Scientific Research Publishing

Energy and Power Engineering, 2022, 14, 327-342
https://www.scirp.org/journal/epe
ISSN Online: 1947-3818
ISSN Print: 1949-243X

An Application of Decision Trees Algorithm to Project Hourly Electricity Spot Price as Support for Decision Making on Electricity Trading in Brazil

Cosme Rodolfo R. dos Santos[1*], Roberto Castro[2], Rafael Marques[3], Luiz Carlos Pereira da Silva[1]

[1]State University of Campinas, School of Electrical and Computer Engineering, Campinas, Brazil
[2]Independent Consultant, Vinhedo/SP, Brazil
[3]Brazilian Association of Photovoltaic Solar Energy (ABSOLAR), São Paulo, Brazil
Email: *cosmerodolfo@gmail.com

How to cite this paper: dos Santos, C.R.R., Castro, R., Marques, R. and da Silva, L.C.P. (2022) An Application of Decision Trees Algorithm to Project Hourly Electricity Spot Price as Support for Decision Making on Electricity Trading in Brazil. *Energy and Power Engineering*, **14**, 327-342.
https://doi.org/10.4236/epe.2022.148018

Received: July 21, 2022
Accepted: August 20, 2022
Published: August 23, 2022

Abstract

Estimating the price of a financial asset or any tradable product is a complex task that depends on the availability of a reasonable amount of data samples. In the Brazilian electricity market environment, where spot prices are centrally calculated by computational models, the projection of hourly energy prices at the spot market is essential for decision-making, and with the particularities of this sector, this task becomes even more complex due to the stochastic behavior of some variables, such as the inflow to hydroelectric power plants and the correlation between variables that affect electricity generation, traditional statistical techniques of time series forecasting present an additional complexity when one tries to project scenarios of spot prices on different time horizons. To address these complexities of traditional forecasting methods, this study presents a new approach based on Machine Learning methodology applied to the electricity spot prices forecasting process. The model's Learning Base is obtained from public information provided by the Brazilian official computational models: NEWAVE, DECOMP, and DESSEM. The application of the methodology to real cases, using back-testing with actual information from the Brazilian electricity sector demonstrates that the research is promising, as the adherence of the projections with the realized values is significant.

Keywords

Artificial Intelligence, Machine Learning, Price Estimation, Energy Planning, Spot Electricity Market, Spot Prices Forecast

1. Introduction

The use of computational techniques in time series forecasting is one of the most active fields in academic research. With the advent of machine learning techniques, new algorithms have been developed and often made available through libraries for large-scale use.

This study discusses some characteristics of electricity price formation for the spot electricity market in Brazil, also known as by the initials SPD (Settlement Price for the Differences), and its applications through a new methodology for time series forecasting, using Machine Learning techniques, more specifically the XGBoost algorithm.

When XGBoost is compared to ARIMA [1], it is observed that XGBoost has advantages, such as the lack of a need to preprocess the data, a fast operation speed, complete feature extraction, a good fitting effect, and high prediction accuracy. When XGBoost is compared to Deep Learning techniques [2], the main conclusion points out that boosting methods demand fewer data and features. Other studies [3], beyond time series scope, also point out that XGBoost is more suitable for tabular data, in the case of this study. The main reasons are specific features of tabular structure: irregular patterns in the target function, uninformative features, and non-rotationally-invariant data where linear combinations of features misrepresent the information.

In this study, the time series features will be passed to the model as input variables, allowing the representation of seasonality and time cycles. A variable representing the characteristics of the Brazilian Interconnected Electricity System (BSIN), from the official models NEWAVE, DECOMP, and DESSEM, is also inserted in the model. Different forecast horizons will be provided, and especially for the very short and short term, the forecast performance will be measured [4] and presented.

2. Characteristics of the Brazilian Electricity Price Definition

The BSIN is characterized by the interconnection, through the Basic Power Transmission Network, of four of the country's subsystems: Northeast, North, South, and Southeast, the latter together with the Center-West. The interconnected system is operated centrally by the National Electric System Operator (ONS) and is integrated by different electricity generation and transmission companies, which may be public or privately owned.

The Brazilian electricity generation profile makes the system predominantly hydrothermal; that is; most of the energy consumed is generated in hydroelectric powerplants, in addition to thermal plants (nuclear, natural gas, biomass, coal, and fuel oil) and renewables such as wind and solar plants.

Data from ANEEL, published in January 2022, indicate that 60.12% of the electric energy demand is met by hydroelectric powerplants (HPP), 8.95% by thermal power plants (TPP) operating on natural gas, 5.02% by TPPs operating on liquid petroleum-derived fuels, and 1.97% by coal-fired TPPs. These Hydro

and Thermal sources meet about 76% of the energy demand [5].

Renewable sources, such as wind, small hydro, and solar power plants, play an important role in the composition of the electricity generation matrix, but in the operational optimization process, the expected renewable generation from these plants is deducted directly from the expected Gross Demand for electricity and is not explicitly considered in the centralized operation optimization process.

Due to the stochastic nature of the Gross Demand and the renewable generation, this operation allows a single parcel to be presented to the hydrothermal optimization model, and this parcel is known as the Net Energy Demand, adopted as deterministic (single scenario) in the SPD calculation process.

The system's hydrothermal characteristic impacts the electricity generation's price since the cost of generating electricity is a function of the optimal dispatch of the hydraulic and thermal sources to minimize the cost over an operation horizon.

In pricing the SPD, the computer models used by ONS to operate the BSIN optimize the operation and calculate the Marginal Cost of Operation (MCO). These models take into account the constant changes in the operating condition of the system, especially related to meteorological issues—favorable or not of the Affluent Natural Energy (ANE) and Energy Stored (ES) in the reservoirs of the hydroelectric power plants—as well as the Unit Generation Costs (UGC) of each thermal power plant.

