | 6.4E-23 | 0.92 | 2.2 | 357.1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Wind | 0.96 | 0.95 | 1.07E-18 | 0.80 | 2.1 | 115.8 |
| Biomass | 0.99 | 0.99 | 5.67E-28 | 0.66 | 2.3 | 6.00 |
| Geotherm | 0.99 | 0.98 | 7.28E-27 | 0.89 | 2.1 | 156.3 |
| Hydro | 0.98 | 0.98 | 1.08E-28 | 0.96 | 1.5 | 78.5 |

Evaluating the metrics, one can see the linear regression had the best metrics for the biomass case. Nevertheless, due to the P(value) results, which should be smaller than 0.05, the Linear Regression method shall have a further review of its application.

### 3.2 Machine Learning Algorithms

Figures 4 and 5 show the feature importance output for the solar and wind cases. The vertical axis represents the parameter or indexes ratio, and the horizontal axis the importance of the parameter (%) explaining the data. For the first case, the most important feature is the GNP/area, with almost 25%, and for the second the ratio GNP/epw, close to 23%.

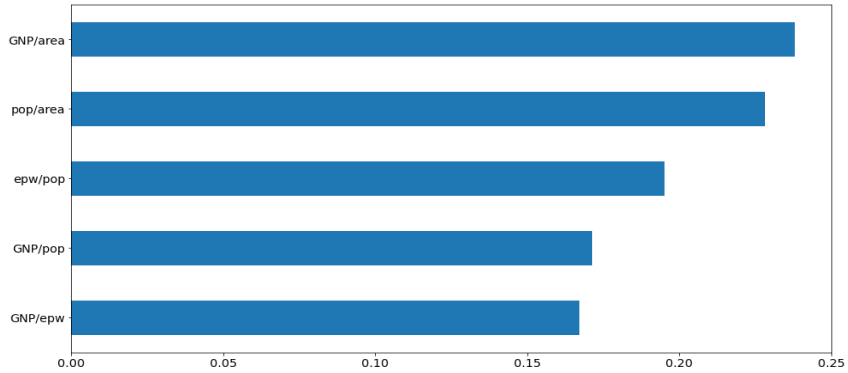


**Fig**.6: Feature importance distribution for the ‘solar’ case from USA data.

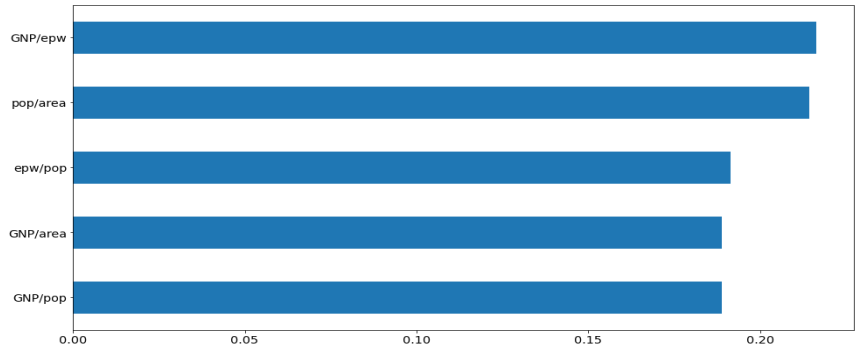

**Fig**.7: Feature importance distribution for the ‘wind’ case from USA data.

It is worth to note that the indexes ‘pop/area’ and ‘epw/pop’ have the same second and third position for both energy cases, near to 20%.

The Mean Square Error (MSE) is a common metric used to evaluate a prediction model, the lower MSE the better the model. In this task, the MSE considered the values between the target values taken for the ‘test’ phase and the predicted values with the input values for the same phase. No values used in the ML ‘training’ phase were used for this purpose. However, one has not found a reference MSE to compare, like in a similar previous work, which is advisable when the MSE values remain within a large band. Thus, to compensate this special case, this work also considered the MSE with the target values of the ‘test phase’ and the mean target value of the ‘training phase’, considering the ‘mean value’ as an expected first prediction output. This latter value may be considered as an overall prediction value. Equation 1 shows the metric ‘R’ between the two MSE described above.

