Aalto-yliopisto
Aalto-universitetet
Aalto University

Master's Programme in Innovative Sustainable Energy Engineering

Assessing Electricity Prices and their driving mechanisms in Brazil with Neural Networks

---

Henrique Costabile

Master's thesis
2023

Aalto-yliopisto
Aalto-universitetet
Aalto University

Author Henrique Costabile

Title of thesis Assessing Electricity Prices and their driving mechanisms in Brazil with Neural Networks

Programme Innovative Sustainable Energy Engineering

Major Energy Systems

Thesis supervisor Prof. Ilkka Keppo

Thesis advisor(s) Túlio Carnelossi, M. Econ.

Collaborative partner XP Inc.

Date 18.08.2023 **Number of pages** 102 **Language** English

Abstract

In general, electricity prices are very volatile and derive from many external variables. In Brazil, this price is determined by computer models developed and operated by government organizations. The supply and demand relationships are not enough to determine prices in Brazilian submarkets. Due to the particularities of the predominance of hydroelectric production in the country and many regulatory factors, electricity prices in Brazil carry a high level of uncertainty to be managed by market participants. The Brazilian electricity Settlement Price is defined by the composition of three models: NEWAVE, DECOMP and DESSEM; for long-, mid- and short-term predictions, respectively. The prices are based on the Operational Marginal Cost, which those models aim to minimize especially by outputting the cheapest hydrothermal operational settings that can attend the electricity demand. To minimize the prices uncertainty, this research proposes investigating the feasibility of developing a predictive model supported by the time series Machine Learning technique, the Long Short-Term Memory (LSTM). This tool is part of a theoretical framework called Recurrent Neural Networks (RNNs). The raw material for this work is the combination of literature on the history of the Brazilian Energy Market and its particularities, in addition to studies on Neural Network technologies and LSTM applications, as well as real historical data related to electricity price in the country. Accordingly, this work compiles data from June 2001 to April 2023, weekly and by submarket, which represents the input variables of the proposed model. The product of this work revolves around a predictive model programmed in Python with support from the Keras library, capable of predicting 4 weekly prices ahead. In addition, a comparative analysis is registered between the results of the LSTM and DECOMP models, which is the one already widely used on the Brazilian market. For this evaluation, performance indicators were used on the assertiveness of the predicted absolute values, the direction of the predicted price, and the predicted volatility. The results show that the LSTM model was significantly more accurate with respect to direction and volatility and less accurate with respect to the absolute values of the predicted prices.

Keywords electricity;prices;Brazil;LSTM;neural;networks;prediction;model.

Acknowledgements

This Master thesis, and my graduation process behind it, would not be possible without multiple people that helped me. Along these lines, I would like to extend my gratitude to my family for their overall support on this journey. Moreover, I wish to highlight a special recognition to my girlfriend, Larissa Loverro, who moved from Brazil to Sweden with me. Obrigado linda!

Additionally, I appreciate my friends that helped me either with information, advice, support, or opportunities: Diego Henrique Lopes (Diegão), Pedro Casella, Guilherme Vilar (Gui), Philipe Gentile (Phil), João Amarante, Cristian Nogueira, Sylvio Saraiva (Sylvão), Bruno Morelli (Brunão), João Soares (Jones), Guilherme Passos (Gui), Matheus Felipe (Gomes), Nathan Luz, Matheus Eide, Ítalo Dias, Túlio Carnelossi (Tulião). Valeu demais!

Moreover, I would like to mention the names of fellow colleagues that I met in Aalto or KTH: Aravind Srinivasachari, Flora Lux, Gabriel Juul, Jaakko Eskola, Kneev Sharma, Lokesh Jayesh Pandya, Manuel Enrique Salas, Rumpee Bora, Zelal Bazancir. Thank you all.

Furthermore, I would like to express my appreciation to my KTH supervisor, Dr. Mohammad Reza Hesamzadeh, and Aalto supervisor, Prof. Ilkka Keppo, for their guidance and shared expertise throughout this research. In addition, thanks to Innovative Sustainable Energy Engineering (ISEE) program coordinators and staff, especially Nelli Markula and Prof. Francesco Fuso-Nerini, for their consistent availability to instruct towards academic affairs. Also, I would like to mention the institutions that made this program possible: Nordic Five Tech Alliance, KTH Royal Institute of Technology, Aalto University, Erasmus+ and XP Inc.

Finally, I hope this work can be insightful for you. Have fun with this reading.

Table of Contents

Table of Figures

Table of Abbreviations

ACL	**Ambiente de Contratação Livre -** Free Contracting Environment.
ACR	**Ambiente de Contratação Regulado -** Regulated Contracting Environment.
ANEEL	**Agência Nacional de Energia Elétrica -** Electric Energy National Agency.
ANN	Artifical Neural Networks
CCEE	**Câmara de Comercialização de Energia Elétrica -** Electric Energy Commercialization Chamber.
CEPEL	**Centro de Pesquisas de Energia Elétrica -** Electric Energy Research Centre.
CMO	**Custo Marginal de Operação -** Marginal Cost of Operation.
CMSE	**Comitê Monitoramento Setor Elétrico -** Electricity Sector Monitoring Committee.
DECOMP	Mid-Term Planning Operation prediction model.
EPE	**Empresa de Pesquisas Energéticas -** Energy Research Office.
FCF	Future Cost Function.
MAE	**Mercado Atacadista de Energia Elétrica -** Electric Power Wholesale Market.
MME	**Ministério de Minas e Energia -** Ministry of Mines and Energy.
NEWAVE	Long-Term Planning Operation prediction model.
ONS	**Operador Nacional do Sistema Elétrico -** Electric National System Operator.
PEN	**Planejamento Anual da Operação -** Annual Energy Operation Plan.
PLD	**Preço de Liquidação das Diferenças –** Price of settlement of the differences (referred only as Settlement Price in the report)
PMO	**Programação Mensal da Operação –** Monthly Energy Program Operation.
SDDP	Stochastic Dual Dynamic Programming
SIN	**Sistema Interligado Nacional -** National Interconnected System.

1 Introduction

The electricity market is relevant in most of the economic sectors of a country. Therefore, it is essential for the proper functioning and development of any economy. The past decades in Brazil have shown that reforms have been undergoing with a higher level of liberalization of the electricity sector [1].

The Brazilian energy market restructuring process proposed the emergence of a market for free negotiation of electricity contracts among generating agents, traders, and free consumers. In this market, the main instruments of negotiation are bilateral contracts whose terms define the amount of energy to be delivered, the contract term, and price [1].

In Brazil, there really is not a spot market. A spot market fulfills some functions such as the increase of the dynamism of negotiations; the balance of the relationship between generated and contracted energy; and acts as a reference for long-term contracts. In short, the spot market is an indispensable mechanism of balance between supply and demand [2]. In addition, there is a short-term market whose objective is to provide settlement for differences between the amounts of energy contracted and consumed. These settlements are parameterized in the Settlement Price (PLD, in Portuguese). The PLD is an output of the computational resources of the models managed by the Electric National System Operator (ONS, in Portuguese) and not necessarily a product of the relationship between supply and demand [2].

The system used by ONS is composed of the programs NEWAVE and DECOMP. The first is used for long-term operation planning, up to five years; the second is used as a short-term tool, up to twelve months. The fundamental objective is to minimize the marginal cost of operating in the Brazilian electrical system, as shown by Maceiral [3], [4].

The supply of electricity in the Brazilian market is closely related to hydroelectric power generation, given the current layout of the country's energy technology mix. This form of energy generation is strongly linked to the level of reservoirs and other stochastic variables, as demonstrated in [4]. The entire structure for determining the values of the PLD and the idiosyncratic characteristics of the energy mix brings high volatility to prices, leading market agents to invest in risk management alternatives for their portfolio of contracts. Therefore, there is great interest in models that help energy market agents in their decision processes.

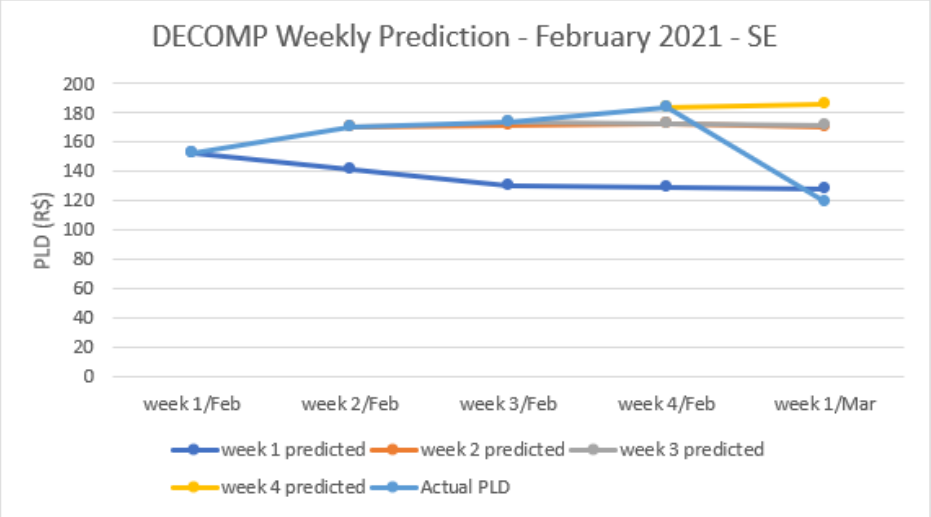
1.1 Motivation

The Settlement Price is used to value energy traded in the short-term market. Therefore, it is a strategic information for agents in the Brazilian electricity sector. The calculation is based on the Marginal Cost of Operation (CMO, in Portuguese), limited by a floor and a ceiling established annually by the Electric Energy National Agency (ANEEL, in Portuguese) [2].

As mentioned previously, the PLD is the product of processing the NEWAVE and DECOMP models, which require a large volume of input data and are often difficult to obtain, design, format, and compile on a computational platform. In addition to the current PLD for the week under study, the output of the processing cycle of these models presents a projection for the next periods until the end of that month, as demonstrated in [5].

With each weekly review conducted by the ONS, some input parameters of the DECOMP model are modified, such as the inflow forecasts and the starting volumes of the reservoirs [5]. This leads to

different projections from those predicted in the previous week. The graph in Figure 1-1 exemplifies the subject addressed so far. It is possible to analyze the evolution of both the real PLD and its projections over the highlighted period.



	week 1/Feb	week 2/Feb	week 3/Feb	week 4/Feb	week 1/Mar
week 1 predicted	152.1	141.4	130.64	129.65	127.84
week 2 predicted		170.45	171.48	172.56	170.77
week 3 predicted			173.37	173.11	171.71
week 4 predicted				184.15	186.14
Actual PLD	152.1	170.45	173.37	184.15	119.64

Figure 1-1 *- PLD variation in relation to projections - Southeast. Source: Author's elaboration with PLD data from [37].*

In this graph, it is possible to verify the behavior of the settlement price, as well as the forecast values for each week of this month and the first week of March. The predicted values for Weeks 2 to 4 are fictitious, only for explanation purposes. Predictions from Week 1 and the actual settlement prices were taken from the DECOMP model throughout the month of February 2021 in the Southeast/Midwest submarket.

Note the disparity between the predicted price for week 2 during week 1, represented by the dark blue line, and the effectively adopted price, represented by the light blue line. It is also noted that as the weeks advance, the forecasts continue to be strongly inaccurate in relation to the real prices.

The discrepancy between the projections and the PLD variation is a direct consequence of the revision of the input parameters of the model while going forward in time. The values considered for the independent variables in the processing of week 1 are updated throughout the month; that is, there is also a propagation error in the values that feed the model.

With the information presented and knowing the existence of the LSTM technique which can work well with finding non-evident relationships among variations that are affected by seasonality, there is a strong motivation to carry out experiments with it to find the settings that best fit the electricity settlement prices prediction in the Brazilian market.

1.2 Objective

The objective of this work is the development and evaluation of the usage of a Recurrent Neural Networks' method known as the Long Short-Term Memory (LSTM) model when it is used as a tool to forecast the Settlement Price (PLD) in the Brazilian short-term electricity market.

1.3 Methodology

Once the objective of the work has been established, the applied methodology can be presented as follows:

1. Research about the history and actual context of the Brazilian energy market.

2. Research on the processes and models used in planning the operation of the Brazilian electricity sector (NEWAVE and DECOMP).

3. Research into the Long short-term memory (LSTM) forecasting method in the field of Recurrent Neural Networks (RNNs).

4. Extract, organize, and process information that will serve as the input of the developed model.

5. Develop the addressed LSTM computational tool capable of automatically managing the entire procedure of consuming the input data, training its forecasting capability, and presenting results.

6. Verify the reliability of the results obtained, through a qualitative analysis and validation of the results with historical data and the outputs of the currently used programs.

7. Search for means that provide optimization of the techniques applied in the system, with the aim of improving their execution and efficiency; such means may include the best selection of input variables and the adjustment of the parameters of the neural networks applied.

This work is organized into this Introduction and eight more chapters. Chapter 2 reviews the existing literature on the behavior of the Brazilian electricity market and focuses on presenting the theoretical foundation that this work uses for the development of the proposed model. Chapter 3 presents the methodology used in this research for the development and evaluation of the model. Chapter 4 shows the results and their evaluation. Additionally, Chapter 5 refers to general discussion, Chapter 6 to conclusions, and Chapter 7 suggests potential future research. Finally, there are the References and Annex chapters, numbered 8 and 9, respectively.

2 Electricity Market in Brazil and Neural Networks

The objective of this chapter is to describe and characterize the object of study of this work – price of short-term electricity in the Brazilian market - in addition to presenting the theoretical fundamentals of this ecosystem and to emphasize the main characteristics of Artificial Neural Networks and their use for forecasting complex variables, mainly focused on LSTM techniques.

2.1 Electricity Market in Brazil

In this section, the main agents of the Brazilian electricity market and the main forms of energy commercialization were shown to be able to operate the short-term market. Finally, the objective is to clarify the particularities involved in the formation of the PLD.

2.1.1 Historical Context

The Brazilian electricity sector has undergone major changes since the 1970s. From there, it was defined that electricity tariffs should be governed by the so-called 'cost of service', that is, they should cover the costs associated with generation, transmission, distribution, and the remuneration at a level between 10% and 12% per year [1].

However, due to differences in generation and distribution costs in the various Brazilian regions, many companies began to show negative financial results, forcing the government to create a mechanism that would make it possible for tariffs to be equal between companies in the electricity sector. This tariff equalization mechanism determined that consumers in all regions of the country had to be overruled under the same tariff level in the same consumption class. Moreover, the legislation stated the transfer of resources from companies with positive balance sheets to those with deficits in their accounts [1].

In this verticalized and state-owned model, electric power companies were responsible for generation work, transmission, and distribution of electricity. Therefore, all activity related to electricity was a monopoly: consumers were captives, obliged to buy from just one company. Furthermore, the market was fully regulated and there were tariffs for all consumer segments: industrial, commercial, and residential, as discussed in [1] and [6].

This modus operandi persisted until the mid-1990s, when the sector began to show signs of stagnation. Public resources were drastically reduced and there was a need for measures to increase the supply of energy and revitalize the Brazilian electricity sector.

The federal government then extinguished the current tariff equalization and instituted supply contracts between generating agents and distributing agents, moment recognized as the beginning of the second wave of reforms of this market by Losekann in [6].

Subsequently, the government stimulated the participation of the private sector in the energy generation sector through the figure of the independent energy producer, allowing private companies to produce and sell electricity - an activity previously restricted to state-owned companies [6]. This regulation also established the figure of the free consumer, who could have the freedom to choose their electricity supplier [6].

2.1.2 Structure

In the late 1990s, the implementation of the restructuring plan for the Brazilian electricity sector came along with laws and decrees that created relevant entities, such as the regulatory body called Electric Energy National Agency (ANEEL, in Portuguese), the energy system operator - Electric National System Operator (ONS, in Portuguese), and an environment for carrying out activities of purchase and sales of energy, the Electric Power Wholesale Market (MAE, in Portuguese) as detailed in [7], [8], and [9], respectively.

In 2001, the electricity sector suffered a serious supply crisis [10], which had as a direct consequence a rationing plan and the realization of the need for a review to improve the current model. In 2002 a committee was created for this purpose, which at the end of the work published three reports suggesting changes in various segments of the electricity sector [6].

During 2003 and 2004, the federal government instituted a new model for the Brazilian electricity sector. Among the main changes in this period, there was the creation of new bodies to improve the Brazilian energy market: a company responsible for planning the long-term electricity sector, the Energy Research Office (EPE in Portuguese); an institution with the function of permanently evaluating the security of the electricity supply, the Electricity Sector Monitoring Committee (CMSE in Portuguese); and an institution to ensure continuity of the MAE activities, related to the commercialization of electric energy in an interconnected system, the Electric Energy Commercialization Chamber (CCEE in Portuguese), which is the one responsible for providing the PLD data to the market [7].

Figure 2-1, below, represents the institutions of the current model in the Brazilian electricity sector:

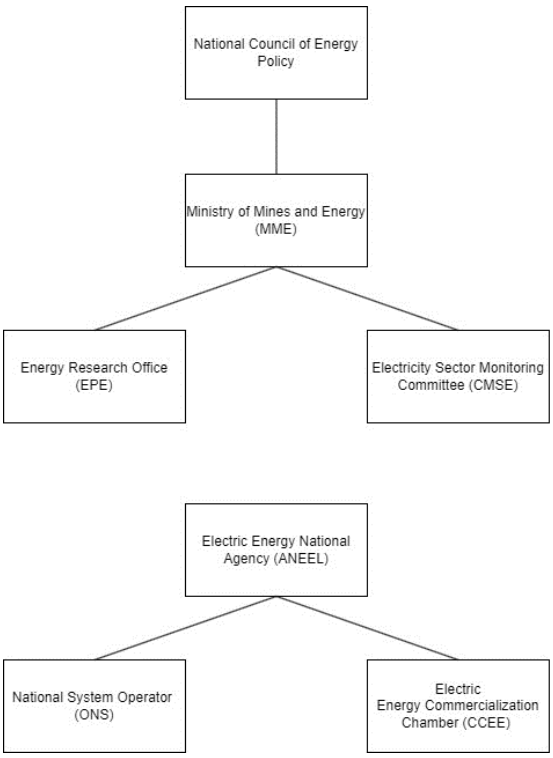


***Figure 2-1** - Current structure of the Brazilian electricity sector. Source: Author's elaboration with data from [7].*

Regarding the commercialization of energy, two environments were created for signing contracts for the purchase and sale of energy, the Regulated Contracting Environment (ACR in Portuguese), where agents of generation, commercialization and distribution of electric energy participate, and the Free Contracting Environment (ACL in Portuguese), involving generation and commercialization agents, energy importers and exporters, and free consumers [9].

The growing complexity of the Brazilian electricity sector in recent decades is remarkable. The model that was previously completely centralized, rigid, and controlled is now changing to a form where agents have more freedom in their decisions. However, the need to plan and monitor the operations of this sector becomes more evident.

The ONS mentioned above is the agent responsible for coordinating and controlling the operation of electricity generation and transmission facilities in the National Interconnected System (SIN in Portuguese). It is under the supervision and regulation of ANEEL and ONS duties include studying the annual electrical operation planning (PEN in Portuguese) and the monthly operation program (PMO in Portuguese) [8].

Certainly, ONS itself is not enough to be responsible for the management of this complex ecosystem. The National Energy Policy Council (CNPE in Portuguese) is an advisory body for the Presidency to develop energy policies and guidelines [7]. This organ periodically reviews the distribution of the country's energy technology mix and proposes targeting programs to develop and improve specific technologies and regions. The CNPE is chaired by the Minister of State of Mines and Energy, whose runs the Ministry of Mines and Energy (MME in Portuguese). The attributions of this Ministry include the formulation, planning and implementation of actions by the Federal Government within the scope of national energy policy [7].

Moreover, there are multiple categories of agents involved in the market that together create a balanced environment. First, there are the Power Generation Agents -- the ones authorized to operate in generation plants. These agents are segregated into three classes: Public services of generation, Independent Electric Energy Producer, and self-producers [7]. Public services hold concessions for the exploitation of energy-generating assets as the name suggests publicly. Then, Independent Producers are individuals or groups who receive concessions or authorization to produce energy intended for commercialization at their own risk. And self-producers are agents with permission to produce electricity for their own use exclusively, eventually selling surplus energy [7].

The second category is the Energy Transmission Agents, who have a concession to carry out the activity of electric energy transmission, through installations in the SIN [7]. Then, another category is Energy Distribution Agents, companies with permission to carry out energy distribution activities in a specific region. These systems must meet the energy demand of consumers with tariffs and conditions determined by ANEEL [7].

Finally, there are Energy Trading Agents whose role is to carry out the purchase and sale of electricity contracts among market participants. In this category there are traders, importers, exporters, free consumers, and special consumers. The traders participate in the energy market by entering bilateral purchase and sale contracts with other players. Energy importers are authorized to import electricity to supply the national market, while exporters are authorized to export electricity to supply neighboring countries. Free consumers can select the supply of energy through the free execution of contracts between energy traders of other agents and generators. The special consumer has a lower minimum demand to participate in electricity contracts and has permission to select suppliers whose energy sources originate from special incentives such as wind energy or solar energy [7].