2.1. Variable Related to the Hydrothermal System

The chain of computer models, more specifically NEWAVE, DECOMP, and DESSEM, are used by both ONS, the system operator, and the Chamber of Electric Energy Commercialization (CCEE), the market operator but for different purposes. ONS seeks the best way to operate the power system to guarantee the supply of demand at the lowest cost. CCEE aims to determine the SPD, for each submarket with hourly granularity, which will be used in the short-term market accounting and financial settlement (*i.e.*, spot market) [6].

To guarantee a connection between the results generated by the two institutions, ONS runs the models first, and then CCEE uses these results and treats the electric constraints and the generating units under tests, processes the models again, and finally publishes the hourly prices. The coupling among the three models occurs through the so-called Future Cost Function (FCF). This function represents a cost associated with each of the several possible trajectories of the state variables and energy storage in the reservoirs of the hydroelectric power plants, together with the respective thermal complements, to satisfy the Net energy demand. **Figure 1** below shows a relation of each model, using FCF as a coupling parameter.

The hydrological scenario fluctuation is the main responsible for the FCF variability, so it is mandatory to include a variable in the Machine Learning model that represents this oscillation between the various scenarios. In this research, it will be used a weekly version of the SPD—also called SPD Week Level—as this

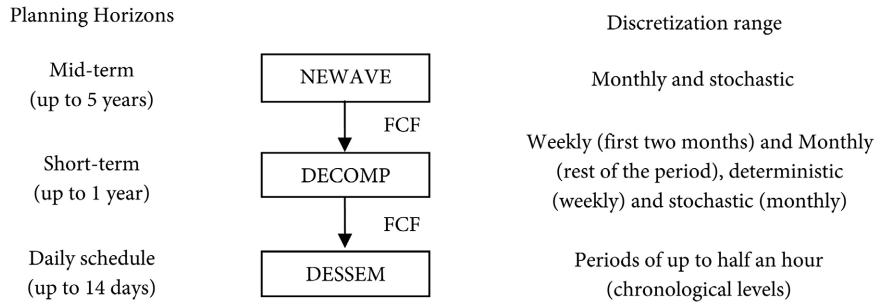

Figure 1. Model coupling structure.

independent input variable. It represents the electricity price level, divided into the blocks of heavy, intermediate, and low electricity consumption throughout the week, as follows:

- Very Short and Short Term:
- SPD Week Level from the MCO of the DESSEM and DECOMP output deck, specifically the deterministic part of the models.
- Medium Term:
- SPD Week Level from the MCO from the DECOMP output deck, specifically the stochastic part of the model.
- Long Term
- SPD Week Level from the MCO of the NEWAVE output deck, presenting a stochastic approach for longer-term hydrological scenarios.

2.2. Temporal Variables

The price forecast model considers and represents the patterns associated with the time profile of the various horizons considered. Thus, the following independent variables were included, as shown in **Table 1**:

- Hour, it can represent the hourly fluctuations (*i.e.*, peak and valley hours of prices), essential for very short-term analysis;
- Day of the Week and Day of the Month, which can represent oscillations within the same week and month, is suitable for short-term analysis (*i.e.*, separation of weekdays and weekends, holidays, beginning of the month, etc.);
- Month, to represent the effects of annual economic and meteorological events (*i.e.*, dry season, rainy season, scholar vacations, year-end festivals, carnival, etc.).

2.3. Hourly Pricing Variable (Hourly SPD)

This is the dependent and target variable of the forecast. To obtain the correlation of this variable with all the possible values assigned to the others, it is part of the database used in the learning process.

The hourly SPD history used starts on April 17, 2018, and extends until July 16, 2021, with daily updates, and the hourly SPD value started to be effectively used in CCEE's accounting and financial settlement as of 1/1/2021. Between

Table 1. Southeast sub Market XGBoost model variables for April 17, 2018.

Hourly SPD (R$/MWh)	Weekly SPD (R$/MWh)	Month	Day	Day of the Week	Hour
40.16	118.17	4	17	1	0
40.16	118.17	4	17	1	1
40.16	118.17	4	17	1	2
40.16	118.17	4	17	1	3
40.16	118.17	4	17	1	4
40.16	118.17	4	17	1	5
40.16	118.17	4	17	1	6
40.16	125.33	4	17	1	7
116.86	125.33	4	17	1	8
119	125.33	4	17	1	9
121.45	125.33	4	17	1	10
121.45	125.33	4	17	1	11
119.07	125.33	4	17	1	12
121.41	125.33	4	17	1	13
122.74	125.33	4	17	1	14
123.81	125.33	4	17	1	15
121.35	125.33	4	17	1	16
119.11	125.33	4	17	1	17
120.76	125.33	4	17	1	18
119.16	125.33	4	17	1	19
119.11	125.33	4	17	1	20
119.1	125.33	4	17	1	21
118.31	125.33	4	17	1	22
114.26	125.33	4	17	1	23

April 2018 and December 2020, the parameter was calculated on an experimental basis, known as "Shadow Operation".

3. XGBoost: A Decision Tree Based Algorithm

Decision Trees are branching structures with three types of nodes used in data classification. The root node represents the entire data set. After it, there are internal nodes, which represent the variables in the data set and the decision criteria for further branching. In summary, each internal node will contain a comparison of a given variable $x_i \in X$—also called the independent or input variable of the model—against a specific value and criterion, such as: "is $x_i \geq 23.7$?" [7].