$$\mathbf{R = MSE\ (y\_test, y\_pred)\ /\ MSE\ (y\_pred, mean\ y\_train)} \quad \mathbf{(1)}$$

Tables I to V show the comparison between the ‘ensemble’ methods in terms of the metrics. As already mentioned, the MSE (y_test, y_pred) values sound too high, not allowing the simple comparison with other cases [11][12]. Thus, one can see how useful the ratio R is: the lower its value, the better is the prediction method. For each energy case, the following tables present the metrics for the US data:

**Table II**: Solar energy & decision tree methods (US data)

| Method | MSE (* $10^{-19}$) | Ratio ‘R’ |
|---|---|---|
| Random_Forest | 2.2 | 2000 |
| Gradient_Boost | 1.7 | 0.94 |
| Light_Boost | 1.8 | 1.00 |
| XG_Boost | 1.2 | 0.67 |
| Ada_Boost | 1.4 | 0.79 |

**Table III**: Wind energy & decision tree methods (US data)

| Method | MSE (* $10^{-11}$) | Ratio ‘R’ |
|---|---|---|
| Random_Forest | 2.4 | 0.4 |
| Gradient_Boost | 4.4 | 0.74 |
| Light_Boost | 6.0 | 1.00 |
| XG_Boost | 1.5 | 0.25 |
| Ada_Boost | 2.7 | 0.44 |

**Table IV**: Biomass energy & decision tree methods (US data)

| Method | MSE (* $10^{-7}$) | Ratio 'R' |
|---|---|---|
| Random_Forest | 3.3 | 0.99 |
| Gradient_Boost | 3.3 | 1.00 |
| Light_Boost | 3.3 | 1.00 |
| XG_Boost | 3.5 | 1.06 |
| Ada_Boost | 3.7 | 1.11 |

**Table V**: Hydro energy & decision tree methods (US data)

| Method | MSE (* $10^{-7}$) | Ratio 'R' |
|---|---|---|
| Random_Forest | 8.6 | 0.96 |
| Gradient_Boost | 8.9 | 0.99 |
| Light_Boost | 9.0 | 1.00 |
| XG_Boost | 8.5 | 0.95 |
| Ada_Boost | 9.1 | 1.01 |

**Table VI**: Geothermal energy & decision tree methods (US data)

| Method | MSE (* $10^{-11}$) | Ratio 'R' |
|---|---|---|
| Random_Forest | 3.0 | 1.23 |
| Gradient_Boost | 2.4 | 0.99 |
| Light_Boost | 2.4 | 1.00 |
| XG_Boost | 1.8 | 0.73 |
| Ada_Boost | 2.3 | 0.94 |

According to this criterion, the best method was the XGBoost, chosen then to predict with Brazil data. Regarding the results, Tables VII to IX present the predicted values for the solar, wind and biomass cases, for the first four higher values.

**Table VII**: XGBoost predicted values for solar energy.

| Brazil´s state | Predicted energy (*$10^{-5}$ GW.h) |
|---|---|
| Espírito Santo | 12.2 |
| São Paulo | 10.1 |
| Rio de Janeiro | 13.5 |
| Alagoas | 7.7 |

**Table VIII**: XGBoost predicted values for wind energy.

| Brazil´s state | Predicted energy (*$10^{-5}$ GW.h) |
|---|---|
| Rio Grande do Sul | 2.5 |
| Amapá | 4.5 |
| Maranhão | 4.5 |
| Piauí | 4.4 |

**Table IX**: XGBoost predicted values for biomass energy.

| Brazil´s state | Predicted energy (*$10^{-3}$ GW.h) |
|---|---|
| Amazonas | 8.2 |
| Tocantins | 8.2 |
| Pará | 8.1 |
| Mato Grosso | 8.1 |

The predicted values for the hydro and geothermal energies presented values not consistent, with high module changes among the Brazilian states, and must have a further review.

### 3.3 Deep Learning

Due the same reason of the wide data variability, the use of neural networks seems suitable to deal with the purpose of this work. The basic topology considered one input layer, 8 intermediate layers and one output layer. Several activation functions were tested and the best one was the 'Relu'. The 'kernel_initializer' was the 'He_Uniform' and the output activation function was the 'linear'. The total number of 'epochs' was of 100 to generate stable errors. For the solar option, figure 8 shows the MSE evolution as a function of the epochs.