Last, not least, it is relevant to mention the Captive consumers. This category includes consumers who cannot freely negotiate the supply of energy. These are restricted to the distributor access, who has the concession to operate in their region. The supply to these agents has the above-mentioned tariffs and conditions regulated by ANEEL [7].

2.1.3 Energy Technology Mix

Electrical energy represents approximately 20.8% of total industrial and 46.4% of total domestic energy consumption. Among all sectors, both together represent 50% of total energy consumption in Brazil, according to the EPE's National Energy Balance report of 2021 [11].

The demand for electricity has been increasing faster than the supply capacity. Consumption is expected to grow at a rate of about 3.0% per year, supported by a 2.9% Brazilian GDP annual growth until 2031, according to the EPE forecasting studies in [12] and [13]. Meeting this growth requires large investments in electricity generation, transmission, and distribution systems.

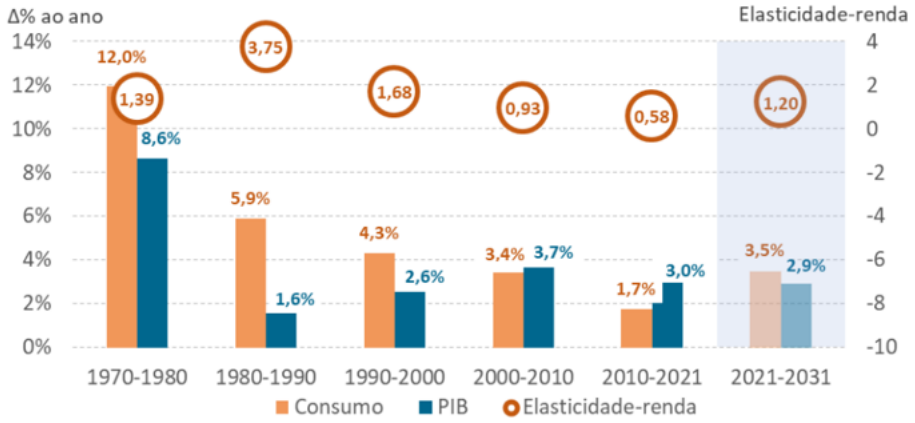


***Figure 2-2** – Income-elasticity in electricity consumption: History x Forecast. Source: [12]*

In addition to the high financial and social costs, the expansion of the electricity system inevitably causes damage to the environment, caused by waste emissions from the generation of thermoelectric plants and from the areas that are flooded when hydroelectric plant reservoirs are constructed [14].

Brazil is characterized by a hydrothermal electricity generation system, with a predominance of hydroelectric sources. As of May 2023, the Brazilian generating complex had a set of 1,481 hydroelectric projects, with a capacity of approximately 110 GW (57% of total) and 3,115 thermal projects with a capacity of approximately 46 GW (24% of total). Additionally, it has 3 nuclear power plants, 1,526 wind farms and 20,674 with combined capacities of around 36 GW (around 19% of the total) according to ANEEL's open-source data [15]. Therefore, Brazil has around 192 GW of total energy capacity and the surplus of energy consumed in the country comes from import contracts [15].

The following chart shows Brazil's energy technology mix. It includes Central Hydroelectric Power (CGH), Wind Power (EOL), Small Hydroelectric Power (PCH), Solar Photovoltaic Energy (UFV), Large Hydroelectric Power (UHE), Thermoelectric Fossil Power (UTE), and Nuclear Energy (UTN).

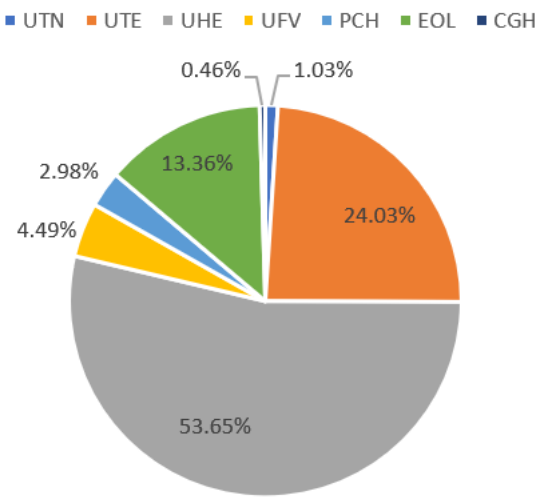


***Figure 2-3** - Brazilian Electricity Production by Source. Source: Author's elaboration with data from [15]*

The national electricity generation matrix differs from the average matrixes in the world (Figure 2-4). This fact, to some extent, makes it difficult to import planning and control models that already exist in other countries.

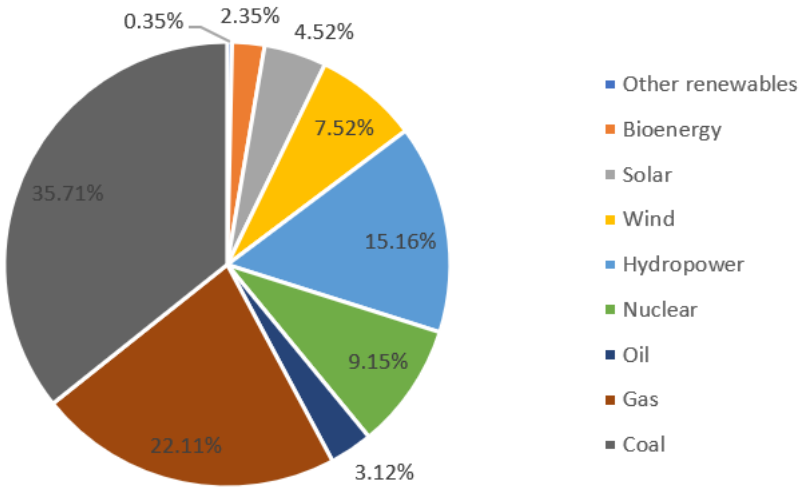


***Figure 2-4** – World Electricity Production by Source. Source: Author's elaboration with data from [38]*

Hydroelectric power plants are arranged in cascades along the course of rivers, which means that the consumption or retention of water in a dam directly affects the other generating units downstream. In addition, the use of water in a plant in each period implies a lower availability for the following period, which may lead to the use of other sources that are not of hydraulic origin [2] [5]. These characteristics demonstrate factors of spatial and temporal interdependence of the Brazilian electricity generation matrix.

The location of hydroelectric plants, normally far from large consumption centers, required the development of a complex transmission system, which is also used to import and export electricity.

The configuration of these transmission lines imposes restrictions on the flow of energy among the main regions and is the foundation of the National Interconnected System (SIN) [16].

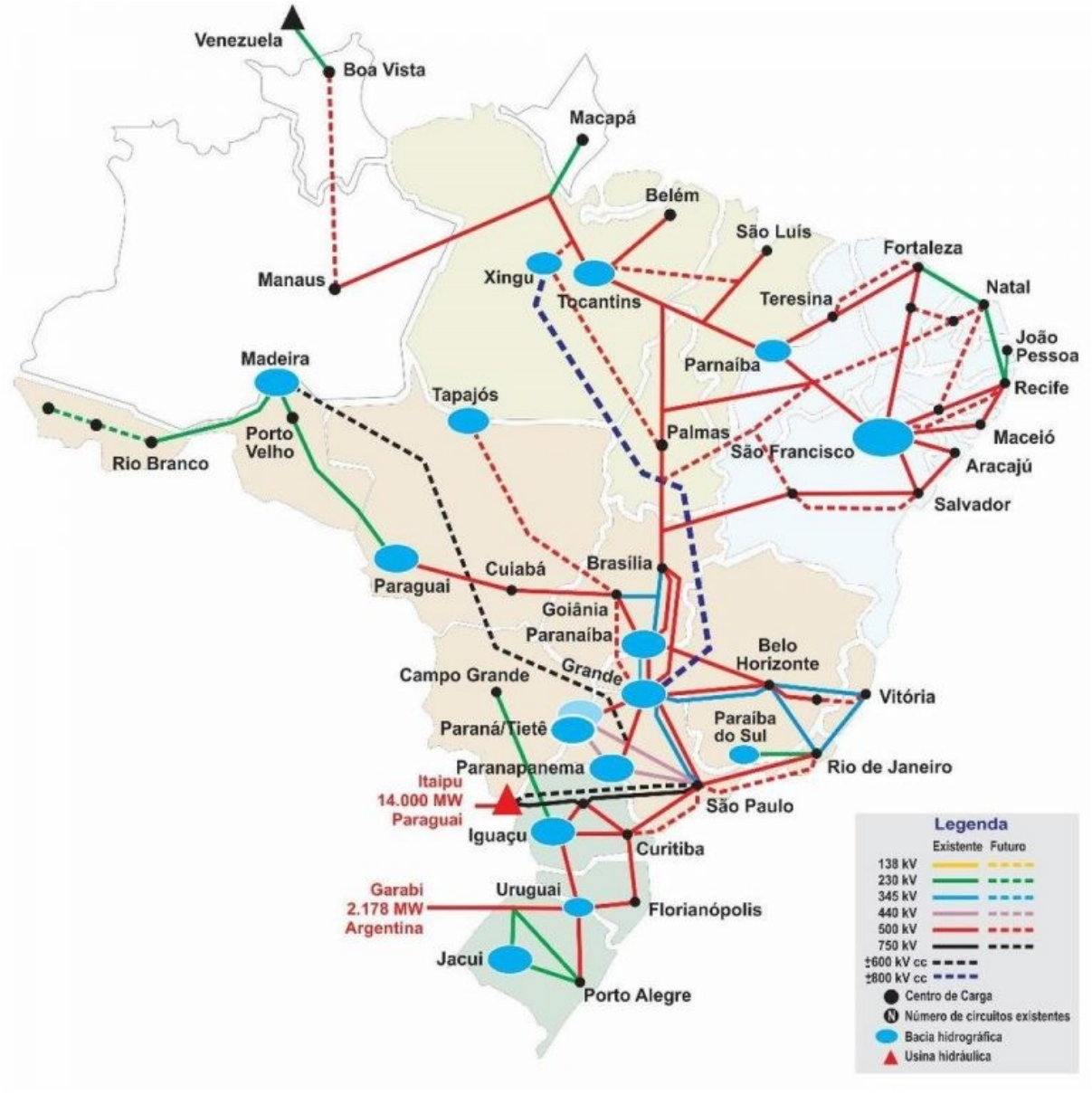


Figure 2-5 *- National Interconnected System – SIN (2021). Source: [16], English translation: "Centro de Carga": Load Center; "Número de circuitos existentes": Existent Circuit Number; "Bacia hidrográfica": Hydrographic Reservoir; "Usina Hidráulica": Hydroelectric Plant*

For reference, Figure 2-6 shows how was the Brazilian Interconnected System in 2007, with its neighbor countries in the map:

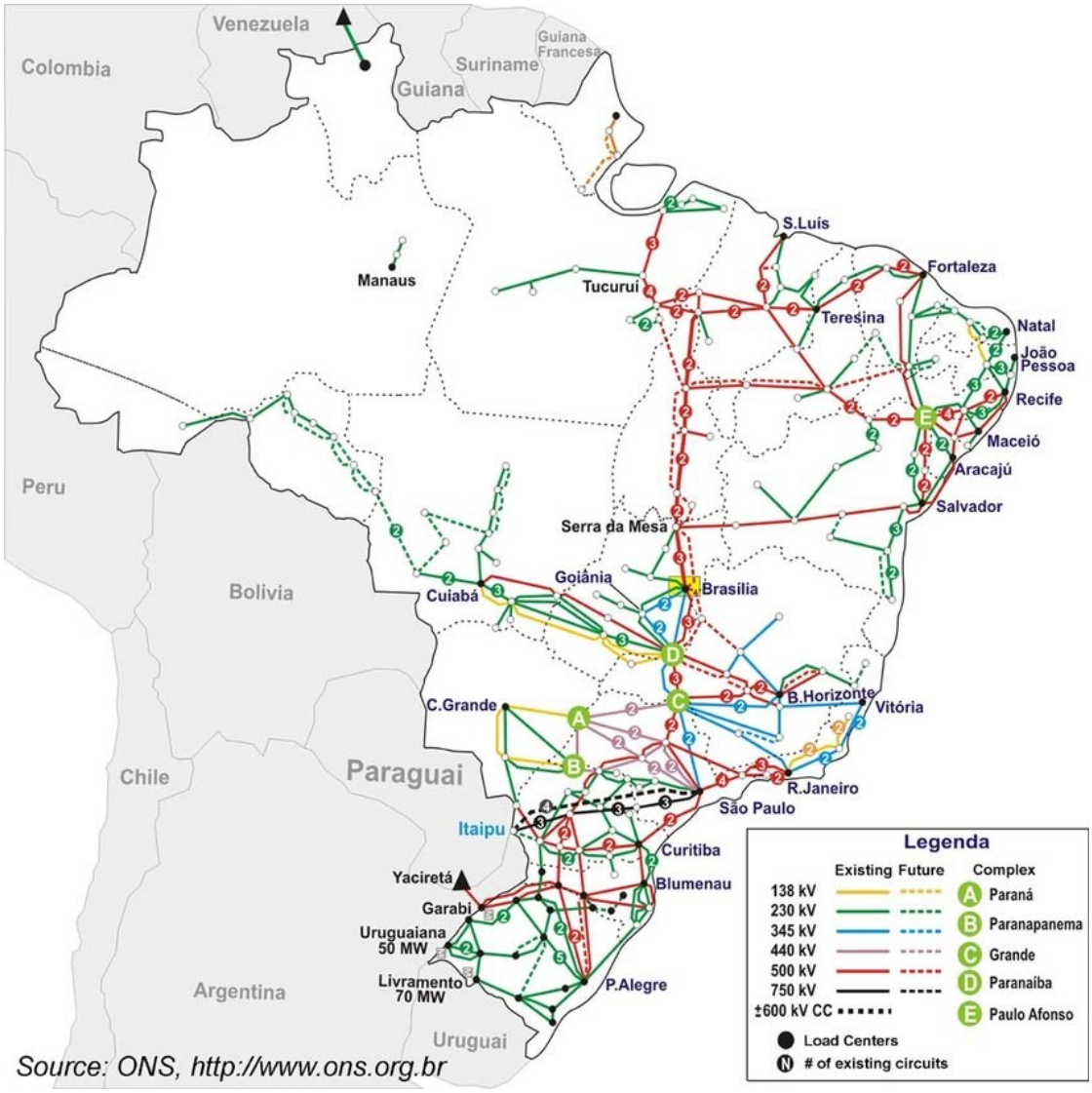


***Figure 2-6** - National Interconnected System – SIN (2007). Source: [16]*

The SIN is formed by companies from the South, Southeast, Midwest, Northeast, and part of the North region. Only 3.4% of the country's electricity production capacity is outside the SIN, in small isolated systems located mainly in the Amazon region, which borders with Guiana Francesa, Suriname, Guiana, Venezuela, Colombia, Peru and Bolivia [16].

The division into submarkets is based on the diversity of hydrological regimes (Figure 2-7) and their own characteristics [16] [17], which can be highlighted as follows:

• Southeast/Midwest: The largest consumer market in Brazil, imports energy from other regions for most of the year and has a high storage capacity located in multiple reservoirs.

• South: Stores with greater volatility throughout the year; Energy exchange with the Southeast region undergoes changes in direction throughout the year, but the tendency is to export energy.

• Northeast: It has an energy demand that has been increasing over the last few years; Part of the energy consumed comes from the Southeast and North regions.

• North: It exports energy during most of the year; this characteristic tends to increase due to new generation projects.

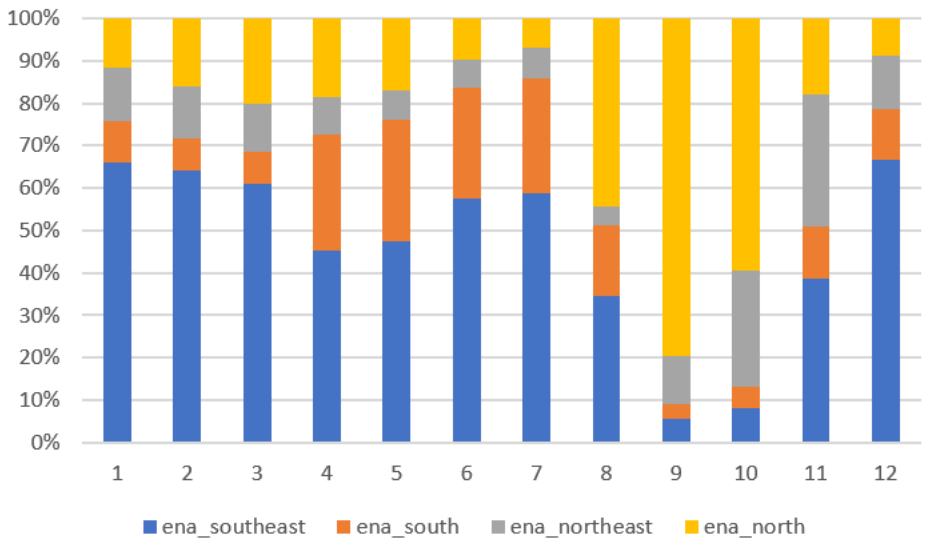


***Figure 2-7** - SIN Hydrological Behavior Diversity. Source: Author's elaboration with data from [39]*

The chart above presents the In-Flow Natural Energy (ENA in Portuguese) by submarket during the year, considering its monthly average from 2001 to 2023. ENA represents the potential energy to be generated from the inflow of the reservoirs in that region, which is more detailed in Section 3.1.2.4. The hydrological behavior of each region is directly reflected in this indicator, which is a factor of great importance in the study and planning of the operation. Because it is directly related to natural factors such as climate, temperature, amount, and frequency of rainfall, this flow goes through cyclical periods of humidity and drought. This seasonal behavior is difficult to predict and implies difficulties in the operation and control of the electric power generation system.

In summary, the Brazilian national electricity generation system can be defined with the following relevant characteristics: (1) it is predominantly hydroelectric; however, it has a large number of thermal plants that offset the growth in demand; (2) the expansion of the system to supply the demand is a challenge, since it represents high financial, social and environmental costs; (3) it is different from other world systems due to the high use of hydraulic sources for generation; (4) it has a complex transmission system with large geographical extensions, mainly due to the great distances between the generating centers and the main consumer centers; (5) it is characterized by several potential areas for hydroelectric use along the same rivers with reservoirs strongly dependent on inflow cycles; and (5) it is divided into four large regions with different hydrological and rainfall characteristics and regimes that were presented above.

Considering the characteristics presented, one can see the importance of studying methodologies for planning and controlling the operation of electric power systems that can provide maximum use of existing generating units, as well as optimal use of the resources involved in generation.

2.1.4 Operation Planning

Allied to natural difficulties, such as order of magnitude of the systems, high number of input variables, limited resources, and uncertainties, the planning of the operation of the Brazilian electric matrix must also contemplate the coordination of the operations of several companies to find a point of operation where the cost is as low as possible. The expected result is a sequence of decisions that seeks to find this optimal point, in addition to reliably meeting all electricity demand.

This type of planning is a large and highly complex problem. It is necessary to subdivide the problem into smaller steps, thus constituting what is called a planning chain. The main criterion for this subdivision is the size of the planning horizon, which must cover a period of years ahead in terms of daily scheduling and real-time monitoring of the operation. The following subsections will present the two parts of this process: (1) annual planning; and (2) the operation of the monthly planning of the energy systems.

2.1.4.1 Annual Planning of Energy Operation

The Annual Energy Operation Plan (PEN, in Portuguese) covers a five-year analysis horizon, from May of the first year to December of the fifth year of study, with monthly details being revised due to the holding of the Energy Auctions. The data and information used in the studies for the annual planning of the energy operation are also used in the processing of the medium-term model for updating the future cost function, within the scope of the development of the Monthly Energy Program Operation – PMO, which will be more detailed in the next Section 2.1.4.2. Over time, significant changes can arise in the load to be served, in the generation offer, in fuel availability, in the transmission work schedule, in the limits of interchange between subsystems, and other factors mentioned in Section 2.1.3 and by ONS in [18]. Thus, to ensure the use of up-to-date information, these data and information are reviewed periodically.

The objective of the PEN is to define what hydraulic and thermal generation portions define the optimal point of operation with minimal cost. Furthermore, it intends to perform the SIN to verify compliance with all the criteria and standards defined by ANEEL's Grid Procedures, such as attendance to energy demand [18].