From the answer to this comparison presented at this inner node, which can assume "TRUE" or "FALSE" values, there will be a new branch to the left or right. This branching will continue until there is no more possibility to proceed, either because all variables have been entered into the model or because all samples have been correctly classified. The last node of the tree is called the leaf and consists of the target variable of the prediction commonly called the dependent variable and notated as y_i.

The most common and used structures are called CART (Classification and Regression Trees) because of their applicability in several classes of problems, which involve classification and regression.

As advantages of using this type of structure, one can list, as presented by [7]:

- Non-linearity, due to the possibility of representing complex data classification boundaries through logical branching;
- Support for categorical variables, which result in binary outcomes, TRUE or FALSE, for example;
- Interpretability, due to the easily built of a self-explanatory structure;
- Robustness, due to the possibility of exponential growth of new variables and possible tests;
- Application in regression problems, in addition to the traditional classification applications in the Data Science universe. This advantage is crucial to obtain a prediction of a numerical value, the target of this study.

The XGBoost stands for Extreme Gradient Boosting and is a scalable and distributed gradient-boosted decision tree (GBDT) machine learning library. It is built upon distributed computing and parallel tree boosting, and the model uses the following concepts: supervised machine learning, decision trees, ensemble learning, and gradient boosting.

In this study, two XGBoost model metrics will be assessed. The first is called Weight and represents the frequency of use of which variable to generate new branches. A higher value of Weight for a variable also indicates its greater importance in the definition of the forecast model. A higher value of Weight for a variable indicates its greater importance in the definition of the forecast model.

The second is called Gain split, based on Equation (1) and indicates the average Gain value of a variable used in the creation of new branches. A higher Gain value for a variable, relative to the others also means its greater importance in generating forecasts.

$$\begin{aligned}\mathcal{L}_{split} &= \mathcal{L}_E + \mathcal{L}_D - \mathcal{L}_R - \gamma \\ &= \frac{1}{2}\left[\frac{\left(\sum_{i\in I_E} g_i\right)^2}{\sum_{i\in I_E} h_i + \lambda} + \frac{\left(\sum_{i\in I_D} g_i\right)^2}{\sum_{i\in I_D} h_i + \lambda} - \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda}\right] - \gamma\end{aligned} \tag{1}$$

4. Results

To facilitate the presentation of graphs and results, only the Southeast submarket will be addressed, but the machine learning model can also be applied to other

electrical submarkets, with very similar results. The main metropolitan regions of Brazil are part of the Southeast submarket, making this chosen scenario a good example to discuss the proposed methodology.

Additionally, the performances of the very short-term and short-term forecasts were evaluated with two metrics commonly used in linear regression assessment, which are MAPE and RMSE (Mean Absolute Percentage Error and Root Mean Square Error, respectively).

A third metric was specially created by the authors and was named TAI (Trend Accuracy Index), defined according to Equation (2). This index represents the adherence between the actual hourly SPD variation, verified between subsequent hours, with the forecasted variation in these same periods. To exemplify, consider that between 00:00 and 01:00 on a given day, an increase in the SPD was predicted, and when verifying the actual SPD for the same period, it was found that the actual SPD also showed an increase. In this example, a score of 1 will be assigned for this hit, and if it is not similar, a score of 0 will be assigned.

This metric aims to provide the decision-maker with the sensibility on what to expect from the price behavior in the analysis horizon, answering with certain accuracy the question: In this horizon will the SPD go up or down?

After the analysis of all intervals, all scores are added up and the result is divided by the maximum possible score, which is 24 (number of hours in a day). The result, multiplied by 100, will represent a percentage score of the adherence between the Forecast and the Realized, in the variation of the SPD.

$$score_i = \begin{cases} 1, \text{if } For_i > For_{i+1} \text{ and } Real_i > Real_{i+1}, \text{ or } For_i < For_{i+1} \text{ and } Real_i < Real_{i+1} \\ 0, \text{otherwise.} \end{cases}$$

$$TAI = \frac{\sum_{i=1}^{24} score_i}{24} \times 100 \quad (2)$$

where:

For_i = Variable of the predicted hourly SPD increase or decrease, for period *i*.

$Real_i$ = Variable of the real hourly SPD increase or decrease, for period *i*.

$score_i$ = grade assigned for the forecast accuracy, for period *i*.

4.1. Weight Criterion

The first model indicator is called Weight, and Figure 2 shows an example obtained for Southeast. As per the results, the SPD Week Level variable was the highlight, with a utilization varying from 607 times for the Northeast to 717 times for the South. And with the worst performance in this aspect, the Day of the Week variable presented a usage varying from 86 for the Southeast to 108 for the North.

The second variable of note here is Month for Southeast, where it was relevant to define new branches. The same result was obtained for South as well, but for Northeast and North, the variable Day was the second more relevant.

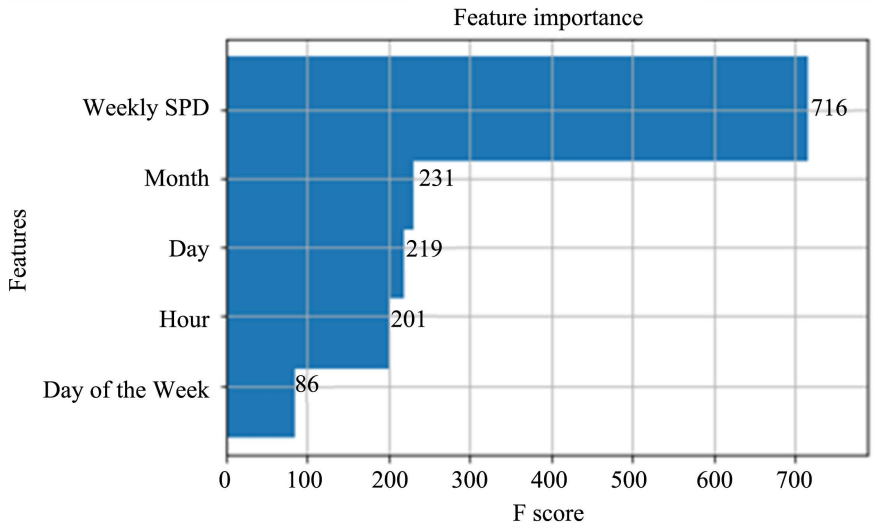

Figure 2. Importance of the variables (Weight Criterion)—Southeast submarket.