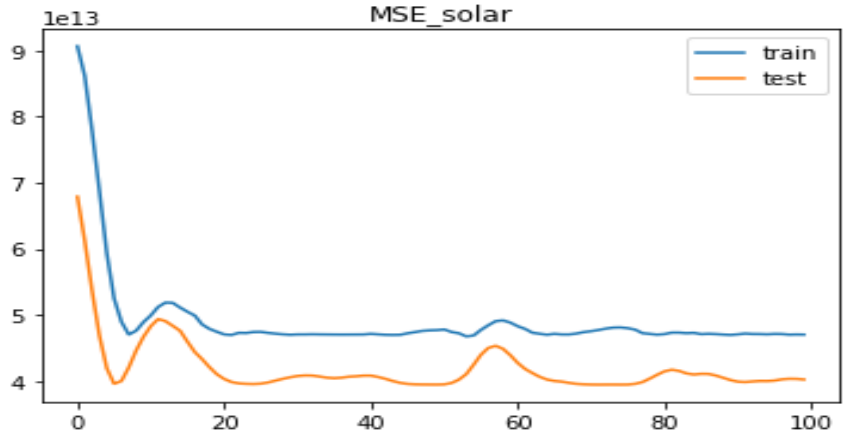


**Fig**.8: ANN MSE for the solar energy as a function of number of epochs.

As one can note, the MSE from the ANN is $10^6$ smaller than the values from the XGBoost algorithm. Figure 9 shows the MSE for the wind energy case.

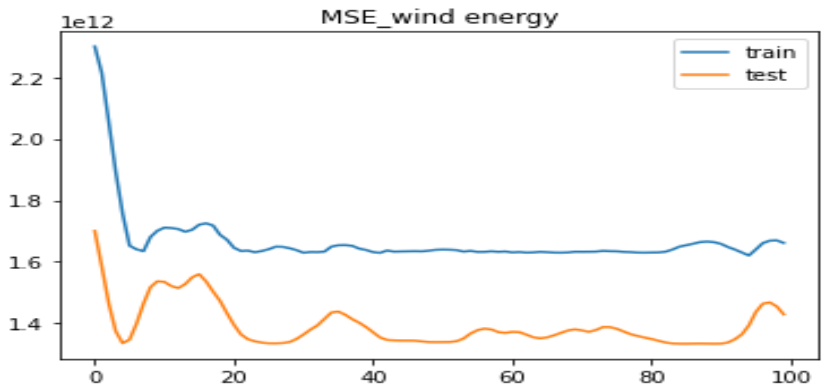


**Fig**.9: ANN MSE for the wind energy as a function of number of epochs.

Note the change in the profile in comparison with the MSE solar energy, and an order of magnitude 10 times smaller. Figure 10 covers the biomass energy case.

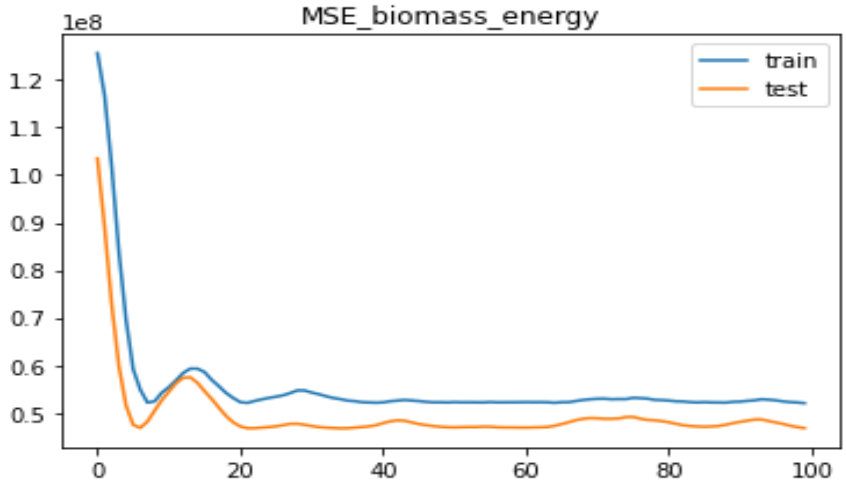


**Fig**.10: ANN MSE for the biomass energy as a function of number of epochs.

The MSE profile of the biomass case sounds closer to the solar energy, but its amplitude $10^4$ smaller than the two previous cases. Table IX presents the MSE for all the ANN models for a comparison, considering the USA test data.