The data used to process the model are sent by all agents involved (detailed in Section 2.1.2). Relevant part of the input data is shown in **Table 2-1**:

Data Description	Source	Updated
Initial storage forecast per reservoir	ONS	Monthly
Verified and predicted Inflow Natural Energy	ONS	Monthly
Hold volumes per reservoir	ONS	Yearly
Historical series of monthly average natural flows	ONS	Yearly
Values of consumptive uses of water and evaporation values	ONS	Yearly
Hydraulic operating restrictions of the harnesses	ONS	Yearly

Relationship and operation regime of existing international exchanges	ONS	Quarterly
Consolidated Forecast of Global Energy Load and Demand Load by level and by subsystem	ONS	Quarterly
Update for the first two months of the study horizon of the forecast of energy load and demand load	ONS	Monthly
Transmission limits among the various electrical areas of the SIN and the schedule of transmission works that impact these limits	ONS	Quarterly
Minimum generation for reasons of electrical reliability of thermoelectric plants	ONS	Quarterly
Calculated values of the equivalent rate of forced unavailability and the equivalent rate of scheduled unavailability	ONS	Yearly
Risk aversion curves and penalty for violating the risk aversion curve	ONS	Yearly
Discount rate to be used in models for calculating the present value of costs	ANEEL	Yearly
Deficit cost function	ANEEL	Yearly
Penalty for violation of multiple water use	ANEEL	Yearly
Information on the status of new SIN generation projects	ANEEL	Monthly
Dead volume filling schedule of new SIN reservoirs	ANEEL	Monthly

Expansion schedule of generating units of SIN plants	ANEEL	Monthly
Technical data of the new SIN generation projects	ANEEL	Monthly

***Table 2-1** - PEN Input Data. Source: Author's elaboration with data from [3], [4] and [18].*

The lines marked in grey contain data that is also used by this work (on a weekly basis) as input data for the proposed LSTM model, more detailed in Chapter 3. After processing, the result of the Annual Energy Operation Plan presents the following items: (1) analysis of marginal operating costs; (2) marginal benefits of interconnections; (3) risks of not meeting the energy load, with analysis of the depth and duration of the associated deficits; (4) estimates of the amounts of international exchanges; (5) estimates of thermal generation, which consider aspects of the SIN's electrical energy security, in order to subsidize the formation of minimum operational stocks and strategic fuel stocks, in view of the logistics of purchase, storage and distribution; (6) estimates of total operating costs; (7) estimates of exchanges between subsystems; (8) evolution of subsystem storage; (9) probabilities of violation of risk aversion curves; (10) balance of assured energies; and (11) recommendations for adapting the maintenance schedules of generating units to the results of the study, when necessary [18].

These assessments are presented to the Electricity Sector Monitoring Committee (CMSE) and the Energy Research Office (EPE) and provide the basis for making decisions regarding anticipation and/or implementation of generation and transmission undertakings with the objective of expanding the safety margin of the energy operation of the SIN.

It means that the PEN makes it possible to define actions to solve the problems identified in the study horizon, as well as to evaluate the benefit of new features in the operation of the system, in addition to indicating measures to mitigate risk and to overcome eventual schedule delays. In addition, it is also possible to indicate operational measures so that the operation meets the standards and criteria established in the Grid Procedures, as well as to identify electrical restrictions that prevent the adoption of energy policies that could ensure the lowest cost of the operation [18].

2.1.4.2 Monthly Planning of Energy Operation

The Monthly Energy Operation Program is also prepared by the ONS in a joint meeting with the relevant market agents, and it is revised on a weekly basis. Studies provide short-term electrical energy parameters and guidelines that optimize the use of generation and transmission resources in the National Interconnected System [19].

For short-term planning, the horizon is up to 12 months, and the objective at this stage is to define the generation targets for each power plant in the system, as well as the energy exchanges between each subsystem.

The study produces a weekly report on market recognition. The weeks included in the study, known as operative weeks, correspond to the period that begins at 00:00 AM on Saturday and ends at 24:00 AM on the following Friday [19]. For this study, data from the PEN is used and the system input data is updated weekly. After processing the model, the PMO presents the following information as a result [19]: (1) individualized generation dispatch, by load level and its weekly average value, of hydroelectric and thermoelectric plants; (2) reservoir storage target levels at the end of each operating week; (3)

turbinal and non-turbinal average spilled energy, by load level and their weekly average values: (4) operating balance of instantaneous demand load by subsystem, on a weekly basis; (5) conditions to meet the SIN demand load; (6) maintenance schedules for hydroelectric and thermoelectric generating units; (7) marginal operating costs, on a weekly basis, by subsystem and by load level; (8) energy balances by subsystems, on a weekly basis; (9) energy exchanges between subsystems, by load level and weekly average; and (10) international exchanges by load level and weekly average.

As a result, the above-mentioned outputs provide insights for planning guidelines to be followed by the executive bodies of the daily programming of the electro-energetic operation in the short-term period. It is worth mentioning at this point that one of the main specific results presented by the PMO is the Marginal Cost of Operation (CMO in Portuguese), item (7) in the previous list, which represents the minimum cost to meet an additional MW load in each region of the SIN after meeting all consumption [18]. The resource used to generate this extra MW is what defines the CMO, as can be seen in Figure 2-8:

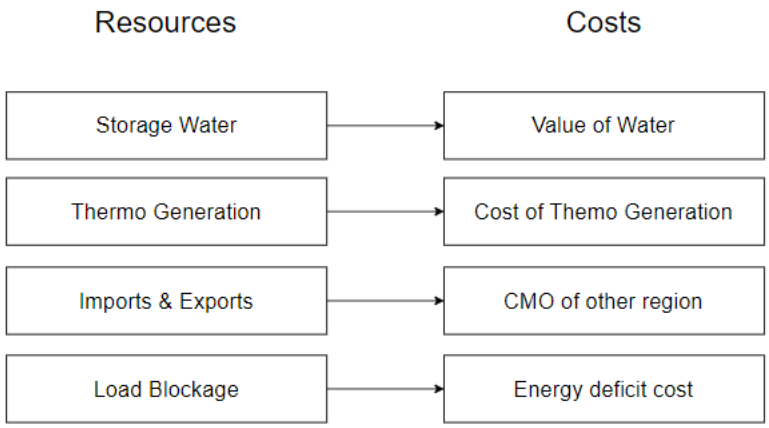


***Figure 2-8** - CMO training Source: Author's elaboration based on [18] and [19].*

Since the generation in Brazil is predominantly hydroelectric, most of the time the CMO is the value of the stored water. The CMO, limited by a ceiling and a floor defined annually by ANEEL, will define the Settlement Price (PLD), which is effectively the price of electricity in the short-term market and the target of prediction in this research.

In this way, the Brazilian market differs from other markets that have undergone a restructuring process in terms of short-term pricing. While in these markets the price stems from a balance between supply and demand, in Brazil it is a function of the CMO calculated in the energy optimization process. The PLD calculation process is detailed in the next section of this work.

2.1.4.3 Daily Planning of Energy Operation

The Daily Energy Operation Program (PDE in Portuguese) aims to ensure the optimization energy generation resources and the operational security of the SIN, establishing the programs load, generation, and exchange logbooks, based on the generation proposal defined by very short-term expectations. This activity is supported by the DESSEM model (Section 2.1.5.3), which was developed by CEPEL and operated by ONS [25].

2.1.5 Pricing

The price of electricity in Brazil is a product of the operation planning carried out by the ONS. One of the planning pillars is to define energy generation goals for each plant belonging to the SIN, with the objective of meeting energy demand and minimizing the cost of operation over the planning horizon.

Therefore, the dispatch of the plants is done in operating cost order, that is, plants with lower operating costs have dispatch preference. The operating cost is determined by some variables, among them the water level in the reservoirs of hydroelectric power plants, the cost of fuel for thermoelectric plants, and the cost of possible interruptions in the energy supply. In this dispatch criterion, as already mentioned in the previous section, the CMO is defined by the additional cost of one unit of energy in the SIN to meet a marginal demand [18].

In a predominantly hydrothermal system like the Brazilian one, the volume of reservoirs of hydroelectric plants is unknown during the planning horizon, given that it depends on future rainfall. This characteristic brings a level of uncertainty to the criterion of the operation planner, as it affects the expectation of future costs [5]. The flowchart below demonstrates the Operation Planner's decision criteria.

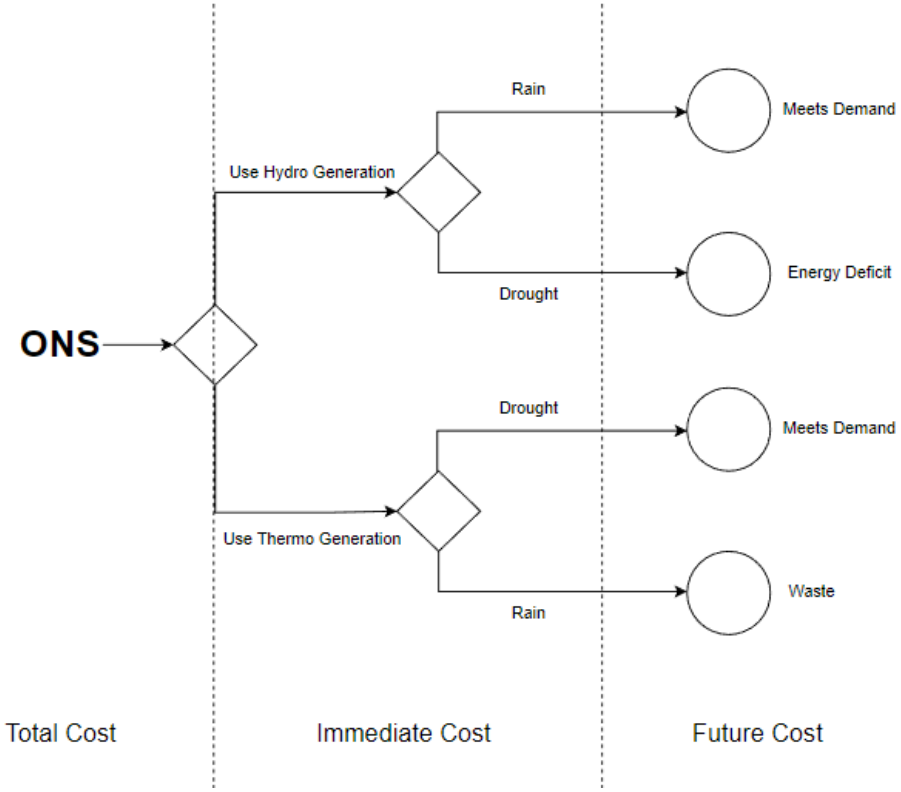


***Figure 2-9** - Operation Planning Flowchart. Source: Author's elaboration based on [5].*

2.1.5.1 Future, Immediate and Total Cost Functions

Decisions made by the operation planner have future consequences that must be considered, as shown in the previous Figure 2-9. At t_0, using available water leads to two future consequences: (1) if future inflows are high, an operation will be economic; (2) if the inflows are low, there will be a deficit in the energy supply. On the other hand, if the reservoirs are kept full at t_0, there are two possible future consequences: (1) if future inflows are low, the operation will occur economically; (2) if the

future inflows are high, the planner must pour the level of the reservoirs, which does not have the same damage as a deficit operation but implies energy waste [5].

Therefore, the operation planner considers the trade-off of the immediate generation utilization hydroelectric plant, given its low cost of production, with its future benefit of storage throughout the planning period. Analytically, the logic is reflected through expressions known as Immediate Cost Function (ICF), which represents the immediate benefit of using water resources, and Future Cost Function (FCF), which represents the benefit of saving the use of resources present for its future use [20].

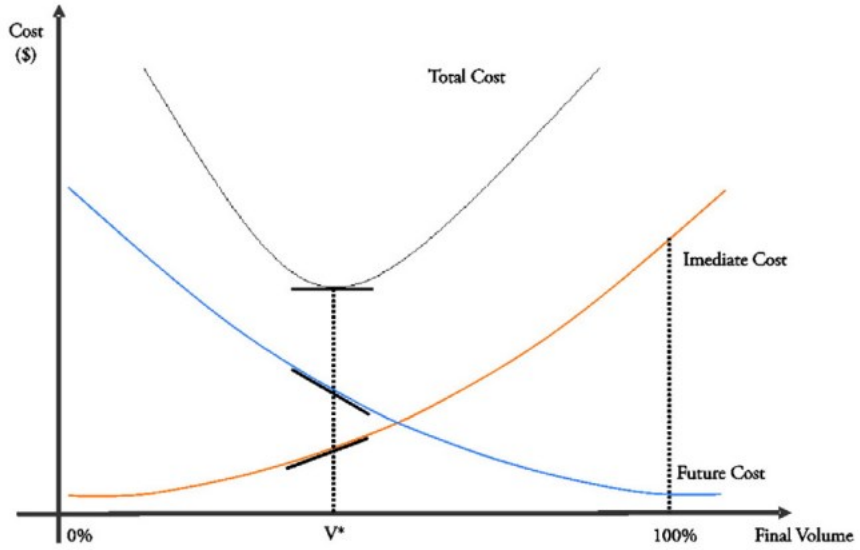


Figure 2-10 *- Total, Future, and Immediate Cost Functions. Source: [20]*

The ICF can be interpreted as the energy generation cost required to meet marginal demand at time *t*. The FCF at each time *t* is defined as the expected energy generation cost over the planning period, depending on the level of the reservoirs at the end of time *t*. As can be seen, the higher the level of reservoirs at the end of time *t,* the smaller the FCF value will be, given the greater availability of water resources at the end of the period. The planner will optimize the use of water resources by minimizing the sum of the cost of generating energy at the beginning versus the expected cost of generating energy at the end of the planned period. Together, these costs make up the so-called Total Cost [20].

Determining the Immediate Cost for a given period is essentially determining the cost related to the activation of certain plants and is directly related to the cost of the fuel and the technology of these plants. In Brazil, this cost is strongly related to the price of fuel consumed to dispatch thermal plants when the Value of Water is too high, represented by *V** in Figure 2-10 [20]. These values are reported to ANEEL and are available to ONS for planning studies. The ONS controls the energy dispatch of all generating agents, to maintain its objective of ensuring the supply of energy to all expected loads [7].

Furthermore, when demand is not met, there is also a cost related to rationing or lack of energy, called deficit cost. The deficit cost is a function determined by ANEEL [7] and is related to the depth of the load cut, as it is intuitive that smaller load cuts have less impact and can be easily managed, while deeper load cuts bring greater losses. Plant activation costs, and the deficit cost make up together the so-called Immediate Cost, as shown in the previous Figure 2-10.

Similarly, there is a portion of the Total Cost that reflects the future impact of the decisions taken. Since water is a renewable natural resource and depends solely on inflows, its Immediate Cost is zero [20]. After all, using the water stored in a reservoir to meet all the demand does not consume any fuel.

However, when the reservoir reaches its minimum level, thermoelectric sources will have to meet the demand throughout the filling of the dam volume [20]. This cost related to the decision to use exclusively water to supply the load is called Future Cost and can be plotted according to the graph in Figure 2-10.

It means that due to the predominance of hydroelectricity in the Brazilian system, as previously mentioned in Section 2.1.3, the Future Cost is influenced by decisions taken in the present towards the use of the available water.

Total Cost is then defined as the sum of Immediate Cost and Future Cost. When tracing its curve, it is noticeable that there is an optimal point of operation, where the Total Cost is minimum (Figure 2-10).

Consequently, the entire operating strategy can be summarized in making decisions in the present so that the target of the reservoir reaches the volume that guarantees the lowest Total Cost at the end of the month. This rationale is also supported by the PMO results, explained in the previous section 2.1.4.2.

An interesting fact observed is that the sum of the slopes of the Immediate Cost and Future Cost curves at the optimal operating point cancel each other out. That is, the sum of the derivatives of these curves is zero when the Total Cost is minimum. This means that the slope of the Future Cost curve varies as a function of the stored water volume. The derivative of this function is known as the Value of Water, already mentioned, and is represented in Figure 2-10 as *V** [20].

On the other hand, the derivative of the Immediate Cost Function (ICF) represents, in ascending order, the costs of thermal generation and the energy deficit. The slope of the curve for each volume reached at the end of the month represents the combined cost of thermal generation and the deficit needed to reach that stored volume (Figure 2-10).

Knowing the FCF and ICF, it is possible for ONS to define optimal energy dispatch, which corresponds to the lowest operating cost, by equaling the Value of Water to the generation cost of the most expensive thermal power plant being activated [20]. The definition of this optimal dispatch must respect the limits of transmission between the submarkets, the avoidance of the energy deficit, and the hydraulic and electrical restrictions to meet the demand for the period. These parameters are regulated by ANEEL, as explained in the previous Section 2.1.2.

In summary, the different possible scenarios of future inflows, which influence the level of reservoirs, attribute stochasticity and dynamism to the problem. The system operation planner, ONS, must use the FCF and the ICF to improve its responsibilities. In this way, the algorithm used must be able to represent stochastic optimization problems as detailed by Maceiral in [3] and [4] and more explored by this report in the following section.

2.1.5.2 Optimization

The optimization problem focuses on minimizing the Total Cost Function (TCF), which is expressed as the sum of ICF and FCF. The global minimum point is found where the derivative of TCF in relation to the reservoir level is equal to zero, that is, the point at which the derivatives of the functions of FCF and ICF in relation to the reservoir level are equal in magnitude [20]. In this context, optimization

models are used when there is a need to find the best solution for a problem that has a predetermined objective. Such models involve the determination of values for a set of decision parameters that will maximize or minimize an objective function, subject to restrictions [3] [4].

For the presented problem, it is possible to use Stochastic Dynamic Programming (SDP) to obtain an optimal operating point. SDP incorporates the randomness of natural phenomena into dynamic deterministic programming [3] [4]. For this, the water storage level in each dam is defined as a state and an interval of the period to be studied as a step.

The hypothesis is also considered that the storage level in any future state depends only on the current value of the water flow rate. That is, it is considered that the probabilities of the inflows into the reservoirs follow the Markov property [21], where the state of the system in any step depends only on the state of the system in the previous step and on the conditional probabilities calculated regressively.

Figure 2-11 show how SDP can be used to find the set of decisions that ensure that the operation of the system is optimal, as well as the Future Cost in each calculated state being optimal. From each state, the decision of the best cost is adopted. Following a procedure in the reverse direction of time, one arrives at the initial step, where the decision to be made and the total cost it entails are optimal.

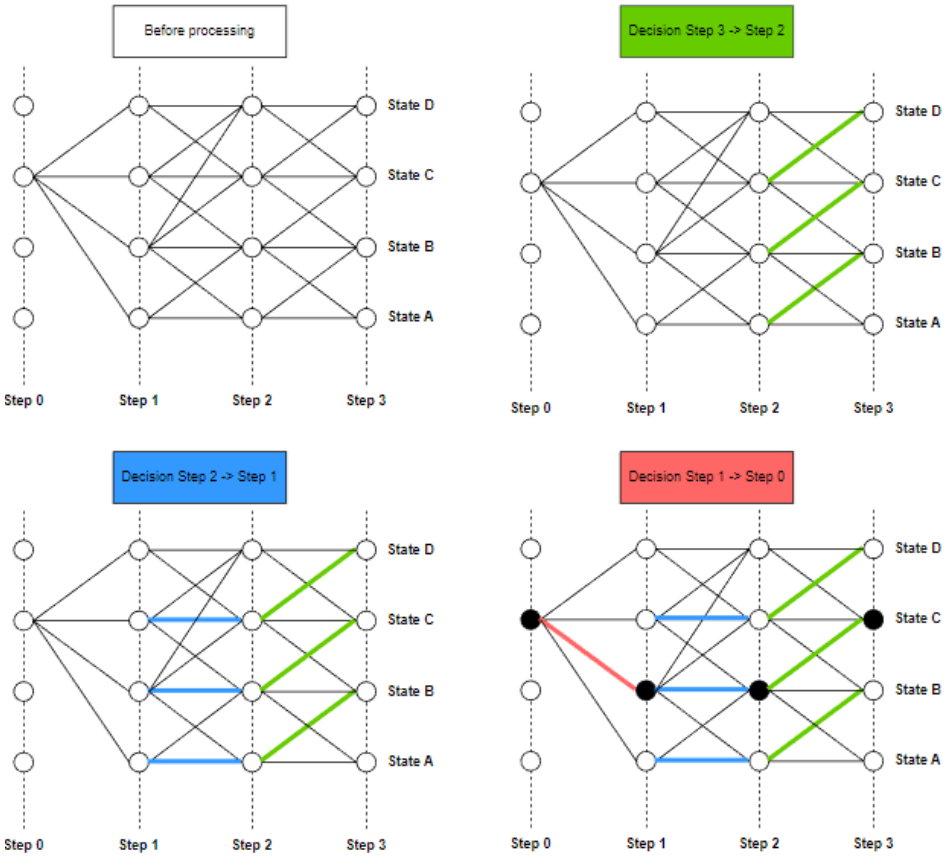


Figure 2-11 *- SDP Processing. Source: Author's elaboration*

The example in Figure 2-11 was built for just a single reservoir, over the course of a few steps. For the case of more reservoirs, the number of states grows exponentially. For example, for 100 storage levels, the calculation for two reservoirs would be 10^4 states, for three reservoirs 10^6 states, and for ten reservoirs it would be 10^{20} states.

This explosion in the number of states is known as the “curse of dimensionality”, first introduced by Bellman in [22] [29] and makes it difficult to use SDP in problems like this, in which it is necessary to plan the operation of multiple interdependent reservoirs. However, it is possible to get around this problem using a very small number of states, while maintaining the possibility of decent planning.