4.2. Gain Criterion

Figure 3 presents the results for the second metric called Gain, obtained for Southeast. It was verified that the SPD Week Level excels the other variables, just like the Weight criterion, with average Gain values ranging from 0.78 for the Northeast to 0.9 for the North.

It is worth mentioning that the Gain values are normalized so that the sum of all the Gains equals 1. Additionally, in **Figure 3**, the model variables are represented by the acronyms f0 (SPD Week Level), f1 (Month), f2 (Day), f3 (Day of the Week), and f4 (Hour).

The second variable of note here is Month, which was relevant to all submarkets. The third more relevant variable was Hour for Southeast and South models and Day for Northeast and North model.

Concerning the Day of the Week, they presented the smallest gains Southeast, South and North. And as the model evolves, with the inclusion of new samples and other independent BSIN variables, this variable may be candidate to be removed, depending on the improvement of two factors: accuracy and processing time.

4.3. 1-Day Ahead Forecast Performance

In this method of performance analysis, the one-day ahead forecast took into consideration the inclusion of the previous day in the Learning Base. This strategy allowed keeping the model as up-to-date as possible regarding the correlations of variables of the BSIN model, embedded in the SPD Week Level variable, with the hourly SPD to be predicted.

The forecast period was divided into hourly discretizations from 00:00 AM on July 17, 2021, to 11:00 PM on July 23, 2021, **Figure 4** shows the forecasted hourly SPD and the actual one for July 19th and the MAPE, RMSE, and TAI partials for the whole week forecast is shown in **Table 2**.

Table 2. 1-day ahead forecast results.

	Southeast			South		
	MAPE (%)	RMSE (R$/MWh)	TAI (%)	MAPE (%)	RMSE (R$/MWh)	TAI (%)
July 17, 2021 Saturday	1.67	11.73	65.22	3.04	20.79	56.52
July 18, 2021 Sunday	5.92	40.32	82.61	4.49	30.75	73.91
July 19, 2021 Monday	1.25	10.6	73.91	1.06	8.06	73.91
July 20, 2021 Tuesday	1.09	7.08	73.91	0.92	6	60.87
July 21, 2021 Wednesday	2.38	15.52	82.61	2.19	16.52	65.22
July 22, 2021 Thursday	1.21	8.25	65.22	1.26	8.65	56.52
July 23, 2021 Friday	0.81	5.15	60.87	0.85	5.67	78.26
Average	2.05	11.3	72.05	1.97	13.78	66.46
	Northeast			North		
	MAPE (%)	RMSE (R$/MWh)	TAI (%)	MAPE (%)	RMSE (R$/MWh)	TAI (%)
July 17, 2021 Saturday	1.33	10.81	47.83	2.18	14.53	60.87
July 18, 2021 Sunday	5.08	32.12	69.57	4.72	30.38	69.57
July 19, 2021 Monday	1.30	9.69	69.57	1.48	12.65	73.91
July 20, 2021 Tuesday	1.49	10.94	73.91	0.56	4.1	73.91
July 21, 2021 Wednesday	2.48	16.63	52.17	1.84	14.97	73.91
July 22, 2021 Thursday	1.67	11.13	47.83	0.81	7.23	65.22
July 23, 2021 Friday	1.86	16.31	69.57	1.00	6.66	73.91
Average	2.17	15.38	61.49	1.80	12.93	70.19

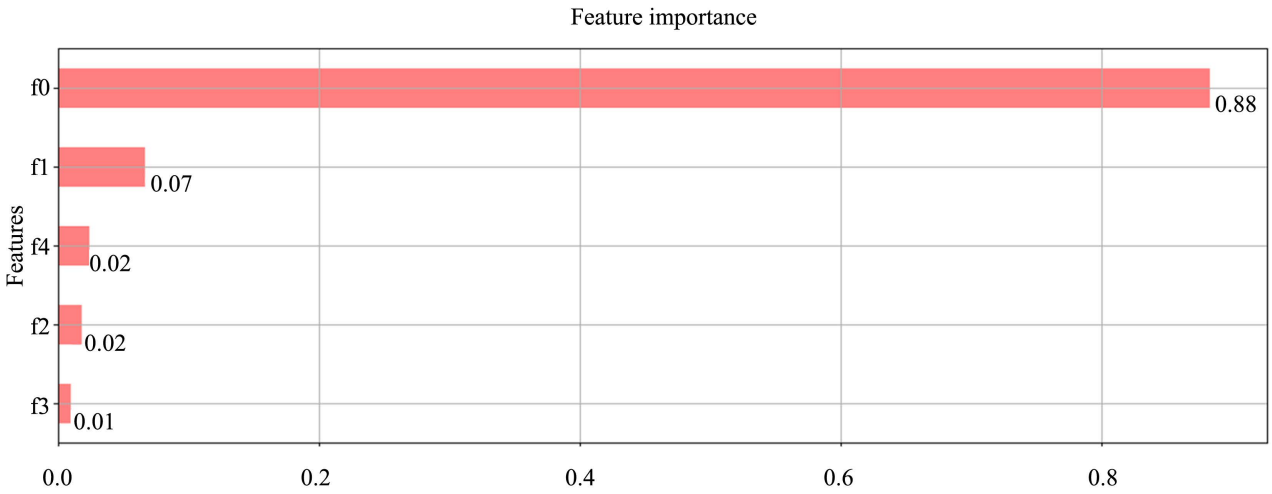

Figure 3. Importance of the variables (Gain Criterion) – Southeast submarket.