**Table IX**: MSE for each ANN application.

| Energy option | ANN MSE – USA test data |
|---|---|
| Wind | 1.4 E13 |
| Solar | 4 E13 |
| Geothermal | 3 E11 |
| Hydro | 0.4 E8 |
| Biomass | 0.5 E8 |

The MSE profile of the biomass case sounds closer to the solar energy, but its amplitude $10^4$ smaller than the two previous cases.

The ANN predicted values for the Brazil´s states did not show variation, for the same energy case, which must be investigated later. In addition, the order of magnitude among the values must be review either. Table X summarizes the predicted values for each energy.

**Table X**: ANN energy predicted values for Brazil´s states.

| Energy type | Predicted energy ($*10^{-3}$ GW.h) |
|---|---|
| Solar | 6950 |
| Wind | 800 |
| Biomass | 8.1 |
| Hydro | 8.1 |
| Geothermal | 650 |

## 4 Conclusions

The key hypothesis of the work relays on some resemblance of geographical and natural resources between the USA and Brazil. Data from the states of the USA were used to build up prediction models on the potential use of renewable sources in Brazil, to provide hints in an electric matrix change to cope with the new environment goals.

This work used multivariable linear regression, Machine Learning and Deep Learning algorithms, using as input data some overall indexes based on the GNP, population, surface area and the electric power generation.

Further investigation must be carried out on the multivariable linear regression; other 'Ensemble' methods and the ANN model can be also used and improved. A statistical method shall be considered to check the indexes 'likenesses' between Brazil and the USA numerically.

Among the above methods, the 'XGBoost' presented the best results, in terms of the ratio 'R' (ratio between MSE´s). The best prediction values about the potential use of renewable energies were the solar, wind and biomass cases.

## References


1. Accelerating clean-energy transitions in major emerging economies. htpps://www.iea.org/areas-of-work/programmes-and-parternships/clean-energy-transitions-programme. Accessed: 2020-11-30.
2. Balanço Energético Nacional 2021. https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-601/topico 588/Relatório Síntese. Accessed: 2021-06-16.
3. Current and Future Energy Sources of the USA. https://www.e-education.psu.edu/egee102/node/1930. Accessed: 2021-06-16.
4. André L. F. Marques, Denise Florio, Ulisses A.S. Costa. US-renewable-energy-potential. https://github.com/65-1157/US-renewable-energy-potential-1, December 2020.
5. André L. F. Marques, Denise Florio, Ulisses A.S. Costa. US-renewable-energy-potential. 10.5281/zenodo. 4310214, 4327348,4310206,4310102,4310173. December 2020.
6. Beluco, A., During F°, F.A., Silva, L.M.R., Silva, J.S., Teixeira, L.E., Vasco, G., Canales, F.A., Rossini, E.G., de Souza, J., Daronco, G.C. and Risso, A., 2020. Dataset after Seven Years Simulating Hybrid Energy Systems with Homer Legacy. Data Science Journal, 19(1), p.14. DOI: http://doi.org/10.5334/dsj-2020-014
7. Instituto Brasileiro de Geografia e Estatística. https://www.ibge.gov.br/. Accessed: 2020-11-28.
8. United States Department of Energy – DOE. https:// www.energy.gov/. Accessed: 2020-11-28.
9. National Renewable Energy Laboratory. https://www.nrel.gov/gis/re-potential.html. Accessed: 2020-11-28.
10. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT Press 2016. https://www.deeplearning.org
11. Kalimoldayev, Maksat & Drozdenko, Aleksey & Koplyk, Igor & Marinich, T. & Abdildayeva, Assel & Zhukabayeva, Tamara. (2020). Analysis of modern approaches for the prediction of electric energy consumption. Open Engineering. 10. 350-361. 10.1515/eng-2020-0028.
12. Leme, João Vitor et al. Towards assessing the electricity demand in Brazil: Data-driven analysis and ensemble learning models. Energies, v. 13, n. 6, 2020.