The reduction in the number of states occurs by dividing the original problem into sub-problems. Such a technique was introduced by the mathematician J. F. Benders in 1962 in his research available in [23] and it became known as Stochastic Dual Dynamic Programming (SDDP).

In compensation for the reduction in the number of states, it is now necessary to calculate the rate of future cost variation for every decision from the calculated state. This is possible by calculating the derivative of the future cost curve, also known as Value of Water, already mentioned in the previous Section 2.1.5.1.

The Future Cost Function, already mentioned earlier, can best be defined as a set of line inequalities capable of providing the future cost for any state researched. Through it, it is also possible to optimize each of the trajectories in each step. In contrast, there is the possibility of activating a thermal plant. With this scenario, it is better to save water and spend fuel, that is, the amount of water used for generation is reduced and the final level of storage at the end of the step increases.

At this point, it is interesting to highlight the fact that the balance between the Value of Water and the cost of generating from thermal plants, which is the derivative of the immediate cost curve, corresponds to the initial premise of minimizing the total cost, as also mentioned in Section 2.1.5.1.

In addition to meeting the future cost function aiming at minimizing the total cost, the final formulation of the problem that CEPEL models should optimize must also include the water balance, the supply of the load, the restrictions of the multiple use of the waters and the electrical restrictions.

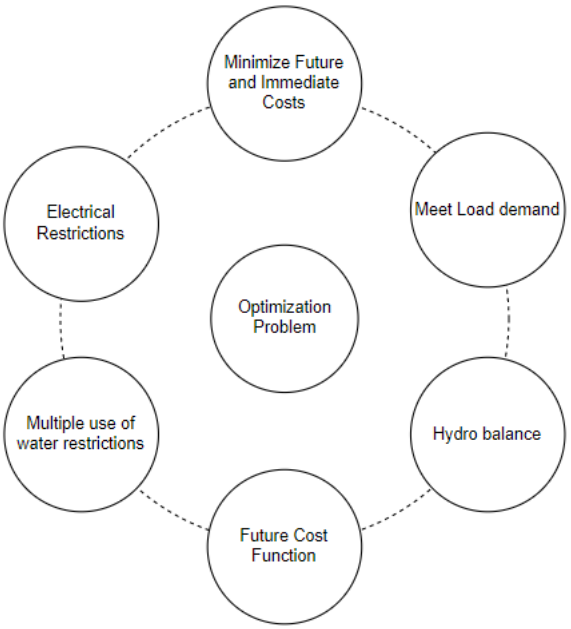


Figure 2-12 *- Overview of the optimization problem. Source: Elaboration of the author.*

The following equations mathematically represent the constraints related to the above-mentioned problem to be solved:

Future Cost Function

$$min\, FC(V) \begin{cases} FC(V) \geq FC(B)+(V-B)B' \\ FC(V) \geq FC(A)+(V-A)A' \end{cases} \quad \text{Equation 2.1}$$

Where A and B are states, V is the final volume of water storage of a step, and A' e B' are derivatives of the FCF in A and B, equivalent to the Value of Water in the respective states.

Hydro Balance

$$S_F = S_I + IF - G_H - W \quad \text{Equation 2.2}$$

Where S_F is the final level of water storage, S_I is the initial level, IF is the hydric inflow, G_H is hydro generation and W is the water waste.

Load Energy Attendance

$$G_H + G_T + I - E + D = L \quad \text{Equation 2.3}$$

Where G_H is hydro generation, G_T is thermoelectric generation, I is imports, E is exports, D is energy deficit and L is Load Energy.

Multiple Use of Water Restriction

Flood Control

$$s_F \leq s_{max} \quad \text{Equation 2.4}$$

Where S_F is the final level of water storage and S_{max} is the maximum level of water storage.

Minimum Outflow

$$G_H + W \geq OF_{min} \quad \text{Equation 2.5}$$

Where G_H is hydro generation, W is the water waste and OF_{min} is the minimum water flow that must be met.

External water usage (irrigation, supply, etc.)

$$IF_f = IF_g - EU \quad \text{Equation 2.6}$$

Where IF_f is final inflow, IF_g is gross inflow and EU is external water usage.

Electrical restrictions

$$Exc(x \rightarrow y) \leq F_{max}(x \rightarrow y) \quad \text{Equation 2.7}$$

Where Exc is exchange and Fmax is maximum flow.

Maximum generation by plant

$$G_i \leq G_{imax} \quad \text{Equation 2.8}$$

Where G_i is individual plant generation.

Maximum generation by group of plants

$$G_g \leq G_{gmax}$$ Equation 2.9

Where G_g in the generation of a group of plants.

These equations are presented to show which variables are relevant to the ONS decision-making process. Therefore, once operation planning have a direct impact on final electricity prices, these are also the variables that influence the price.

In summary, to solve the complexities that the system planner faces in his optimization problem, there is decomposition of the problem into smaller problems. Moreover, ONS uses different models for different period ranges to be studied. These models are called NEWAVE and DECOMP and were developed by the Electric Energy Research Centre (CEPEL in Portuguese) [3] [4]. Their iterative algorithms solve the optimization problem for different timeframes and complement each other. The operation of each of these scanning algorithms will not be detailed in this work but briefly explained in Section 2.1.5.3.

2.1.5.3 NEWAVE, DECOMP, and DESSEM Models

NEWAVE (Strategic Model of Hydrothermal Generation with Equivalent Subsystems) is an optimization program developed by CEPEL, which solves the problems of planning the interconnected operation of hydrothermal systems using the stochastic dual dynamic programming technique. It is used in medium-term operation planning. The main characteristics are planning with a horizon of five years or more, discretized monthly. From NEWAVE, the system planner sets the FCF to subsequently define the CMO of each SIN subsystem [3] [4].

NEWAVE represents thermoelectric generators individually and hydroelectric generators aggregated in equivalent energy reservoirs. That is, all the reservoirs in a submarket are grouped into a single reservoir as a single plant, whose generation capacity is equal to the sum of the generation capacities of all the hydroelectric plants that comprise that submarket [4].

Such simplification is convenient since studies involving NEWAVE have the main objective to obtain multiannual consumption attendance indexes among other information that will support decisions that involve, for example, plannings for generation and transmission expansions [4].

One of the main results obtained in studies with the NEWAVE model is the Future Cost Function, already mentioned in the previous Section 2.1.5.1. Through this function, coupling with the short-term model is convenient. This makes the short-term operating guidelines compatible with the medium and long-term operating policy [24].

Then comes DECOMP, which is the short-term planning model with weekly discretization. It was developed to optimize the operation of up to one month of horizon. The main inputs come from NEWAVE outputs [24].

DECOMP's function is to individually determine the generation goals of each plant, to meet the demand, and minimize the expected value of the operating cost throughout the planning period. The model is formulated as a linear programming problem, representing the physical characteristics and operating restrictions of hydroelectric plants individually [24]. That is, unlike the NEWAVE model, the DECOMP model now represents each plant individually.

It is important to point out that both NEWAVE and DECOMP are systems that depend on future operating scenarios. These scenarios are constructed from many variables, such as: future hydrological conditions, demand for energy, prices of fuels used in thermoelectric plants, energy deficit costs, entry of new energy generating plants in the SIN, availability of energy transmission between the different subsystems of the SIN, among other variables [4].

Both ONS and CCEE use the computational models NEWAVE and DECOMP; however, the objective of each institution is different. ONS conducts studies with the objective of finding the best sequence of operations that will supply demand safely and at the lowest possible cost. The CCEE, on the other hand, aims to calculate the Settlement Price, the PLD in Portuguese acronym, by load level and submarket. Due to these different objectives, the CCEE makes some changes to the input data received from the ONS [24]. NEWAVE data are updated monthly, while DECOMP data are updated weekly.

The changes made by the CCEE are [25]:

- Availability data from generating units in the testing phase are discarded.
- Data on internal operating restrictions for each submarket are removed.

DESSEM is a model for planning very short-term operation of hydrothermal systems (daily and hourly). The purpose of DESSEM is to determine the generation dispatch of hydroelectric and thermoelectric plants that minimizes the cost of operation during the planning period, given the most detailed information possible as input to the model, such as load forecasts, inflows, wind generation, availability, transmission limits between subsystems, Future Cost Function, among other inputs [41].

The following Figure 2-13 shows a diagram that represents the usage of these models and how they interact with each other.

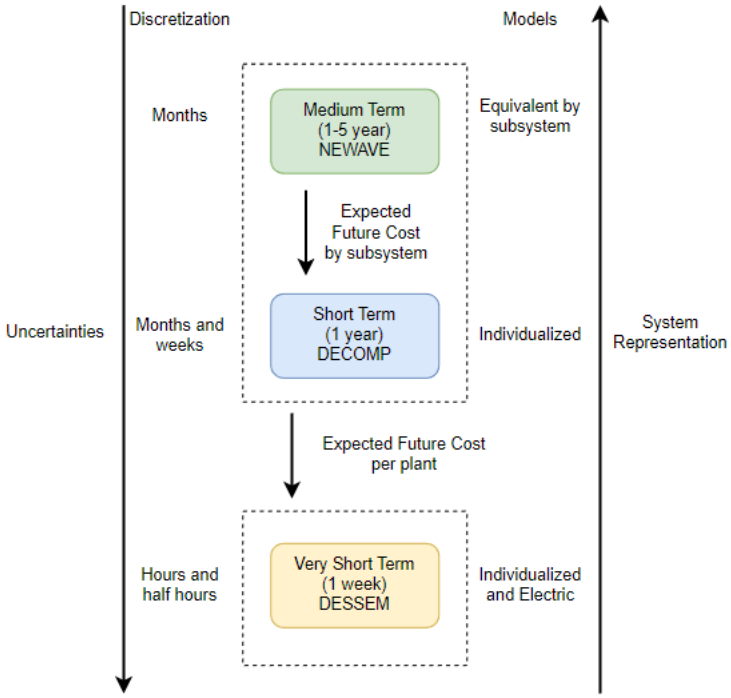


Figure 2-13 *– Brazilian Energy System Model's Chain. Source: Author's elaboration*

2.1.5.4 Settlement Price

Following the results of the Brazilian electricity market reform in 2004 [6], it was established that the CCEE would be responsible for the accounting of the total traded electricity in the SIN and to carry out the financial settlement of the amounts resulting from the purchase and sale of electricity in the short-term market [7].

To calculate the value of these amounts settled in the short-term market, CCEE uses the Settlement Price (PLD).

The weekly PLD is then defined for the first week of the planning horizon, considering load level, submarket, and its respective Operational Marginal Cost (CMO in Portuguese) as follows [25]:

$$PLD_{s,l,w} = min\left(max\left(CMO_{s,l,w}, PLD_{MIN}\right), PLD_{MAX}\right) \quad \text{Equation 2.10}$$

Where $PLD_{s,l,w}$ is the Settlement Price of the subsystem *s*, the load level *l* and the week *w*. Moreover, $CMO_{s,p,w}$ is the marginal operating cost calculated by DECOMP in a state to the subsystem *s*, the load level *l* and the week *w*. Then PLD_{MIN} is the minimum value that the PLD can assume and PLD_{MAX} is the maximum value that the PLD can reach. Both are defined annually by ANEEL.

Since 2021, CCEE has also started to make use of the hourly PLD, which is currently defined as being equal to the PLD of the threshold to which the hour belongs. With semi-hourly discretization for the first day and a planning horizon of up to 7 days, the very short-term stage aims to determine the daily schedule of the hydrothermal operation. This step considers the determination of variations of intermittent sources, the representation of operational constraints of thermoelectric units, and safety constraints [25] [41].

$$PLD_{Hs,j} = PLD_{s,p,w}, \forall\ j\ \in p, w \quad \text{Equation 2.11}$$

Where $PLD_{Hs,j}$ is the Hourly Settlement Prices for the subsystem *s* and the hour *j*. $PLD_{s,p,w}$ is the PLD of the subsystem *s*, the load level *l* and the week *w*.

The minimum PLD is calculated based on the estimate of the variable operating cost of the Itaipu Binacional hydroelectric power plant, considering the apportionment of the energy transferred from Paraguay to Brazil, valued by the daily geometric mean of the US dollar closing quotes, published by the Brazilian Central Bank (PTAX, in Portuguese reference acronym), in the period from December 1st of the previous year to November 30th of the calculation year [25].

On the other hand, the maximum PLD corresponds to the lowest value between the maximum PLD of the previous year corrected by the variation of the General Price Index - Internal Availability (IGP-DI in Portuguese) and the structural price of the most expensive thermoelectric plant, with installed capacity greater than 65 MW, included in the PMO (explained in Section 2.1.4.2) for the month of December of the current year, since the January value will only be available in the last week of December [25].

2.2 Neural Networks

In this section, the theoretical framework applied to the work is presented. First, it will show introductory concepts about Artificial Intelligence, later emphasizing the technique of Artificial Neural Networks and in particular the networks of the Long-Short Term Memory type (LSTM) that will be

applied to forecast electricity spot prices in the Brazilian market. Finally, the performance metrics that will evaluate the proposed models are explained.

2.2.1 Machine Learning

According to Norvig and Russell [26] Artificial Intelligence (AI) is a field of knowledge that addresses a series of subareas that aim to automate and systematize complex and intellectual tasks.

One of the possible resources to apply to AI is Machine Learning (ML), which represents a collection of techniques and algorithms based on statistics, algebra, and optimization, whose objective is to compile knowledge through data. Some widely used ML techniques are decision trees, random forests, Artificial Neural Networks (ANNs), among others [26]. In general terms, the different ML techniques seek to detect patterns from input information or past states to model systems and design future information as output. Accordingly, many algorithms are developed for the most diverse applications. These are mainly data-oriented and aggregate the main goals of predictive performance and process automation [27].

Machine Learning can be divided into three large groups: Supervised Learning, Unsupervised Learning, and Reinforcement Learning. Supervised learning models have the objective of making predictions based on data and the presence of uncertainty. These algorithms use a known sample source as input and output data to employ them for training and then design reasonable responses to new data. This category can be subdivided into classification algorithms and regression algorithms [27]. On the other hand, Unsupervised Learning does not have a learning variable response used to supervise the model, being widely used for classification [27]. Finally, Reinforcement Learning brings together the trial-and-error-based algorithms that are trained iteratively. Unlike Supervised Learning, this technique does not require the analysis by an expert to analyze input and output data and model adherence. In Reinforcement Learning, algorithms learn from their own interactions with the environment [27].

2.2.2 Artificial Neural Networks

Artificial Neural Networks (ANNs) were developed in the mid-1940s by mathematician Walter Pitts and neurophysiologist Warren McCulloch with the aim of proposing a computational system model whose functioning follows the fundamentals of biological neurons, capable of simulating synaptic connections using variable resistors and amplifiers [28].

In the field of ANNs, the first academic publications occurred in the mid-1960s. However, in 1990 the subject began to be widely studied with many applications in several areas [28]. Artificial Neural Networks are used in classification and regression problems. The greatest benefit generated by this technique is to capture the nonlinearity of complex problems involving time series [28]. Therefore, ANNs are capable of extracting information from large data samples, helping in solving complex problems, for example, forecasting asset prices.

ANNs are composed of interconnected neurons that simulate structured synapses, referenced in biological models. As a result, ANNs have important characteristics such as learning through training, fault tolerance, and the ability to generalize complex systems [28]. An artificial neuron can be analyzed as shown in Figure 2-14 below:

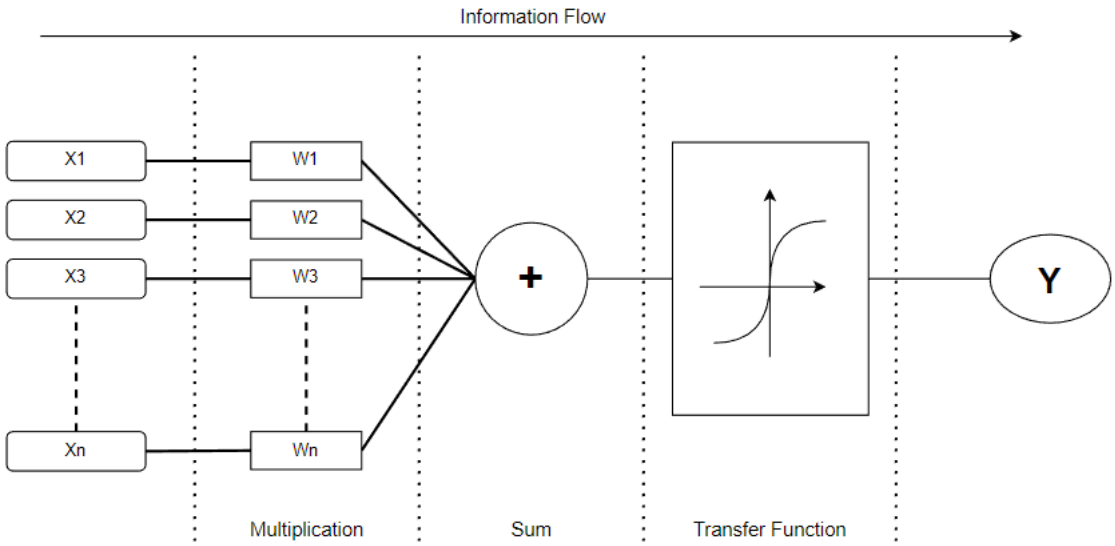


***Figure 2-14** – Working principle of an artificial neuron. Source: Author's elaboration based on [28]*

In the above illustration, X_n represents the input of the artificial neuron, and W_n represents the weights, specifically referred to as synaptic weights in the literature, that are assigned to each input data. The Sum in the center of the figure, also called the activation threshold, represents the combination line of the product between the weights and the input data, returning an output known as 'synthesized data'. The Transfer Function, or the Activation Function, transforms the synthesized data into the final output Y in a determined interval that can be assumed by the network [28].

The main Activation Functions used in the different ANN methods are the logistic function, the hyperbolic tangent function, the Gaussian function, and the linear function. Each function has a more appropriate application in relation to the type of problem in which ANN techniques are applied. The logistic function, defined in Equation *2.12*, will always assume inputs as real values between zero and one. The hyperbolic tangent function, defined in Equation *2.13*, will assume real values between -1 and 1. The Gaussian function, defined in Equation *2.14*, will assume symmetrical values for inputs that have the same deviation from the sample mean. Finally, the linear function, defined in Equation *2.15*, will assume the input exactly as they are output from the activation threshold [28].

$$g(u) = \frac{1}{1 + e^{-\beta \cdot u}} \qquad \text{Equation 2.12}$$

$$g(u) = \frac{1 - e^{-\beta \cdot u}}{1 + e^{-\beta \cdot u}} \qquad \text{Equation 2.13}$$

$$g(u) = e^{\frac{(u-c)^2}{2\sigma^2}} \qquad \text{Equation 2.14}$$

$$g(u) = max\ (0, u) \qquad \text{Equation 2.15}$$

Where β represents the slope of the Activation Function with respect to its inflection point; c and σ represent the center and standard deviation of the Gaussian function, respectively; u and g(u) represent the input and output of the Activation Function, respectively.

With the steps forward, ANNs can be subdivided into layers of neurons. These layers are classified as input, hidden, and output layers. The input layer is defined by the X_n inputs. In the hidden layers are the neurons that process network information. Finally, the output layer includes the neurons responsible for generating the output Y (referred to Figure 2-14) [28].

The architecture of the ANNs is what it is called, the different dispositions of the networks. It refers to the number of layers and neurons inserted in them [28]. The interconnections between artificial neurons within the network can vary according to the number of layers in the network, the number of neurons in each layer, the different activation functions existing in the different layers and the way the layers are connected, either in whole or in part [28]. The basic functioning of ANNs consists of learning the relationship between data through several interconnected neurons in different layers.

Over the years, different ANNs have been developed, with different architectures and forms of training, seeking the resolution of different problems found in multiple areas of study. Accordingly, some common applications are approximation of functions, process controlling, classification, and pattern recognition, data grouping, forecast models, and optimization of processes [28].

The main architectures of ANNs can be grouped into Simple Feed-Forward Networks, Multilayer Feed-Forward Networks, and Recurrent Networks [28].

Simple Feedforward Networks have only the input layer and an output layer, no hidden layers inside the network. On the other hand, Multilayer Feedforward Networks have one or more hidden layers. Feedforward topologies have data flow in a single direction, from input data to output data. Unlikely, the Recurrent Networks (RNNs) outputs have a feedback characteristic, that is, they serve as input data to other neurons. Recurrent Networks are applied to structures that vary over time, such as time series and dynamical systems [28].

For the construction of an ANN, the input data available is usually divided into two samples: approximately 60% to 90% of the original sample is separated for training and the rest is used to test and analyzing the trained model [29].

Specifically for supervised models applied to time series, which is the scope of this work, the training stage will serve to teach the network the history of the input data and its thresholds and synaptic weights. In this way, it can reach appropriate values with the answers inserted as the goals to be achieved. For the analysis and testing stage, the network will simulate the responses for the next model steps. From there it is possible to compare the values predicted by the network with the actual real data and verify the capacity of network forecasting using performance indicators, which will be more detailed in Section 2.2.4.