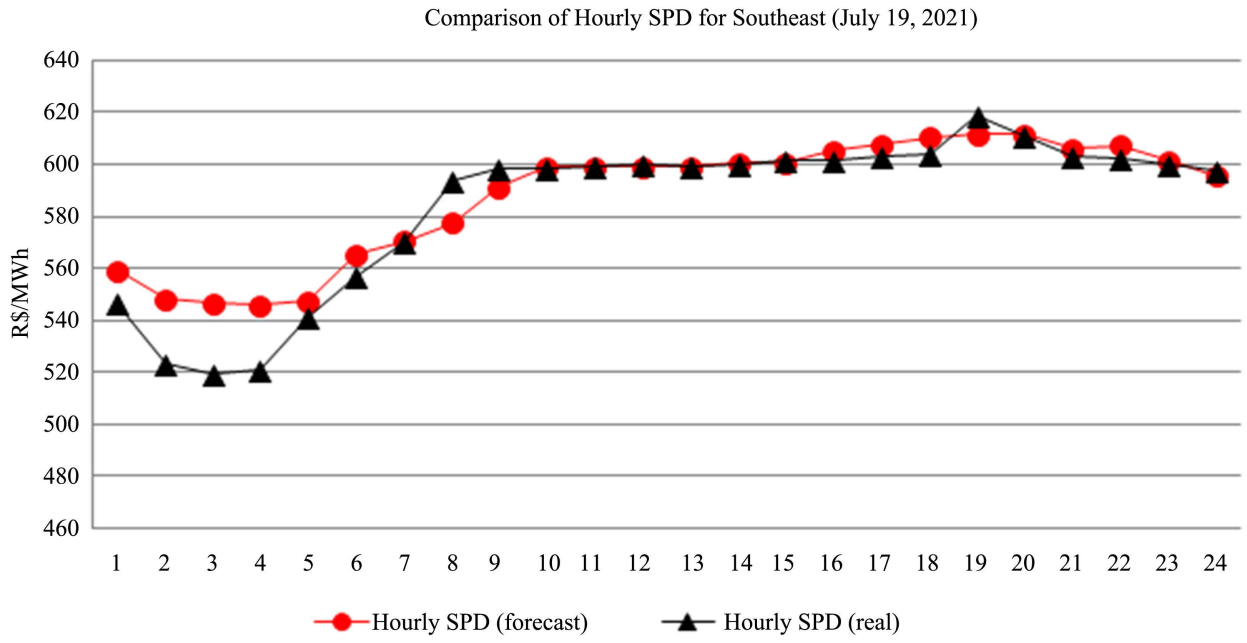

Figure 4. 1-day ahead forecast example, for Southeast submarket.

4.4. 7-Days Ahead Forecast Performance

In this method, an hourly SPD forecast was performed with a horizon of seven days ahead. The model was trained only at the beginning of the process; that is, as the days evolved within the horizon, there were no new model updates. At the end of the seven days, its assertiveness was measured based on the initial view. Figure 5 shows the results obtained for the Southeast submarket.

Table 3 shows the results of the forecast performed on July 16, 2021, for the period from July 17 to July 23, 2021.

4.5. 1-Month Ahead Forecast Performance

Like the seven-day ahead forecast and performance analysis, in this method, the model update occurs only on the day the forecast is calculated, and the model is not updated during the next 30 days. The model's performance is measured at the end of the first month period. For the analysis, a one-month forecast limit was set, however, this analysis can be extended to a period up to two months, according to public files provided by CCEE. From the third month on, the forecast acquires an informative character of conjectural trends in the sector.

For this horizon, the publication date coincides with the Monthly Operation Programming (MOP), held by ONS monthly on the last Friday of the month before the validity of this programming. In the case of the forecast below, the MOP date was June 25, 2021, for the programming of July 2021. Figure 6 shows how the forecast and real SPD movement behaved on this horizon.

The results indicate that the short-term forecast presents higher volatility than the very short-term forecast due to the updates of the SPD Week Level that occur between the availability of the short-term study, known as revision 0 or *rv0*

Table 3. 7-days ahead forecast results.

	MAPE (%)	RMSE (R$/MWh)	TAI (%)
Southeast	1.77	15.10	73.81
South	2.14	17.22	69.64
Northeast	2.08	17.32	60.71
North	1.84	14.56	72.62