2.2.3 Long Short-Term Memory

Neural Networks called Long Short-Term Memory (LSTM) are in the group of RNNs architecture. These networks have the objective of learning complex patterns in structures with time dependence and multiple stages of processing. This means that LSTMs are good at learning from experiences that have time delays of unknown duration, which explains its name. It deals with decreasing the collateral effect

of ANN training of losing information from the beginning of the process when in mid- and end-phases, due to several interactions in their hidden layers [30].

Therefore, LSTM networks have the same properties as conventional RNNs; however, they can store information for long periods of time when processing a temporal sequence. A key feature of LSTMs is that they prevent backpropagated errors from vanishing or exploding, which can be an issue with traditional RNNs. Instead, errors can flow backwards through an unlimited number of "virtual layers" unfolded in space. That is why LSTM networks are very good at capturing long-term dependencies in time-series data [30].

For a more detailed view of its functioning, the memory points of an LSTM network are called cells. Cells can carry information to the end of a sequence or identify information that the network must forget after some processing step [30]. A LSTM structure is described by the following Figure 2-15 and its operation by the equations below.

$$f_t = \sigma\big(w_f \cdot [h_{t-1}, x_t] + b_f\big) \quad \text{Equation 2.16}$$

$$i_t = \sigma(w_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{Equation 2.17}$$

$$\dot{C}_t = tanh(w_c \cdot [h_{t-1}, x_t] + b_c) \quad \text{Equation 2.18}$$

$$C_t = f_t \cdot C_{t-1} + i_t . \tilde{C}_t \quad \text{Equation 2.19}$$

$$o_t = \sigma(w_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{Equation 2.20}$$

$$h_t = o_t \cdot tanh\,(C_t) \quad \text{Equation 2.21}$$

Where x_t is the network input at time t; h_t is the cell output at time t; σ denotes the logistic activation function, C_t denotes the state of the cell at time t; and $\dot{C}_t$ represents the candidate for the state at time t. Additionally, there are 3 gates in LSTM cells which are the forgetting f_t, input gate i_t, and output gate o_t. The constants w_f, w_i, w_o and w_c are the weights of the forgetting, entry, exit and cell gates, respectively. The constants b_f, b_i, b_o represent the thresholds of the forget, input and output gates, respectively. Finally, b_c represents the cell state [30].

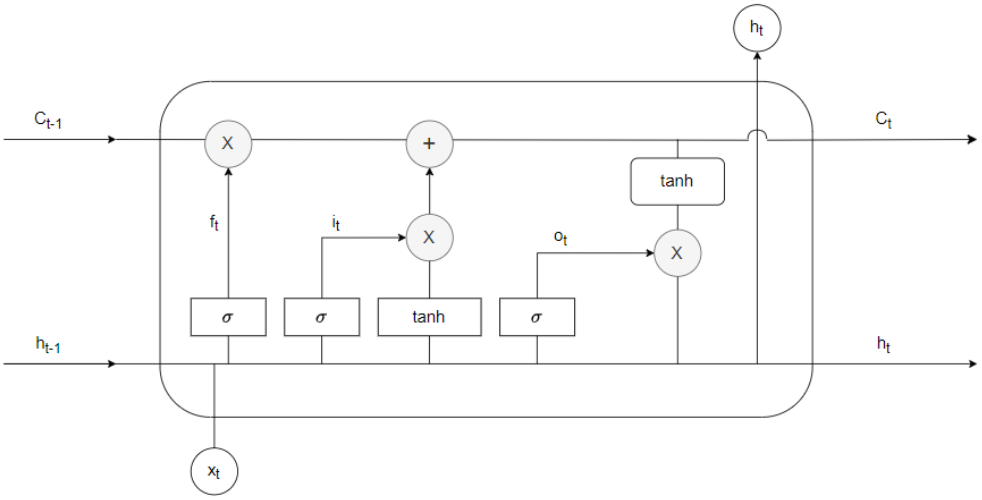


Figure 2-15 *- Representation of an LSTM cell. Source: Author's elaboration based on [30]*

Whether the input information will be memorized in the cell or not is a decision made by the input gate. The output gate defines whether the information will be discarded at time t. The state of processing will be memorized in the cell and the output data will be processed by the output gate. Through this cell design, LSTM networks can learn long-term dependencies from temporal data. In general, LSTM networks have good results for temporal forecasts [30]. Figure 2-16 demonstrates the temporal structure of LSTM networks.

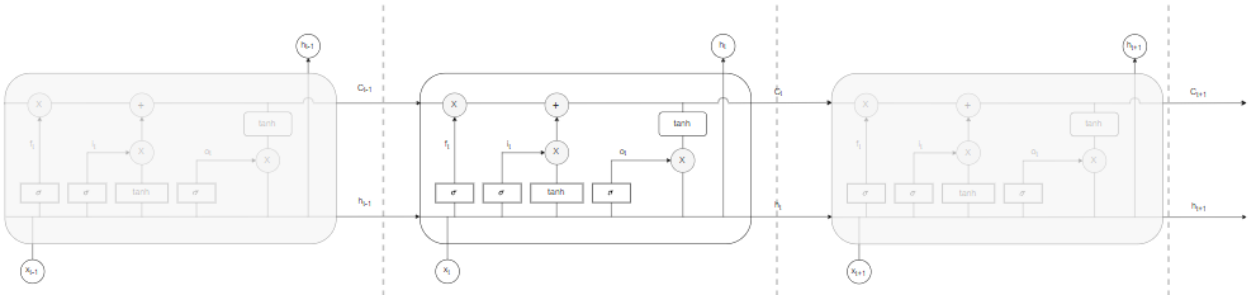


Figure 2-16 *- LSTM replicated over time. Source: Author's elaboration based on [30]*

To summarize, the structure of an LSTM unit is as follows.

1. **Forget Gate**: The forget gate's job is to determine how much of the previous state should be forgotten. This decision is made based on the current input and the previous hidden state. The forget gate is a sigmoid function.
2. **Input Gate**: The input gate updates the cell state with new information. It has two parts. A sigmoid layer, called the "input gate layer", decides which values to update, and a *tanh* layer creates new candidate values that could be added to the state.
3. **Cell State**: This can be thought of as the "memory" of the LSTM unit. It is updated with the information decided on by the forget and input gates. The cell state runs through the entire chain, with only minor linear interactions.
4. **Output Gate**: Finally, the output gate decides what the next hidden state should be. This output will be based on our cell state but will be a filtered version. The hidden state is computed based on the current input, the previously hidden state, and the current cell state.

The above processes result in a system that can learn and represent long-term dependencies in the data. It is particularly useful for applications like machine translation, speech recognition, and time series forecasting. Accordingly, LSTM (Long Short-Term Memory) networks have shown a lot of promise in time series forecasting, particularly due to their ability to capture long-term dependencies, handle variable-length input sequences, and manage data with complex temporal behaviors.

The following are some of the key advantages and applications of LSTM in time series forecasting.

1. **Capturing Long-Term Dependencies**: LSTMs are explicitly designed to avoid the long-term dependency problem. Recalling information for long periods is practically their default behavior, not something they struggle to learn. This makes them excellent for handling data where the temporal dependencies span various lengths of time.
2. **Sequence-to-Sequence Forecasting**: One powerful application of LSTMs is in sequence-to-sequence forecasting. This is particularly useful in scenarios where the output is a sequence of data points (such as predicting the next sequence of weather patterns, stock prices, etc.) instead of a single data point.
3. **Multivariate Time-Series**: LSTMs can model complex relationships across multiple input variables or series (also known as multivariate time series). This is useful in scenarios where multiple measures are recorded over time and the interactions between these measures could impact future forecasting.
4. **Anomaly Detection**: LSTMs can be used in time series anomaly detection, which is important for detecting fraud, managing system health, and identifying outliers in any other context. They can learn a 'normal' pattern from historical data and then identify any deviations from this normal pattern as anomalies.

However, despite these advantages, LSTM models can be quite complex and may require significant computational resources and time to train, especially for larger datasets. They can also be more difficult to interpret compared to simpler, more traditional time series models, such as ARIMA.

Like all models, it is important to consider the specific characteristics of the task at hand when deciding whether to use an LSTM. Depending on the situation, simpler statistical methods, traditional machine learning methods, or other types of neural networks might be more appropriate. But in the right situations, LSTMs can be a powerful tool for time series forecasting.

2.2.4 Performance Indicators

The purpose of regression models is to predict numerical values for the problems studied. To quantify the performance of Neural Networks in time series problems, it is common to use metrics that compare series of projected values and series of values real. The main performance metrics used in the literature will be described in the following.

It is not possible to say that there are superior and inferior metrics, but some are more used than others. Most are widespread in the field of statistics and have unknown authors. This section will highlight the metrics investigated.

2.2.4.1 Root Mean Squared Error - RMSE

The Root Mean Squared Error (RMSE) indicates the size of the average error obtained between the projected and real series [31]. Its calculation is defined in Equation 2.22.

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(x_t - \tilde{x}_t)^2} \qquad \textit{Equation 2.22}$$

Where x_t represents the actual value of the series at time t, $\tilde{x}_t$ represents the projected value at time t, and *N* represents the total length of the sample.

2.2.4.2 Trend Direction Accuracy Measurement

The Trend Direction Accuracy Measurement (TDAM) is the indicator that will check whether the predicted values follow the same trend direction as the actual values. It is represented as the percentage of the total predicted values that corresponded to the direction of their respective real values.

$$TDAM = \frac{R}{T} \cdot 100\ [\%] \qquad \textit{Equation 2.23}$$

Where *R* represents the number of predicted values of the correct trend and *T* represents the total amount of predicted values.

3 Model Application

The approach to the problem was made through quantitative research, because in this work data collection and treatment procedures were carried out. According to Roger Watson in [32] the quantitative research approach is characterized by the quantification of processes, from collection and treatment of information, through simple statistical techniques to the most sophisticated.

Quantitative research proves to be adequate when there is a need to measure variables and to make inferences from sample data of a population through specific instruments. This research method uses numerical evidence to analyze the models and hypotheses proposed in the knowledge construction process [32].

According to the objectives described in Section 1.2, the neural network architecture used for projecting energy prices is the Long Short-Term Memory (LSTM). This technique has a convenient framework for problems involving time series analysis and forecasting, and its deep learning characteristics deal appropriately with problems with many variables.

Predicting the price of electricity can be a complex task due to the numerous factors that affect it, such as weather conditions, time (hourly, daily, monthly patterns), demand and supply conditions, and other economic indicators. However, the LSTM model can be a good choice because of its ability to capture long-term dependencies in time series data.

Therefore, the objective of the LSTM model developed in this work is that the network can adequately design the different PLD series by subsystem, proving to be a useful technique in the decision making of the different agents belonging to the Brazilian electricity market.

This chapter describes the methodology used in this work, which consists of five steps referenced by specialists in [33], [34], [35] and [36]:

- **Data Collection**

The collection of the historical data of the subject of study, also called the target variable or dependent variable, and all the factors that present a high correlation with its volatility, also known as independent variables.

- **Data Preprocessing**

This involves several steps:

  - Check for missing values: If there are missing values in the collected data, handling them is needed. This can be done by deleting those time periods or by inputting the missing values using a method such as linear interpolation.
  - Normalization: LSTM models are sensitive to the scale of input data. It is a common practice to rescale the data in the range of 0 to 1.
  - Sequence creation: LSTMs expect data to be in a specific format, usually a 3-dimensional array. The three dimensions of this array are:
    - Samples: One sequence is one sample. A batch contains one or more samples.
    - Time Steps: One time step is one point of observation in the sample.
    - Features: One feature is one observation at a time step.

Therefore, there is the need to create sequences corresponding to the prediction task. For example, if this work intends to predict the price for the next week based on the previous four weeks, each sequence will contain four values and the target variable will be the price for the fifth week.

- **Model Building**

The first step of building the model is to split the processed data into training and validation sets. A common split might be 80% for training and 20% for validation.

Then, using a preferred programming language, it is possible to define and train an LSTM model. An example of general model programming, using Python's library Keras, is the following:

```
model = Sequential()
model.add(LSTM(50, activation='relu', input_shape=(n_steps, n_features)))
model.add(Dense(1))
model.compile(optimizer='adam', loss='mse')
```

This creates a model with one LSTM layer with 50 neurons and one output layer. The input is shaped to receive "*n_steps*" as the length of the lookback window and "*n_features*" as the number of sequences. Additionally, the model is compiled with the Adam optimization algorithm and the Mean Squared Error loss function. [40] [44]. Refer to the Annex for the detailed coding developed during this research.

- **Model Training**

This is the step to fit the model to the training data, and to pass in the validation data for monitoring. It is needed to choose the number of epochs (full passes through the training data) and the batch size (number of samples per gradient update).

- **Model Evaluation and Prediction**

The evaluation of the model's performance on the validation set. If the performance is unsatisfactory, the following step would be to re-analyze the previous steps in an iterative way. Otherwise, if the model performs well, it is functional to predict the target variable.

The diagram below illustrates the model application process that was undertaken in this work.

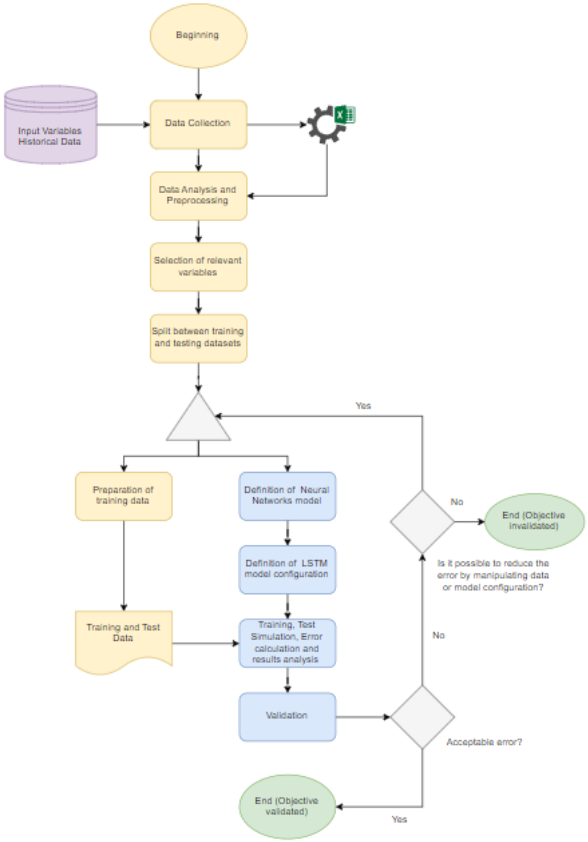


Figure 3-1 *– Model application process diagram. Source: Author's elaboration*

3.1 Data

The data selected for the development of the work were extracted from the data history of the Operation Planning of the Brazilian Electric System, disclosed by [37] and [39]. The series has weekly periodicity and was extracted for the time interval between June/2001 and December/2022, covering 1139 operative weeks (5 weeks were excluded due to lack of all input data for them – refer to Annex – FINAL_INPUTS_v2.xls).

The selected variables are Settlement Price (PLD), Energy Load, Maximum Demand, In-Flow Natural Energy, Stored Energy, Energy Generation by Hydroelectric Sources, Energy Generation by Thermoelectric sources, and finally, Energy Exchange between different subsystems.

The data collected will be inserted as inputs for the LTSM network to be implemented in this work.

3.1.1 Gathering

Forecasting electricity prices in Brazil requires consideration of multiple unique factors of its region. Here are the factors considered for the design of the LSTM model presented in this report:

1. **Regional Electricity Market Structure**: Brazil has a unique electricity market. It has a mix of public and private sector ownership, a large portion of its electricity comes from renewable resources (especially hydropower), and it has two parallel markets (the regulated and the free

market). Understanding the market structure helped to decide what kind of data would be useful to the model.

2. **Seasonality and Weather**: Brazil's electricity demand and prices can be significantly affected by seasonal weather patterns. For example, the country's dependence on hydropower means that periods of low rainfall can cause price spikes. Therefore, including weather and seasonality data in the model should improve its accuracy.
3. **Economic Factors**: economic factors, such as GDP growth and inflation rate, could influence electricity demand and, thus, prices. Incorporating this information might improve the performance for long-term scenarios. This work is focused on short-term prediction; therefore, these data were not included.
4. **Regulatory Factors**: any changes in the regulatory landscape can impact electricity prices. In the same way as economic factors, this was not included in this work due to its predominant relevance with long-term studies, other than short-term.
5. **Energy Generation Mix**: as previously mentioned, a large part of Brazil's energy comes from hydric and thermic resources. Changes in fuel prices or the availability of these impact electricity prices.
6. **Local Time Zone Patterns**: Electricity demand, and therefore prices, can vary across the day and across the week, with peak times often occurring in the evenings and mornings, and higher demand on weekdays compared to weekends. These patterns can be different in different regions and should be considered in the model.

Data for some of these factors were not readily available and even if some were, incorporating them into a model adds more complexity. Therefore, it was carefully considered which factors were most likely to improve the model's performance and were worth the added complexity.

The selection process for the relevant variables to input in the LSTM model was supported by multiple interviews with XP Inc. energy traders and analysts. With their help added to the meticulous analysis of the functioning of the Brazilian energy market, it was possible to separate the best available independent variables, which are described in the following section.

3.1.2 Descriptive analysis

In this section, the concepts, historical time series, and main descriptive statistics of the collected data will be highlighted. These collected data will serve as independent variables (inputs) for the proposed model. The consideration of these variables is of fundamental importance for the results of the research presented. Note that these variables are also used by the ONS model, DECOMP, as highlighted in grey in Table 2-1.

3.1.2.1 Settlement Price

The Settlement Price (PLD), as detailed in Section 2.1.5.4, is the reference for electricity prices in the short-term market. The CCEE publishes weekly prices for the three different existing load levels (Light, Medium, and Heavy) and for each SIN subsystem. The PLD is measured in Reais (Brazilian currency) per Megawatt-hour (BRL/MWh). The prices used in these research experiments are an average between all load levels for each submarket.

Below, Table 3-1 shows the main descriptive statistics by SIN subsystem.

	pld_se	pld_s	pld_ne	pld_n
count	1139.000000	1139.000000	1139.000000	1139.000000
mean	167.844030	154.399921	170.071361	149.306356
std	191.140861	182.380095	193.716973	181.507898
min	4.000000	4.000000	4.000000	4.000000
25%	30.250000	27.130000	20.920000	20.930000
50%	93.250000	80.390000	91.910000	70.010000
75%	229.125000	201.905000	242.530000	200.115000
max	822.830000	822.830000	822.830000	822.830000

Table 3-1 *– Statistical values of PLD for each submarket. Data in R$/MWh. Source: Author's elaboration*

The '*std*' stands for Standard Deviation in Table 3-1, and it shows that the PLD series has high volatility, regardless of the subsystem, and this is a strong characteristic of electricity prices in general.

Below, Figure 3-2 shows the PLD history by submarket. Note that there is no significant difference among the submarkets in regard to the price's volatility, or its absolute values, or its direction, which can be confirmed analyzing the similarity between the mean prices in Table 3-1, represented in the second row.

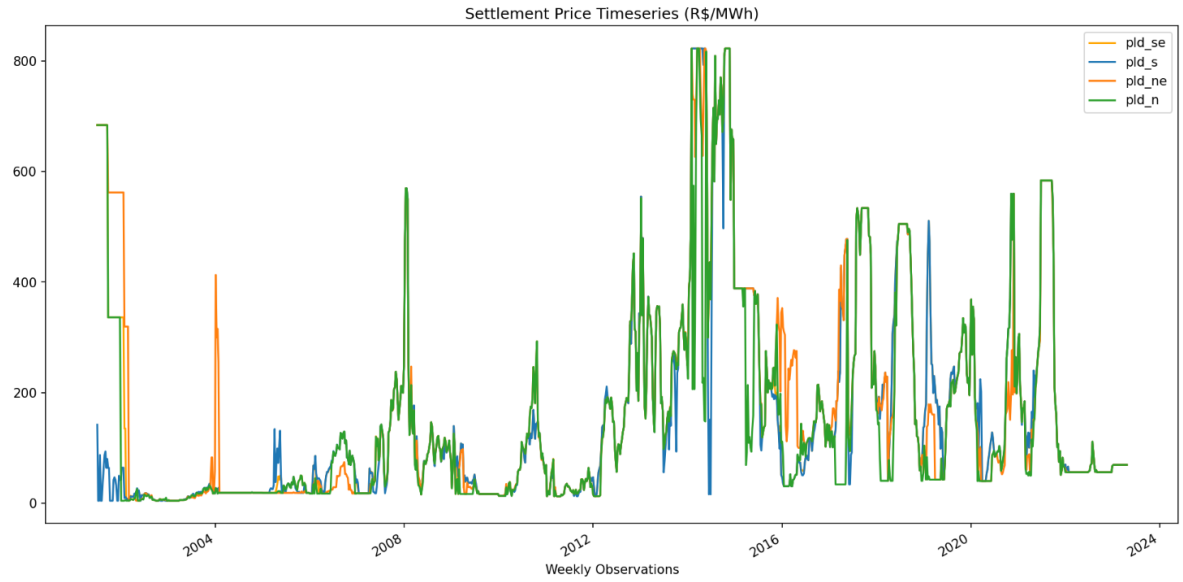


Figure 3-2 *– Historical PLD for each submarket. Data in R$/MWh. Source: Author's elaboration and CCEE historical PLD data [37].*

It is relevant to mention that the Settlement Price (PLD) is almost flat during the last years, 2022 and 2023, because the Operational Marginal Cost has been lower than the minimum PLD established by ANEEL for these years (refer to 2.1.5.4).

3.1.2.2 Load Energy

The Load Energy variable determines the average demand for electrical energy in a specified period. The data used in this work refer to the period of each operative week. Load Energy is measured by average MW and published by the SIN subsystem. Table 2 presents a summary of the main descriptive statistics of the variable and is followed by a graph with its historical values separated by subsystem.

	load_energy_se	load_energy_s	load_energy_ne	load_energy_n
count	1139.000000	1139.000000	1139.000000	1139.000000
mean	33150.064248	9481.258969	8583.696196	4325.177198
std	4984.521116	1864.365893	1879.668547	1204.923951
min	15799.787835	4538.646434	4136.362923	1973.738690
25%	29662.703521	7951.628958	7021.634256	3406.992688
50%	33589.828762	9676.330643	8680.493738	4068.782143
75%	36706.679792	10831.426622	10125.161649	5352.172789
max	45550.973137	14832.823821	12328.605756	6890.044661

Table 3-2 *– Statistical values of Load Energy for each submarket. Data in MWmed. Source: Author's elaboration and ONS historical data [39].*

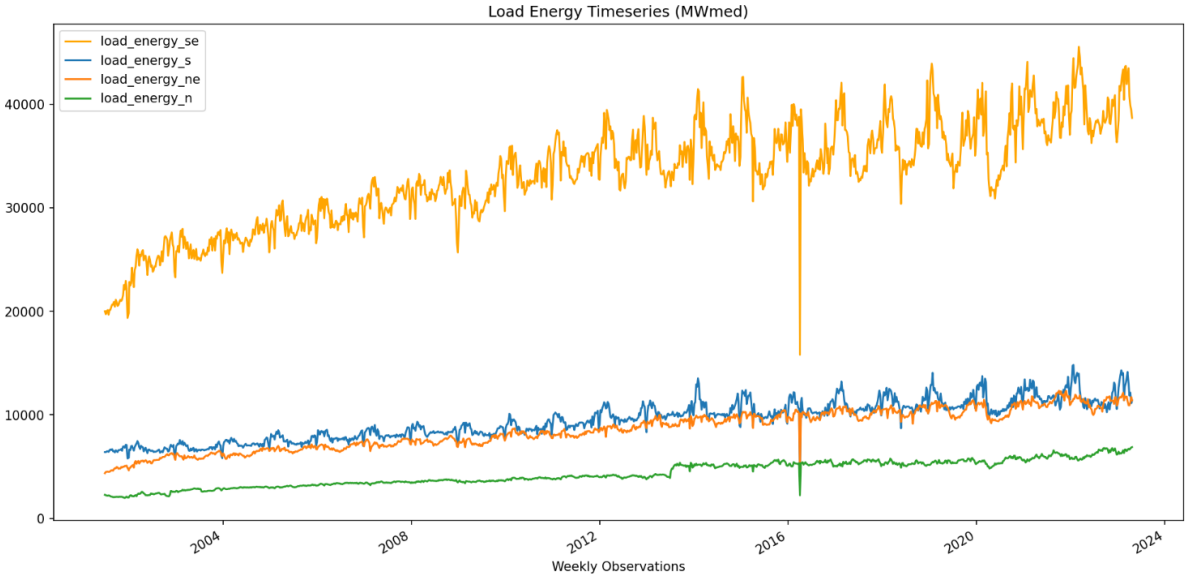


Figure 3-3 *– Historical Load Energy for each submarket. Data in MWmed. Source: Author's elaboration and ONS historical data [39].*

From the load energy data, it is possible to verify the stronger representativeness of average demand in the Southeast/Midwest subsystem, due to its higher population density, when compared to the others.

3.1.2.3 Maximum Demand

Maximum demand values are published by the subsystem and reflect the maximum values of electricity demand in the region, also known as load peaks, for a certain time interval. Values are

measured in GWh. In this work, the maximum demand values for each operative week of the sample were considered. Below is the table with the main descriptive statistics and a graph with historical values by subsystem.

	max_demand_se	max_demand_s	max_demand_ne	max_demand_n
count	1139.000000	1139.000000	1139.000000	1139.000000
mean	40.962220	12.410803	10.108622	4.871068
std	5.661713	2.437362	2.015741	1.386358
min	24.994000	7.889000	5.330000	2.234000
25%	37.279500	10.421500	8.549500	3.814500
50%	41.558000	12.497000	10.219000	4.508000
75%	45.039000	14.268500	11.751000	6.080500
max	52.889000	18.925000	15.049000	7.860000

Table 3-3 *– Statistical values of Maximum Demand for each submarket. Data in GWh. Source: Author's elaboration and ONS historical data [39].*

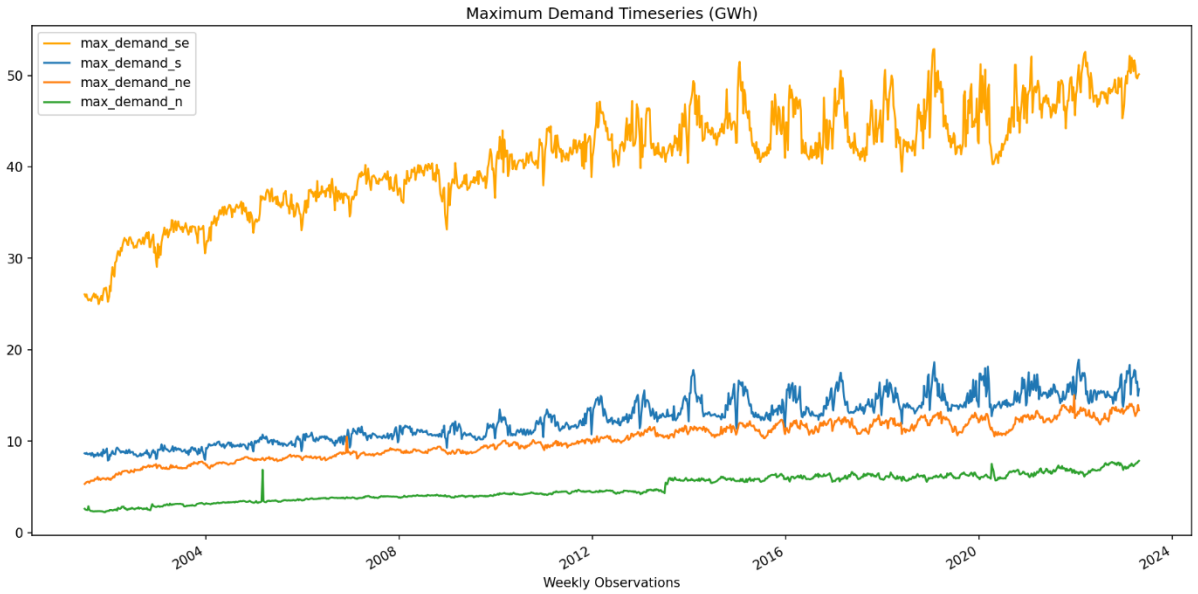


Figure 3-4 *– Historical Maximum Demand for each submarket. Data in GWh. Source: Author's elaboration and ONS historical data [39].*

As seen in the previous variable, higher values can be verified in the Southeast/Midwest subsystem, also because of its higher population density.

3.1.2.4 *Inflow Natural Energy*

Inflow Natural Energy (ENA, in Portuguese acronym) represents the energy generated from the flow of water from the hydroelectric plants, that originated from a basin or a river to the reservoirs. The ENA is measured in Average Gigawatt (GWmed). This indicator depends on some factors, such as the volume of rain, because the greater the volume of rain, the greater the generating capacity of the plants. Below, there is the table with the main descriptive statistics and a graph with the historical values for this variable by subsystem.

	ena_se	ena_s	ena_ne	ena_n
count	1139.000000	1139.000000	1139.000000	1139.000000
mean	33.840746	8.809184	5.590695	7.302315
std	17.927634	7.164354	4.813774	7.357011
min	9.756000	0.776000	0.706000	0.758000
25%	19.678500	4.149000	2.177000	1.743000
50%	28.664000	6.664000	3.634000	4.103000
75%	44.784000	10.972500	7.384000	11.011500
max	105.696000	54.757000	28.828000	38.762000

Table 3-4 *– Statistical values of In-Flow Natural Energy for each submarket. Data in GWmed. Source: Author's elaboration and ONS historical data [39].*

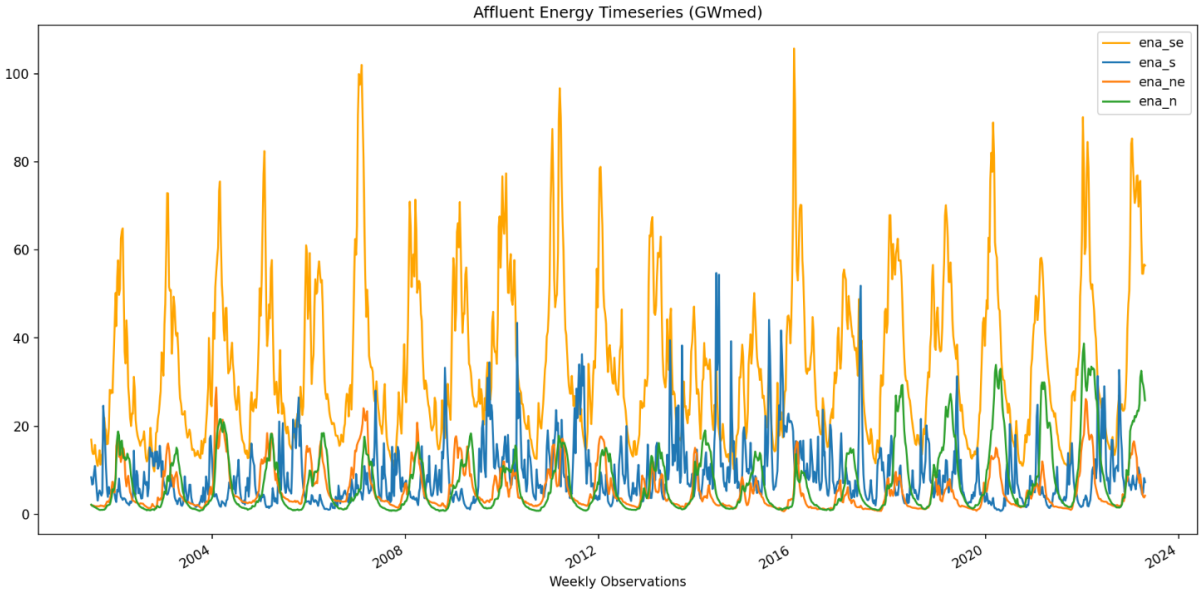


Figure 3-5 *– Historical In-Flow Natural Energy for each submarket. Data in GWmed. Source: Author's elaboration and ONS historical data [39].*

Given the dependence of the ENA values on the volume of rainfall, it is possible to verify the seasonality of the data in relation to the rainy and dry periods of the year.

3.1.2.5 *Hydroelectrical Generation*

Hydroelectric Power Generation corresponds to the total power generated within the Brazilian electric power system by hydroelectric sources, grouped by subsystem and measured by Average Megawatt (MWmed). Below is the table with the main descriptive statistics and a graph with the historical values by subsystem.

	hidro_gen_se	hidro_gen_s	hidro_gen_ne	hidro_gen_n
count	1139.000000	1139.000000	1139.000000	1139.000000
mean	28105.716681	7180.659701	4378.886918	5229.472783
std	3946.418203	2775.555671	1595.214874	3140.910520
min	12449.800000	453.400000	1032.900000	1394.300000
25%	25676.100000	4984.750000	3158.750000	3096.350000
50%	28370.600000	7212.800000	4348.700000	4066.000000
75%	30770.200000	9261.950000	5742.750000	6546.400000
max	38234.000000	13709.000000	7918.600000	16988.600000

Table 3-5 *– Statistical values of Hydrological Generation for each submarket. Data in MWmed. Source: Author's elaboration and ONS historical data [39].*

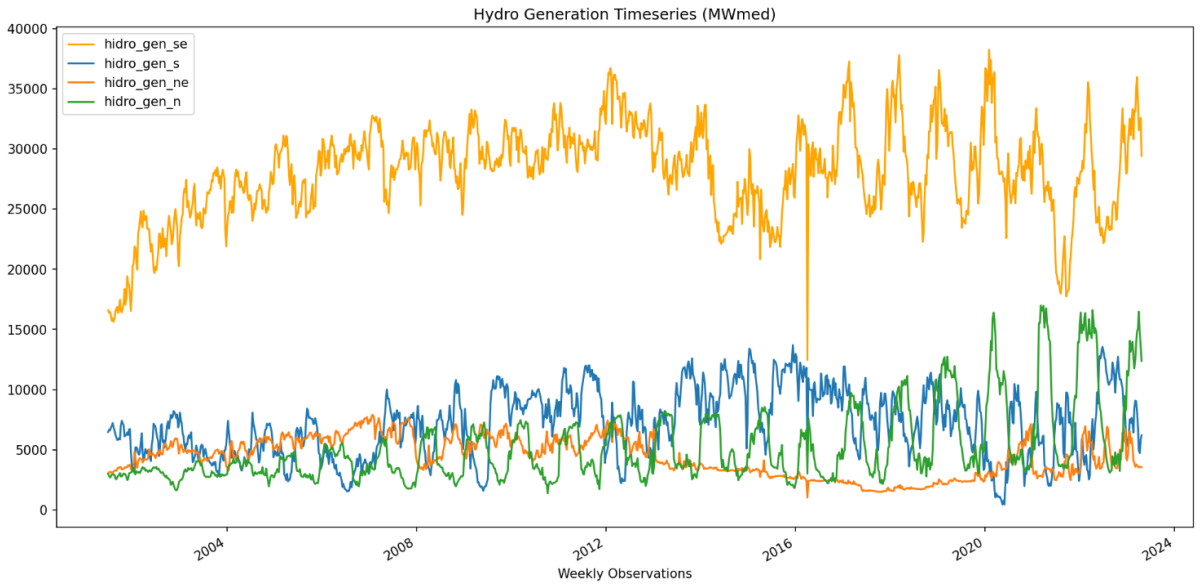


Figure 3-6 *– Historical Hydrological Generation for each submarket. Data in MWmed. Source: Author's elaboration and ONS historical data [39].*

Based on historical data, it is possible to verify the importance of the Southeast/Midwest subsystem in the generation of hydroelectric power for the SIN, which is responsible for more than 60% of the total average produced in the collected sample.

3.1.2.6 *Thermoelectrical Generation*

Thermoelectrical Generation corresponds to the energy generated through plants that use the burning of fuels, fossil or not, to generate electricity. This kind of energy generation plays a fundamental role in guaranteeing the supply of energy in moments of low levels in the reservoirs of hydroelectric plants. Below is the table with the main descriptive statistics and a graph with historical values by subsystem.

	thermo_gen_se	thermo_gen_s	thermo_gen_ne	thermo_gen_n
count	1139.000000	1139.000000	1139.000000	1139.00000
mean	3174.529939	1093.687182	1343.122651	863.69324
std	2344.620522	393.490661	1351.098904	1034.90443
min	286.500000	98.100000	0.100000	0.00000
25%	1120.750000	816.850000	200.650000	0.00000
50%	2720.700000	1048.000000	678.100000	0.00000
75%	4696.900000	1313.900000	2440.500000	1820.40000
max	24107.800000	2252.200000	5022.200000	3282.30000

Table 3-6 *– Statistical values of Thermoelectric Generation for each submarket. Data in MWmed. Source: Author's elaboration and ONS historical data [39].*

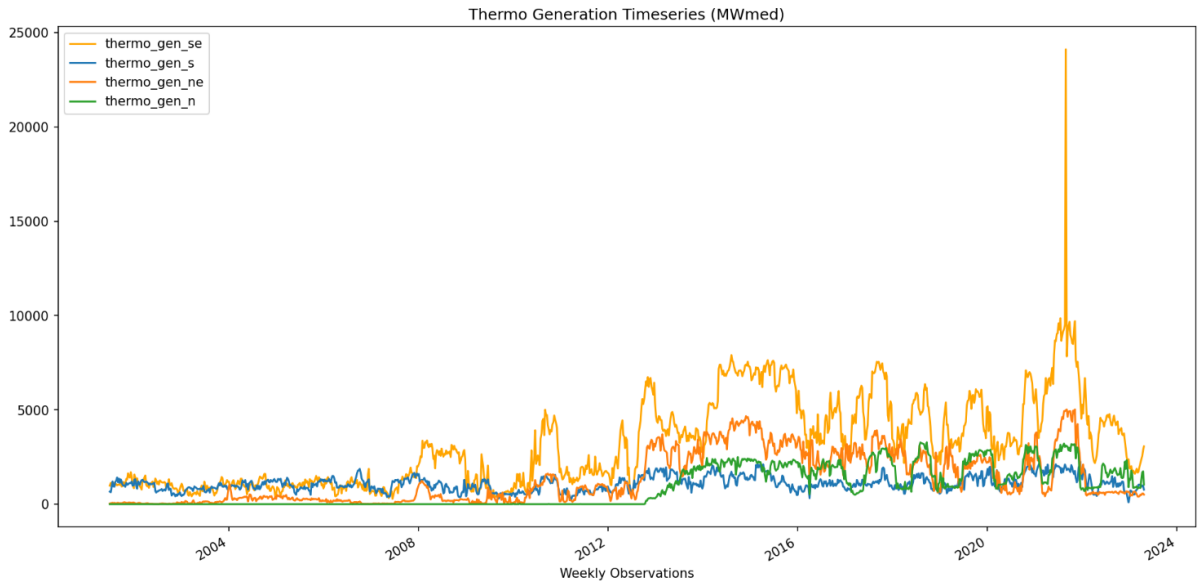


Figure 3-7 *– Historical Thermoelectric Generation for each submarket. Data in MWmed. Source: Author's elaboration and ONS historical data [39].*

From the historical data, it is possible to notice the relevance of the Southeast/Midwest subsystem also in the production of thermoelectric energy, being responsible for approximately 50% of the total

average produced in the selected sample. In addition, in August 2021 there is a high peak of thermoelectrical generation in this region, related to a strong deficit on hydroelectrical energy supply during that period, due to an intense dry period that the country faced.

3.1.2.7 Stored Energy

Stored Energy is the product between the average productivity of hydroelectric plants and their respective reservoir levels, aggregated by subsystem. Values are measured in Average Gigawatt (GWmed). Below is the table with the main descriptive statistics and a graph with the historical values per subsystem.

	stored_energy_se	stored_energy_s	stored_energy_ne	stored_energy_n
count	1139.000000	1139.000000	1139.000000	1139.000000
mean	101.625991	12.151600	26.743268	8.529153
std	37.680309	4.054031	13.158907	3.422340
min	30.928000	2.963000	2.252000	1.502000
25%	71.522500	8.988000	16.295000	5.482500
50%	101.497000	11.963000	26.609000	9.046000
75%	133.811000	15.182000	37.646000	11.497000
max	179.055000	19.822000	51.399000	15.229000

***Table 3-7** – Statistical values of Stored Energy for each submarket. Data in GWmed. Source: Author's elaboration and ONS historical data [39].*

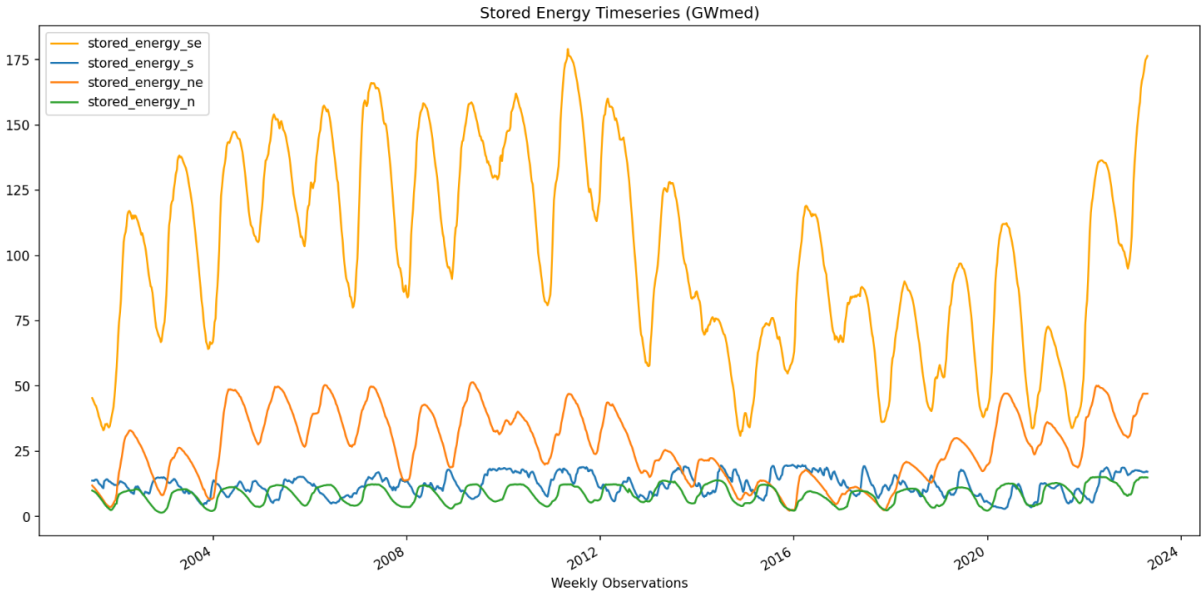


***Figure 3-8** – Historical Stored Energy for each submarket. Data in GWmed. Source: Author's elaboration and ONS historical data [39].*

3.1.2.8 Imports and Exports

Imports and Exports are mechanisms that the System Operator uses to transfer energy produced between the different subsystems through the transmission networks that interconnect them. The purpose of these transfers is to meet the demand for energy between the different SIN submarkets.