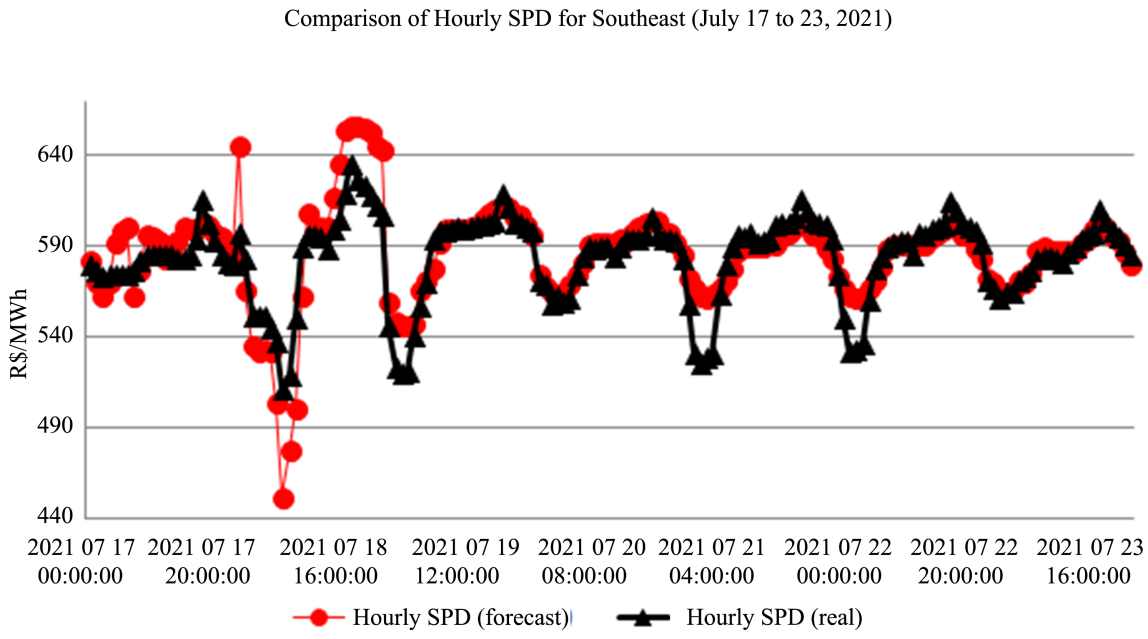

Figure 5. 7-days ahead forecast example, for Southeast submarket.

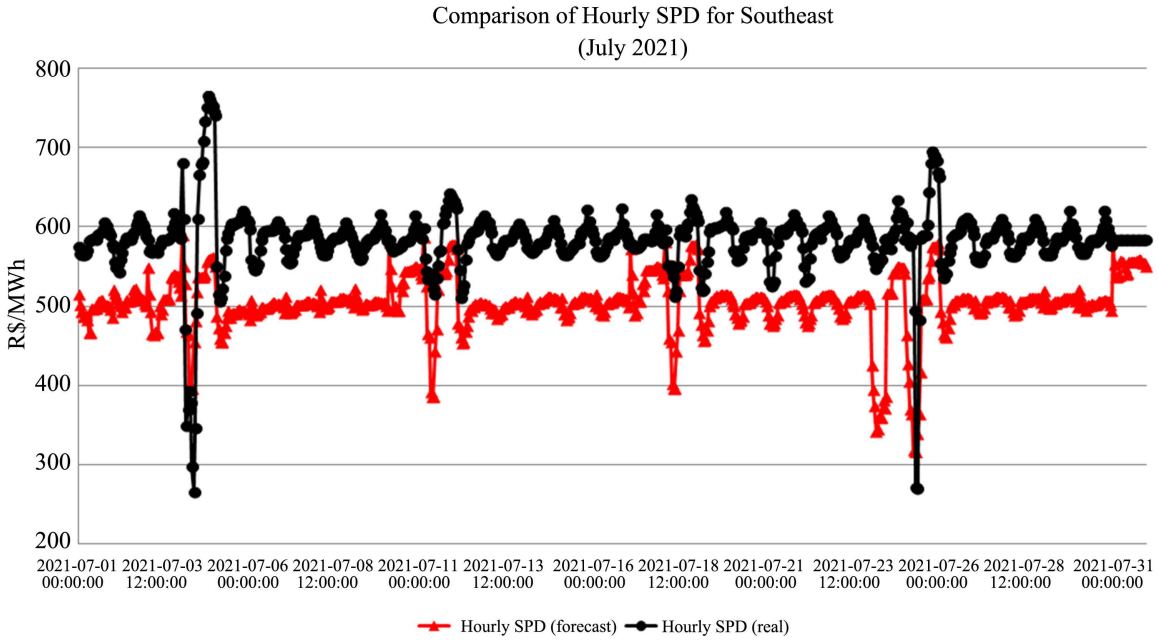

Figure 6. 1-month ahead forecast result, for Southeast.

until the actual disclosure of the SPD Week Level that will be in effect for the week in question. The results for all submarkets can be seen in Table 4.

4.6. Medium-Term Projection

The medium-term forecast starts from the end of the horizon defined by the short-term analysis; in other words, it starts in the third month and extends until the 14th month. It represents a prospective study carried out by CCEE, based on the Market Operator view of SPD Week Level for the horizon.

In the forecasting strategy without coupling short-term results, the Learning Base included only the actual values of the Hourly SPD, published at the time of this study.

In Figure 7, it is possible to visualize the behavior of the Hourly SPD along the period, specific for working days. For comparison purposes, the SPDs Week Level used in the input was included in the graphics.

In this forecasting strategy with the coupling of short-term results, the Learning Base incorporated the short-term predicted results, so the predicted results for the first and second months were incorporated into the model.

In Table 5, it is possible to visualize the effect of the Average variation of the Hourly SPD after this coupling.

Table 4. 1-month ahead forecast results.