The data are measured in Average MW (MWmed). Negative values mean that, for the operative week in question, the analyzed subsystem exported more energy than imported. Positive values mean the opposite, that is, the subsystem received more energy.

There are four streams of transfers between subsystems: between Northeast and Southeast/Midwest (NE – SE/CO); North and Northeast (N – NE); North and Southeast/Midwest (N – SE/CO); Southeast/Midwest and South (SE/CO – S). The table below contains the main descriptive statistics and is followed by a graph with historical values.

	exports_se	exports_s	exports_ne	exports_n
count	1139.000000	1139.000000	1139.000000	1139.000000
mean	995.056190	655.292362	541.908692	1758.597170
std	1293.920498	1073.013330	1505.922056	2210.655275
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	10.230000
50%	253.000000	0.000000	0.000000	907.000000
75%	1798.500000	1043.500000	0.000000	2729.000000
max	6529.000000	8000.000000	9138.000000	9947.000000

***Table 3-8** – Statistical values of Exports for each submarket. Data in MWmed. Source: Author's elaboration and ONS historical data [39].*

	imports_se	imports_s	imports_ne	imports_n
count	1139.000000	1139.000000	1139.000000	1139.000000
mean	1123.323539	1547.568920	1244.823529	191.584723
std	1917.254426	2013.387116	1029.404526	445.245937
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	213.500000	0.000000
50%	0.000000	574.000000	1168.000000	0.000000
75%	1482.500000	2577.500000	2017.000000	0.000000
max	9885.000000	8904.000000	5560.000000	4000.000000

***Table 3-9** – Statistical values of Imports for each submarket. Data in MWmed. Source: Author's elaboration and ONS historical data [39].*

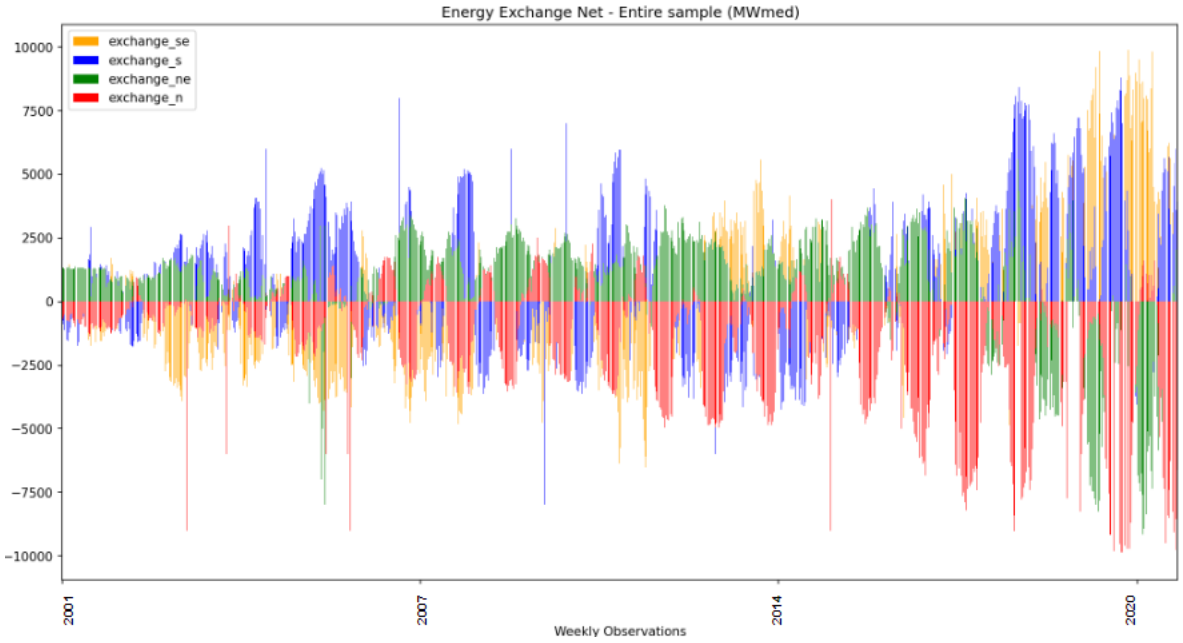


Figure 3-9 *– Historical Energy Imports and Exports for each submarket. Data in MWmed. Source: Author's elaboration and ONS historical data [39].*

Given the above historical data and statistical values, it is noticeable that the North (red bars) is strongly representative in exporting energy, while the Northeast (green bars) and South (blue bars) do more imports. The Southeast/Midwest subsystem vary similarly in between the energy exchanges sides. It is also relevant to note that the overall exchanges increased significantly in the past 5-10 years. This can be explained with the increase of transmission lines in the Brazilian National Interconnected System (refer to Figure 2-5 and Figure 2-6). For reference, the following Figure 3-10 shows a less dense view of this data by zooming in the 2021 period only.

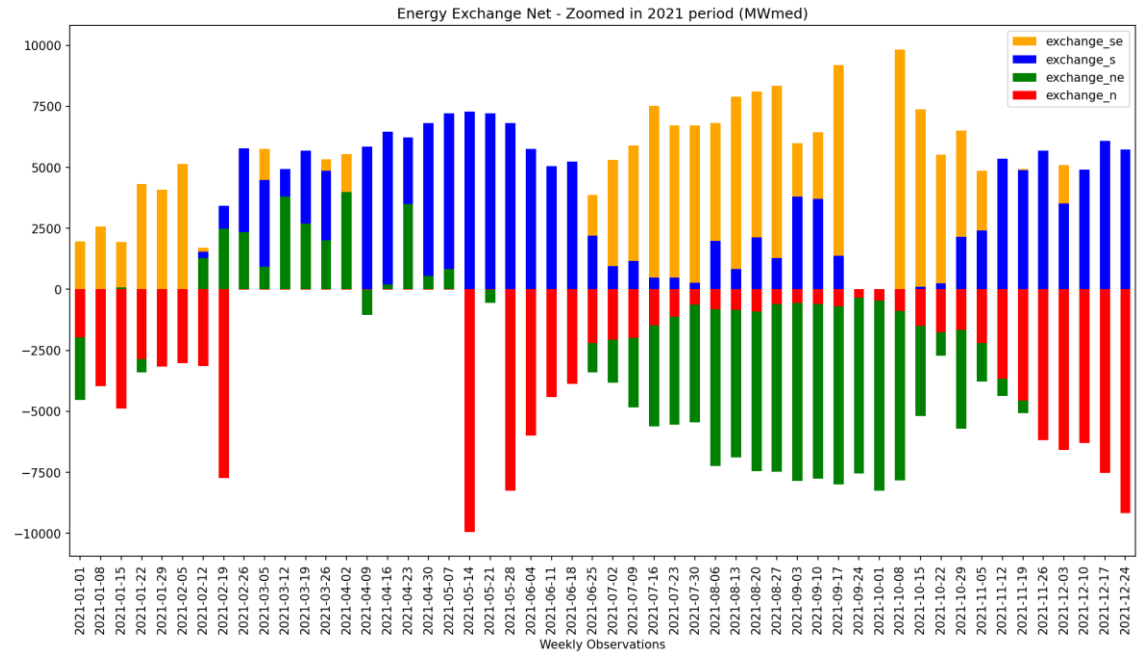


Figure 3-10 *– Historical Energy Imports and Exports for each submarket during 2021 period. Data in MWmed. Source: Author's elaboration and ONS historical data [39].*

During 2021, it is possible to see that the North and Northeast (red and green bars, respectively) express more exports, especially during the second semester. On the other hand, the Southeast and South submarkets (orange and blue bars, respectively) present more imports during the year. This happens since the south of Brazil is much more populated than the north.

3.1.3 Treatment

To make projections of electricity prices for the different analyzed submarkets, all variables underwent a standardization process described by the following equation:

$$X_t = \frac{x_t - min(x)}{max(x) - min(x)} \qquad \textit{Equation 3.1}$$

Where X_t is the normalized value of the series at time t, x_t represents the original variable that will undergo the linear transformation, the max(x) and min(x) values represent the maximums and minimums, respectively, of the original series to be normalized.

The objective of this linear transformation is to adapt the input data to the scale of the neural network's activation functions. The linear transformation proposed by Equation 3.1 does not change the properties of the input data, it just restricts them to the set [0,1].

The standardization defined by Equation 3.1 was performed for all variables described in Section 3.1.2.

Additionally, all independent variables were analyzed for their correlation with the target variable, the Settlement Price (PLD). The following tables show the correlation results for each feature by submarket.

Features	**Southeast/Midwest**	**Northeast**	**South**	**North**
Settlement Price (PLD)	1.0	1.0	1.0	1.0
Load Energy	0.233251	0.253214	0.438806	0.302692
Maximum Demand	0.238038	0.239961	0.437343	0.296453
In-Flow Natural Energy	-0.275795	-0.292015	-0.033271	-0.240569
Hydroelectrical Power Generation	-0.271053	-0.491908	0.179393	-0.124943
Thermoelectrical Power Generation	0.643515	0.675219	0.586428	0.573505
Stored Energy	-0.625240	-0.568609	-0.210196	-0.113380
Exports	-0.229073	0.047778	0.017672	-0.026585
Imports	0.174564	0.055364	-0.017942	-0.040926

Table 3-10 – Pearsons Correlation Analysis for sampled dataset. Source: Author's elaboration.

The Pearsons correlation coefficients above are useful to understand the degree level of correlation between the independent variables and the target variables. If the coefficient value lies in between ± 0.50 and ± 1, then it is said to be a strong correlation. If the value lies in between ± 0.30 and ± 0.49, then it is said to be a medium correlation. When the value is between - 0.29 and + 0.29, then it is said to be a small correlation [42].

Table 3-10 shows that, for the Southeast/Midwest submarket, the Thermoelectrical Power Generation, and the Stored Energy features present strong correlation, while the others are small correlated. Additionally, Northeast is similar, but Hydroelectric Power Generation adds up as another strong

correlated variable. Following, the South submarket presents Load Energy and Maximum Demand with medium correlation, and only Thermoelectrical Power Generation with a strong correlation. Finally, the North presents a strong correlation only for Thermoelectrical Power Generation, while all the other variables are within the small correlation range.

3.2 Model

This section describes the proposed model and its application.

3.2.1 Network Configuration

The data described in the previous section was split into two different subsets: training (70% of the sample) and validation (30% of the sample). The training subset is used to train the network learning, and the validation part is used as backtesting for the trained model.

For any application of neural networks, it is necessary to define the configuration to be used for its architecture, and several methodologies are described in the literature to define the parameters of the network, such as number of hidden layers and number of neurons. For ANN applications involving time series, it is indicated that there are no systematic methodologies to obtain optimal configurations. It means that it is not possible to describe a universal algorithm for forecasting time series, and it will always be necessary to adapt different models that best fit the data analyzed by multiple tests that the researcher needs to perform and organize [43].

The proposed LSTM model was trained with the method of optimization called Adam. This method is an algorithm for optimizing stochastic functions. This method is appropriate for non-stationary problems [44]. The tool used in the training of the proposed network was Python and its library called Keras.

The model's input data is described in Section 3.1.2 of this paper and standardized by Equation 3.1. After some calculation speed tests and predictive gain, 100 intermediate neurons were defined as default. This configuration was trained iteratively for 100 epochs. The network output is the Settlement Price (PLD) per submarket, still on the standardized scale. Finally, the data are unstandardized to obtain all information in real scale. This process is also described in the flow chart in Figure 3-1.

Additionally, multiple tests were performed to decide how many layers the network would be configured with. Figure shows the RMSE (refer to Section 2.2.4.1) results for different layers setups:

```
[138] # get root mean sqaured error (RMSE) - 100 lstm, 1 dense
      rmse = sqrt(mse(actuals, predictions))
      print('Test RMSE: %.3f' % rmse)

      Test RMSE: 76.877

[141] # get root mean sqaured error (RMSE) - 100 lstm, 25 dense, 1 dense
      rmse = sqrt(mse(actuals, predictions))
      print('Test RMSE: %.3f' % rmse)

      Test RMSE: 68.079

[147] # get root mean sqaured error (RMSE) - 100 lstm, 50 lstm, 25 dense, 1 dense
      rmse = sqrt(mse(actuals, predictions))
      print('Test RMSE: %.3f' % rmse)

      Test RMSE: 78.374

[153] # get root mean sqaured error (RMSE) - 100 lstm, 8 dense relu, 1 dense linear
      rmse = sqrt(mse(actuals, predictions))
      print('Test RMSE: %.3f' % rmse)

      Test RMSE: 165.136

[159] # get root mean sqaured error (RMSE) - 100 lstm, 8 dense, 1 dense
      rmse = sqrt(mse(actuals, predictions))
      print('Test RMSE: %.3f' % rmse)

      Test RMSE: 68.233
```

Figure 3-11 – Network layers setup testing results. Source: Author's elaboration.

The green sentences describe the layers and neurons setup that was used for the model training and below it is the RMSE test result. The lower the RMSE, the closer the predicted values were to the actual values. Therefore, it was chosen to configure the model with an LSTM input layer, aggregating 100 neurons, in conjunction with a Dense layer with 25 neurons. The final output goes through another Dense layer with 1 linear neuron to finish the training cycle.

3.2.2 Training Configuration

In the context of predicting electricity prices in Brazil, the following are the factors considered to configure the training process of the LSTM model:

1. **Periodicity**: this depends on the granularity of the data and the predictions one is willing to make. For example, when using hourly data and willing to predict prices for the next day, one might choose a daily periodicity. If using daily data and willing to predict prices for the next month, one might choose a monthly periodicity. In this work, it was chosen to use weekly data to predict the next week, considering the periodicity of factors influencing electricity prices, such as days usage patterns and seasonal weather patterns.
2. **Time Steps**: the choice of time steps, or the number of previous time periods, the model should consider when making a prediction, is a hyperparameter that one may need to experiment with. One consideration might be the typical time it takes for changes in explanatory variables (like weather or economic indicators) to affect electricity prices. If one believes that the price of electricity today has been influenced by the weather over the last three days, then the number of time steps should be at least three.

Another consideration was the amount of data available. LSTMs are capable of learning long-term dependencies, but the longer the sequence, the more data you need. If one has a limited amount of data, it might be more practical to choose a smaller number of time steps.

To find the best configuration for this specific problem, a systematic approach was used to tune these parameters (cross-validation). This process was necessary to achieve the best performance.

It is also important to remember that time series forecasting is inherently uncertain, and all models will have some degree of error. This is why it is also important to consider the confidence intervals around predictions.

In this work, it was chosen to use the input and output data on a weekly basis. In addition, the training process was configured to have a fixed 4-week look-back window and a dynamic look-forward window that varied between 1 to 4 weeks ahead. The figures below support a proper understanding of these windows and how they affect the model.

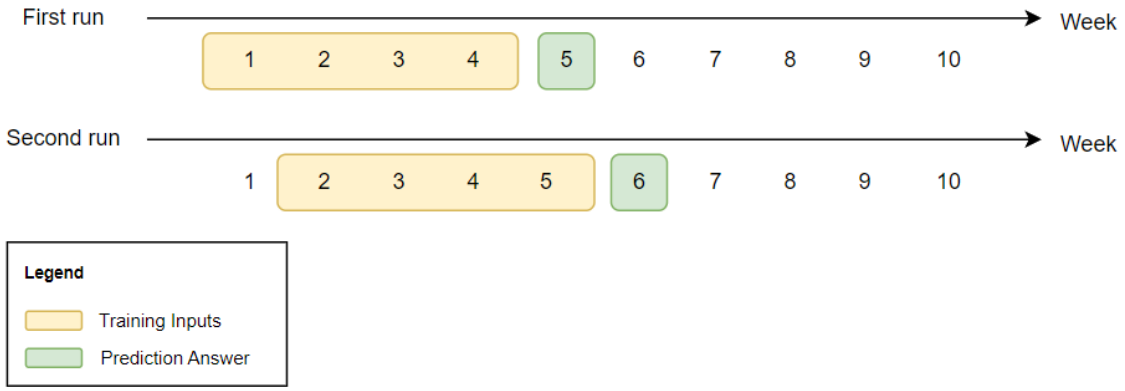


Figure 3-12 *– Illustration of 4 weeks lookback window training to predict settlement price of the first week ahead. Source: Author's elaboration.*

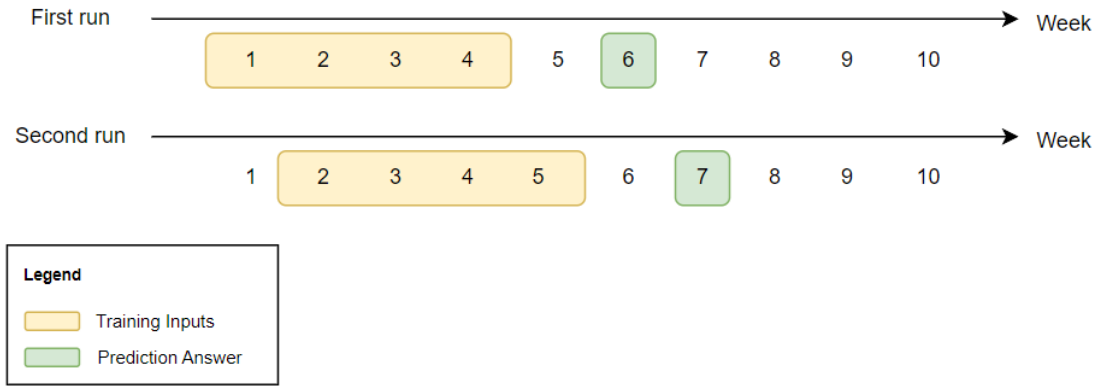


Figure 3-13 *– Illustration of 4 weeks lookback window training to predict settlement price of the second week ahead. Source: Author's elaboration.*

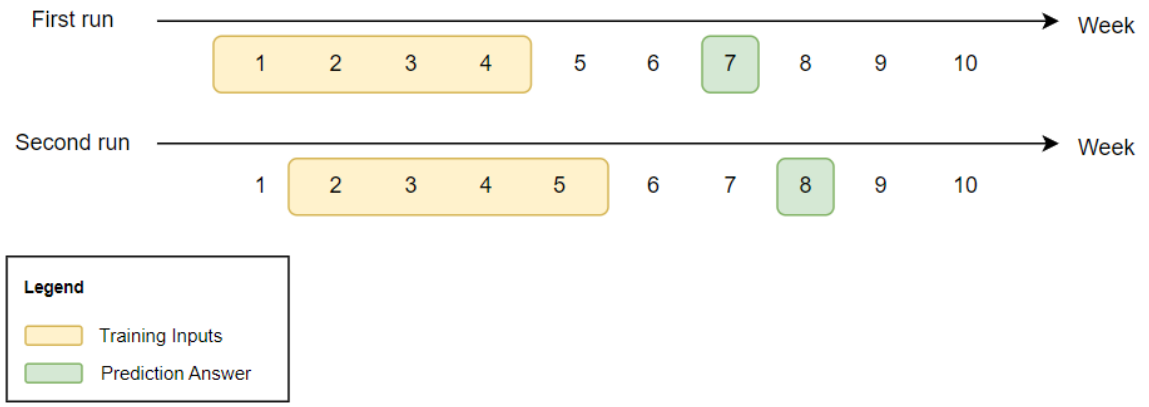


Figure 3-14 *– Illustration of 4 weeks lookback window training to predict settlement price of the third week ahead. Source: Author's elaboration.*

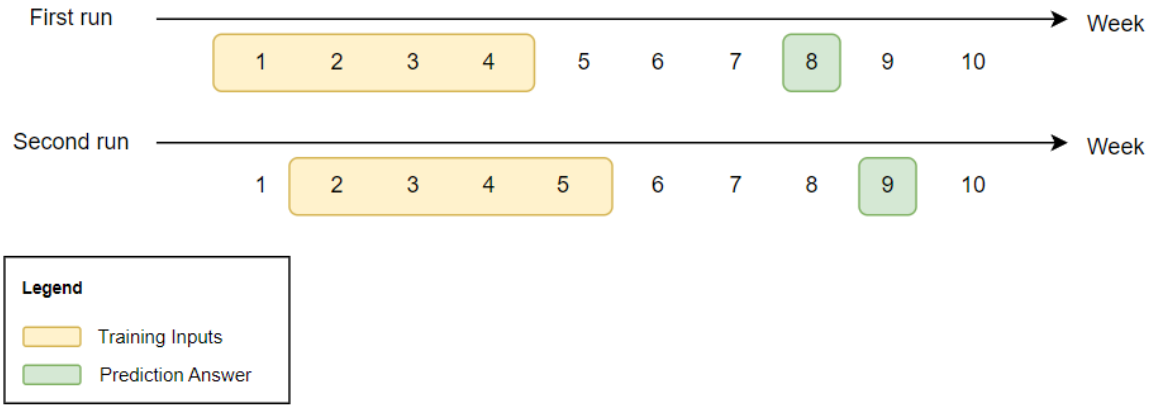


Figure 3-15 *– Illustration of 4 weeks lookback window training to predict settlement price of the fourth week ahead. Source: Author's elaboration.*

The figures above illustrate the difference among running the training to predict the settlement price of the very first, the second, the third and the fourth week ahead starting from the moment in which the model is looking back to the previous four weeks information. Note that all figures demonstrate how the training method goes on the timeline, as it is configured, by showing the difference between the first and second run for each configuration.