	MAPE (%)	RMSE (R$/MWh)	TAI (%)
Southeast	14.00	87.27	60.08
South	13.00	81.97	46.64
Northeast	9.06	63.94	46.24
North	13.16	82.41	54.03

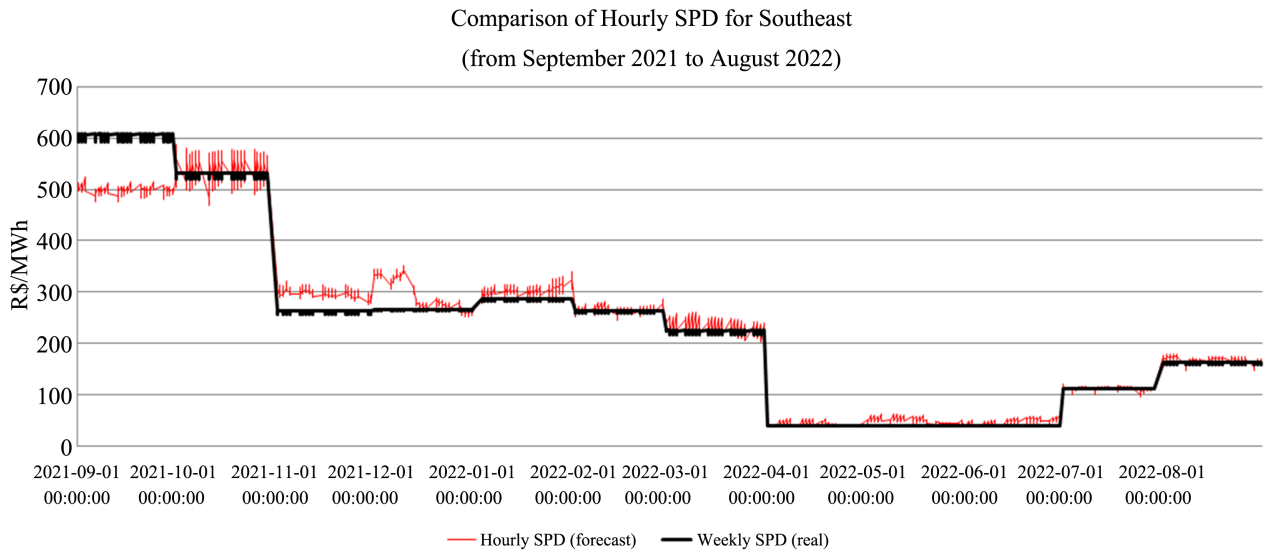

Figure 7. 12-month ahead forecast result, for Southeast.

Table 5. Hourly SPD with and without short-term coupling for Southeast.

Month	Average Hourly SPD with coupling (R$/MWh)	Average Hourly SPD without coupling (R$/MWh)	Variation (%)
Sep-21	573.95	500.85	14.60
Oct-21	528.11	547.98	−3.63
Nov-21	291.72	301.59	−3.27
Dec-21	284.59	293.05	−2.89
Jan-22	320	306.88	4.28
Feb-22	267.51	269.03	−0.56
Mar-22	238.6	236.2	1.02
Apr-22	42.63	41.93	1.67
May-22	45.65	48.12	−5.13
Jun-22	45.84	46.84	−2.13
Jul-22	110.21	111.2	−0.89
Aug-22	165.91	166.24	−0.20
Averages	242.89	239.16	0.24

4.7. Long-Term Projection

The long-term forecast starts at the end of the horizon defined by the medium-term analysis, that is, at the 15th month. However, it is possible to start the long-term horizon as early as the third month, thus dispensing the medium-term prospective study made available by the CCEE or another source.

Given the uncertainty inherent to this horizon, NEWAVE's official input files make available 2000 synthetic series, each representing a possible hydrological scenario, and with this, the series were grouped to create three possible scenarios for the SPD Week Level in each month of this horizon:

- Pessimistic scenario:

Comprised of 20% of the largest weighted average MCOs, calculated from the range between the 80th percentile and the highest MCO value found,

- Probable scenario:

Comprised of 60% of the weighted averages of MCOs, calculated from the range between the 20th percentile and the 80th percentile,

- Optimistic scenario:

Comprised of 20% of the lowest weighted average MCOs, calculated from the range between the 20th percentile and the lowest MCO value found.

As a result, Figure 8 presents the three scenarios for the projection of Hourly SPD. The Probable scenario was represented by the black lines, the Optimistic scenarios by the blue lines, and finally, the Pessimistic scenarios by the red lines.

The coupling technique, using medium-term horizon results, was also applied in the long-term study. And to understand the impact of this strategy, Table 6

Table 6. Variation of Hourly SPD with and without medium-term coupling for Southeast.

Month	variation of Probable scenario (%)	variation of Optimistic scenario (%)	variation of Pessimistic scenario (%)
Sep-22	4.59	3.53	2.55
Oct-22	12.94	14.13	−1.81
Nov-22	−1.78	−1.78	−2.15
Dec-22	1.65	1.65	0.83
Jan-23	−11.19	−11.19	2.92
Feb-23	4.79	4.79	0.78
Mar-23	1.31	1.31	1.69
Apr-23	1.12	1.12	3.49
May-23	7.64	7.64	1.80
Jun-23	5.01	5.01	4.08
Jul-23	−2.58	−2.58	0.30
Aug-23	−1.48	−1.48	−0.26
Sep-23	3.59	3.59	1.16
Oct-23	12.24	12.24	9.07
Nov-23	−1.61	−1.61	−0.83
Dec-23	4.02	4.02	1.88
Jan-24	−11.84	−11.84	−16.73
Feb-24	5.03	5.03	−6.45
Mar-24	2.85	2.85	2.37
Apr-24	−0.33	−0.33	4.55
May-24	8.17	8.17	1.72
Jun-24	5.62	5.62	1.52
Jul-24	−3.72	−3.72	−0.03
Aug-24	−0.92	−0.92	0.47
Sep-24	2.56	2.56	4.50
Oct-24	11.09	11.09	3.84
Nov-24	−0.07	−0.07	−1.78
Dec-24	2.02	2.02	0.86
Jan-25	−10.72	−10.72	−8.36
Feb-25	5.64	5.64	−4.74
Mar-25	2.57	2.57	−1.19
Apr-25	−0.79	−0.79	−4.60

Continued

May-25	8.91	8.91	6.12
Jun-25	4.71	4.71	1.44
Jul-25	−3.07	−3.07	0.03
Aug-25	−0.63	−0.63	1.42
Sep-25	2.47	2.47	7.44
Oct-25	12.31	12.31	−7.37
Nov-25	−1.12	−1.12	−2.96
Dec-25	−0.64	−0.64	−0.76
Averages	2.01	2.01	0.17

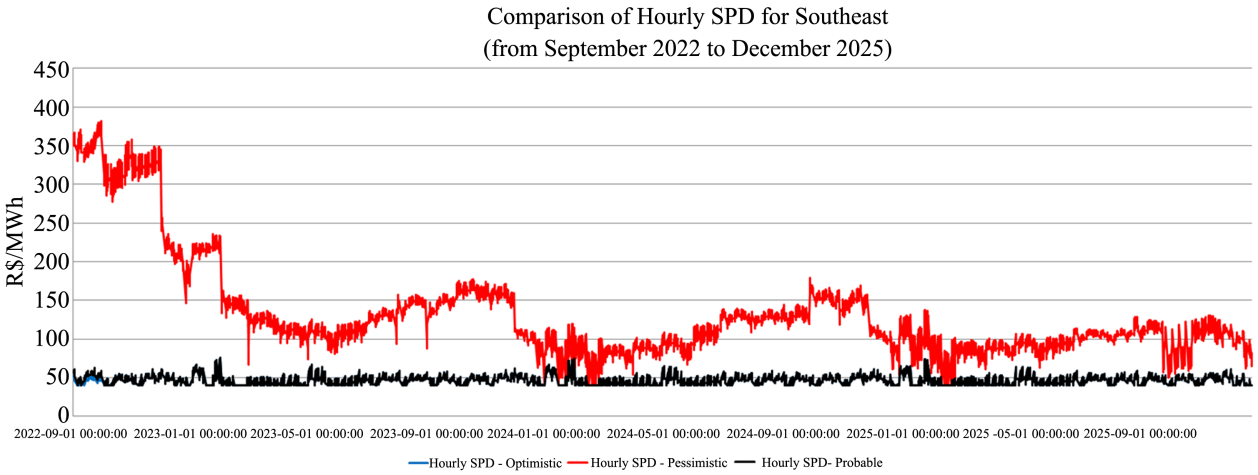

Figure 8. Long-term forecast for Southeast.

displays the differences between the average hourly SPD for a specific month with coupling, minus the average hourly SPD without coupling. Additionally, this strategy was applied to three scenarios created.

In both approaches, it is possible to see that the Probable and Optimistic scenarios are practically the same in most of the months of this horizon. This happens because of the low values of MCO, pointing to a well-known characteristic of the official models: optimism about the future inflows of the rivers. This optimism points to a favorable scenario of rains in the future, thus reducing the CMO by favoring the use of hydraulic sources in energy generation.

It is important to mention that if the values in both scenarios are lower than the floor allowed for the hourly SPD; this value must be discarded and replaced by this minimum value defined by CCEE.

5. Conclusions

This study presented the XGBoost model as an interesting tool in the projection of multivariate time series, with a relevant application in the hourly projection of

the SPD for the Brazilian electricity market.

This projection encompassed different forecast horizons and proved to be flexible when incorporating temporal and BSIN variables, dealing satisfactorily with tabular-type information.

As shown in this study, the official models have an optimistic bias in the medium and long term, so there is much to evolve in the strategy presented of basing the creation of scenarios on a group of series of the SPD Week Level. Potential new studies can be developed to anticipate this bias more consistently by including other variables that can translate potential water crises ahead, such as ANE and ES. Although the official models include these indicators in the SPD Week Level, the inclusion of specific variables for this purpose may be a promising field for further studies.

From a model robustness point of view, with the increase of samples inserted into the Learning Base, the XGBoost model will gradually become more comprehensive, serving as an important decision-making support tool.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Alim, M., Ye, G.-H., Guan, P., Huang, D.-S. and Zhou, B.-S. (2020) Comparison of ARIMA Model and XGBoost Model for Prediction of Human Brucellosis in Mainland China: A Time-Series Study. *British Medical Journal Open*, **10**, e039676. https://doi.org/10.1136/bmjopen-2020-039676

[2] Zhang, L., Bian, W., Qu, W., Tuo, L. and Wang, Y. (2021) Time Series Forecast of Sales Volume Based on XGBoost. *Journal of Physics*: *Conference Series*, **1873**, Article ID: 012067. https://doi.org/10.1088/1742-6596/1873/1/012067

[3] Grinsztajn, L., Oyallon, E. and Varoquaux, G. (2022) Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data? Cornell University, Ithaca.

[4] dos Santos, C.R.R., Castro, R., Marques, R.F. and Pereira, L.C. (2022) Aplicação de Aprendizado de Máquina para projeção do Preço Horário de Liquidação das Diferenças, como suporte às estratégias de comercialização de energia elétrica. *Revista Brasileira de Energia*, **28**, 243-279.

[5] ANEEL (2021) Sistema de Informação de Geração da ANEEL SIGA. https://bit.ly/2IGf4Q0

[6] Centro de Pesquisas de Energia Elétrica (2019) Manual de Referência: Modelo NEWAVE. Centro de Pesquisas de Energia Elétrica, Rio de Janeiro.

[7] Skiena, S.S. (2017) The Data Science Design Manual. Springer, New York.