4 Results

In Chapter 4 the results for simulation tests using the proposed LSTM model are shown by submarket and by training configuration. Moreover, a benchmark comparison is demonstrated between the predictions of this model and the DECOMP model, used by ONS, when forecasting 4 weekly prices ahead.

4.1 Southeast / Midwest

This section shows the results for the Southeast/Midwest (SE) submarket.

4.1.1 First week

Below, the model was trained to predict the first weekly Settlement Price ahead based on all input data of four previous weeks (check Section 3.2.3 for more detailed information about the training methods).

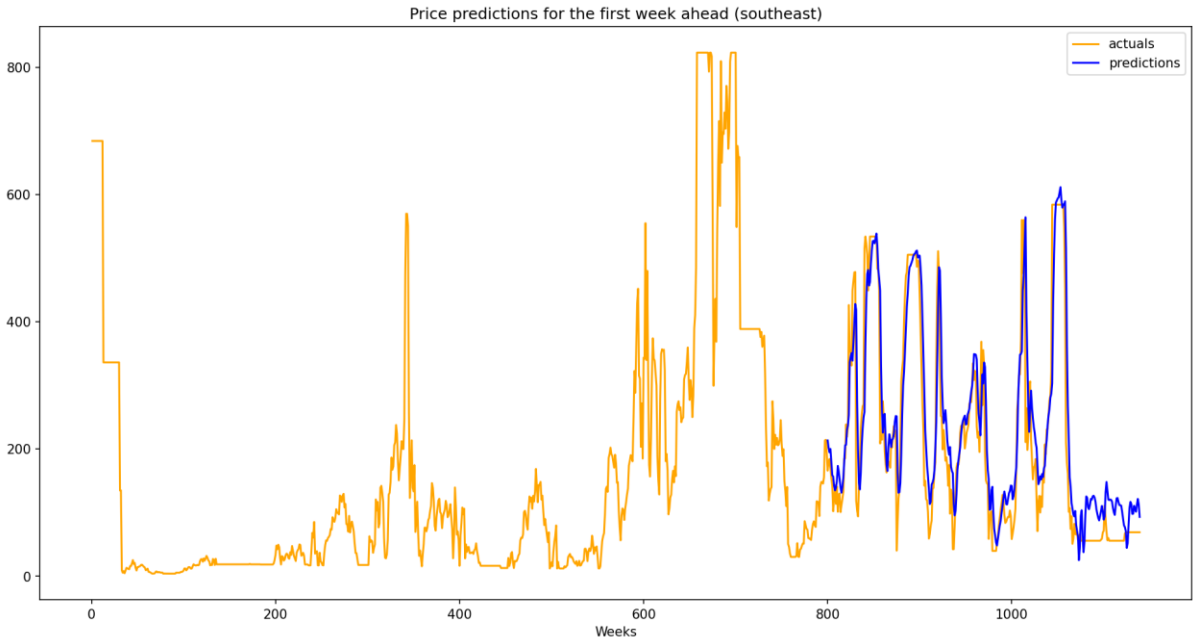


Figure 4-1 *– Simulation test results for predicting settlement price of the first week ahead in Southeast/Midwest sub-market. Source: Author's elaboration and CCEE historical PLD data [37].*

Predictions RMSE Southeast 1: 70.040

This graph illustrates the historical series of the actual weekly Settlement Prices for the 1039 weeks sampled with the orange line and the prediction results with the blue line. Because of the 4 weeks lookback window and 1 week look forward window, the training went over the first 795 weeks and the predictions for the last 340 weeks (check Section 3.2.3 for more detailed information about the training configuration). The Root Mean Squared Error, detailed in Section 2.2.4.1, is $70.040 BRL.

4.1.2 Second week

Below, the model was trained to predict the second weekly Settlement Price ahead based on all input data of four previous weeks (check Section 3.2.3 for more detailed information about the training methods).

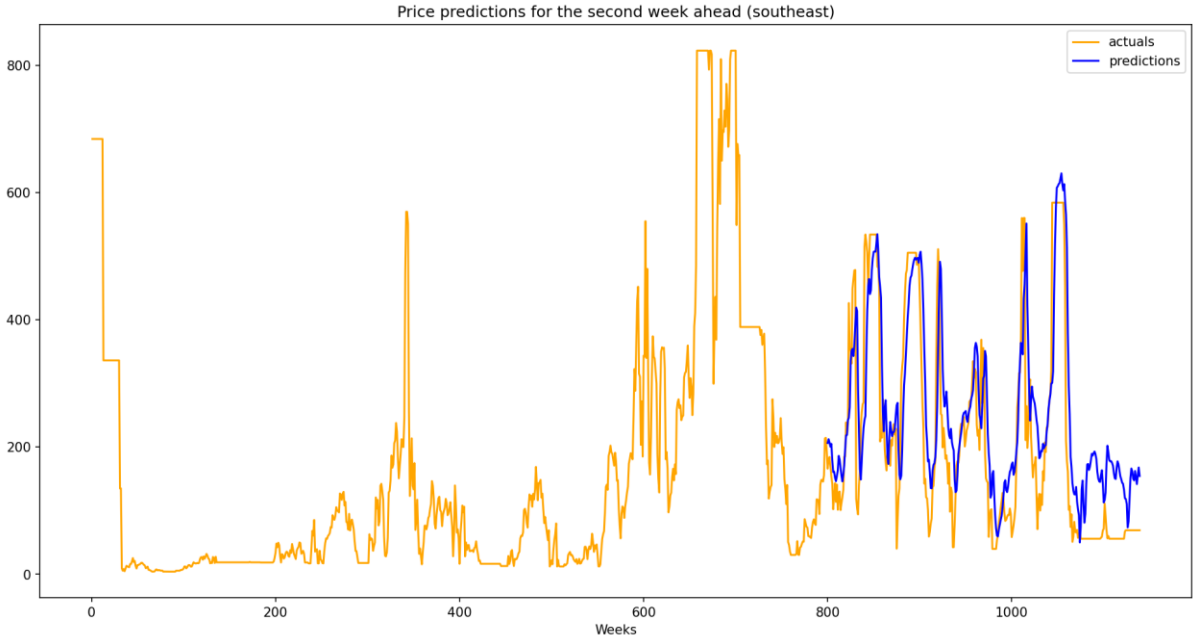


Figure 4-2 *– Simulation test results for predicting settlement price of the second week ahead in Southeast/Midwest sub-market. Source: Author's elaboration and CCEE historical PLD data [37].*

Predictions RMSE Southeast 2: 101.093

This graph illustrates the historical series of the actual weekly Settlement Prices for the 1039 weeks sampled with the orange line and the prediction results with the blue line. Because of the 4 weeks lookback window and 2 weeks look forward window, the training went over the first 795 weeks and the predictions for the last 339 weeks (check Section 3.2.3 for more detailed information about the training configuration). The Root Mean Squared Error, detailed in Section 2.2.4.1, is $101.093 BRL.

4.1.3 Third week

Below, the model was trained to predict the third weekly Settlement Price ahead based on all input data of four previous weeks (check Section 3.2.3 for more detailed information about the training methods).

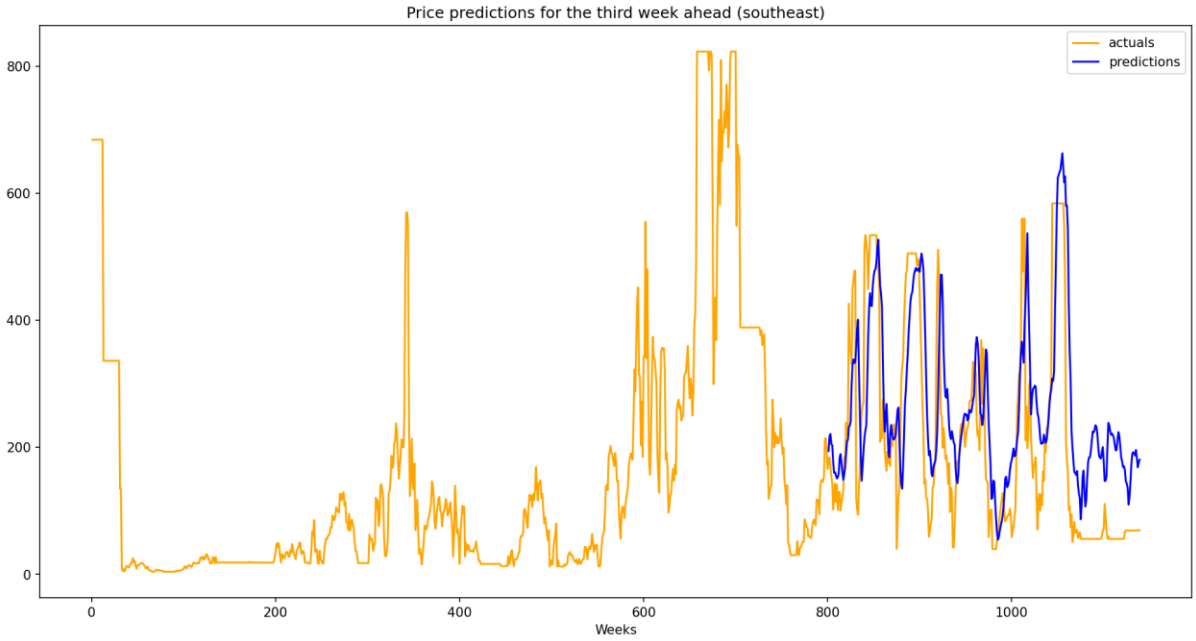


Figure 4-3 *– Simulation test results for predicting settlement price of the third week ahead in Southeast/Midwest sub-market. Source: Author's elaboration and CCEE historical PLD data [37].*

Predictions RMSE Southeast 3: 126.740

This graph illustrates the historical series of the actual weekly Settlement Prices for the 1039 weeks sampled with the orange line and the prediction results with the blue line. Because of the 4 weeks lookback window and 3 weeks look forward window, the training went over the first 794 weeks and the predictions for the last 339 weeks (check Section 3.2.3 for more detailed information about the training configuration). The Root Mean Squared Error, detailed in Section 2.2.4.1, is $126.740 BRL.

4.1.4 Fourth week

Below, the model was trained to predict the fourth weekly Settlement Price ahead based on all input data of four previous weeks (check Section 3.2.3 for more detailed information about the training methods).

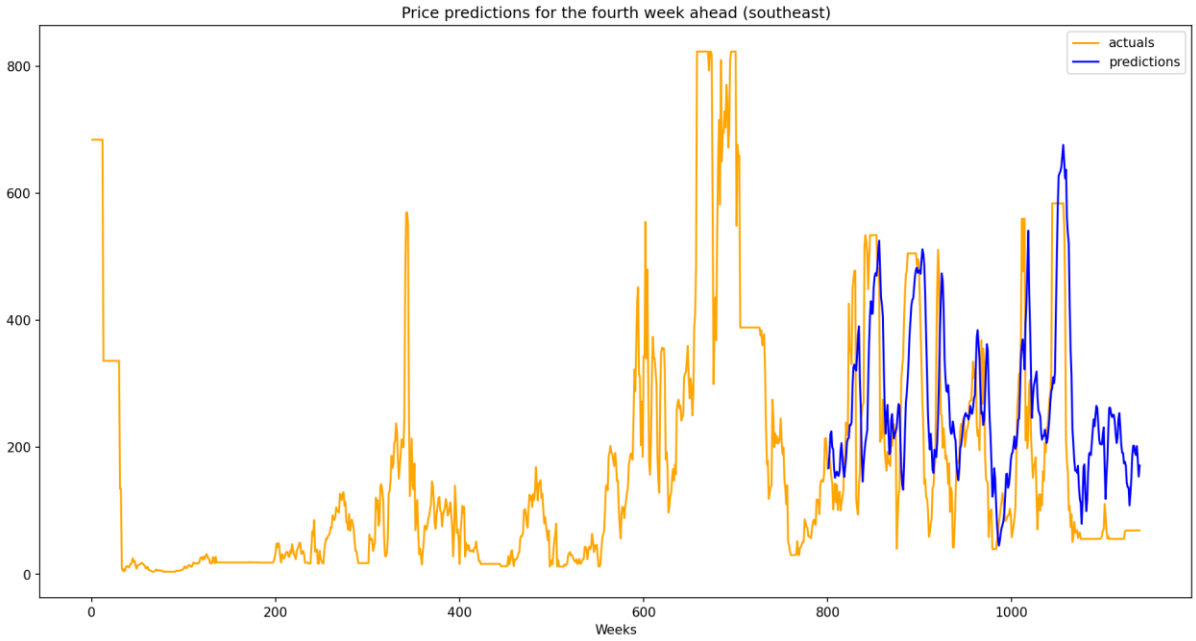


Figure 4-4 *– Simulation test results for predicting settlement price of the fourth week ahead in Southeast/Midwest sub-market. Source: Author's elaboration and CCEE historical PLD data [37].*

Predictions RMSE Southeast 4: 141.511

This graph illustrates the historical series of the actual weekly Settlement Prices for the 1039 weeks sampled with the orange line and the prediction results with the blue line. Because of the 4 weeks lookback window and 3 weeks look forward window, the training went over the first 793 weeks and the predictions for the last 339 weeks (check Section 3.2.3 for more detailed information about the training configuration). The Root Mean Squared Error, detailed in Section 2.2.4.1, is $141.511 BRL.

Note that the RMSE increases once the predicted weekly price gets farther away from the present moment of the model. That is the expected behavior, since any prediction model should present a worse performance once it is trying to predict something that is farther away in the future. In other words, for any prediction model, the further away the target, the worse the performance.

4.1.5 Benchmark Analysis

This section compares the results of the proposed LSTM model against the DECOMP model (refer to Section 2.1.5.3) predictions when forecasting the next four weekly prices ahead in the Southeast/Midwest region. This analysis was done during the first semester of 2021 because that is the closest period from now that the Brazilian electricity prices were showing high volatility. On the other hand, the years 2022 and 2023 have presented floor static prices (see Section 3.1.2.1).

Figure 4-5 *– LSTM vs. DECOMP evaluation on Settlement Price prediction for the next four weeks as of the beginning of every month of the first semester of 2021 in the Southeast/Midwest. Source: Author's elaboration, CCEE historical data [37] and DECOMP results (Annex).*

The graph below aggregates all seven previous prediction slices for the first semester of 2021.

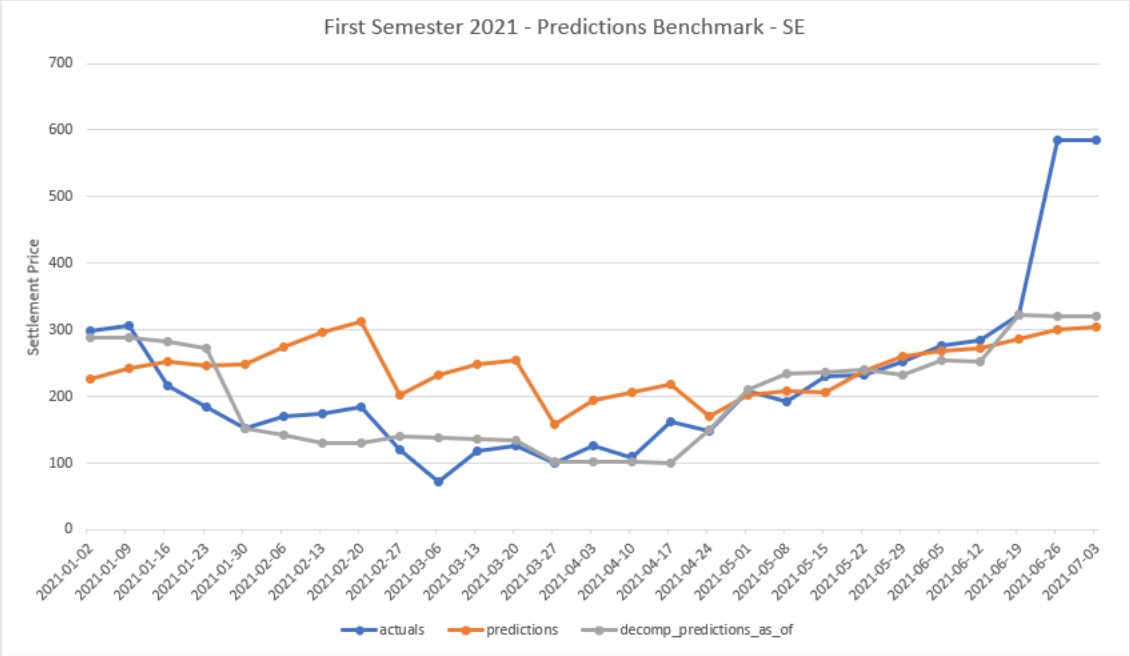


Figure 4-6 *– LSTM vs. DECOMP results evaluation for the first semester of 2021 in the Southeast/Midwest. Source: Author's elaboration, CCEE historical data [37] and DECOMP results (Annex).*

RMSE for LSTM Absolute Values: 107.766

RMSE for DECOMP Absolute Values: 79.395

The LSTM results (orange line) present a worse RMSE compared to the DECOMP model results (grey line). It means that the benchmark is closer to the absolute value of the actual prices (blue line).

The following graph shows the Trend Direction Accuracy Measurement for the results of each model. As detailed in Section 2.2.4.2, it indicates how well the model predicts the direction of the price changes.

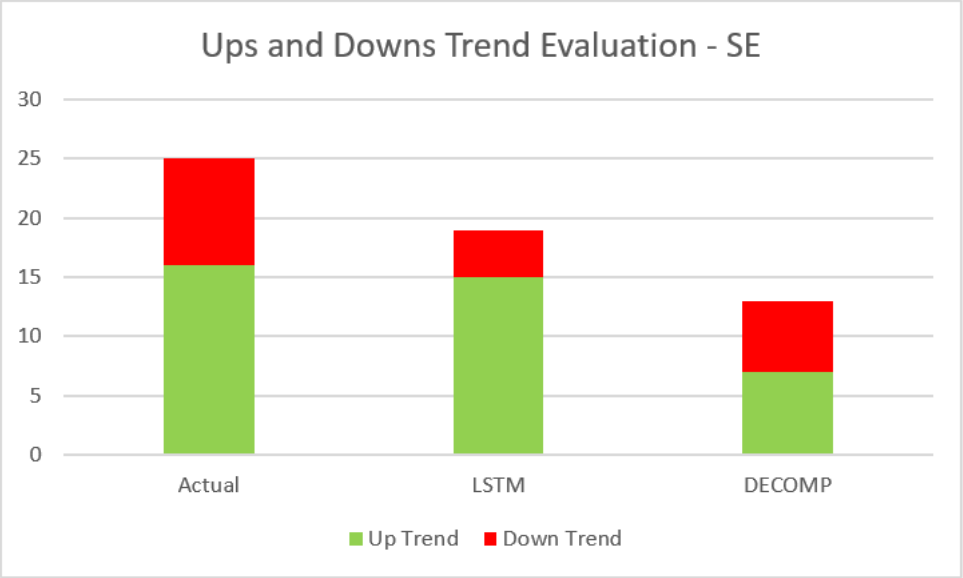


Figure 4-7 *– LSTM vs. DECOMP over trend direction accuracy, Southeast/Midwest. Source: Author's elaboration, CCEE historical data [37] and DECOMP results (Annex).*

LSTM Trend Direction Accuracy: 76.0 %

DECOMP Trend Direction Accuracy: 52.0 %

The results show that the LSTM model presented a better performance in predicting whether the electricity Settlement Price is going up or down, when compared to the DECOMP model. Of 25 trends, the proposed model got 76% of the trends correctly, while the benchmark got 52%.

The following graph shows the actual weekly price volatility with the grey bars, and the predicted weekly volatility of each model, green bars for LSTM and red bars for DECOMP. Volatility is presented as the percentage of the price rising or falling from one week to the next.

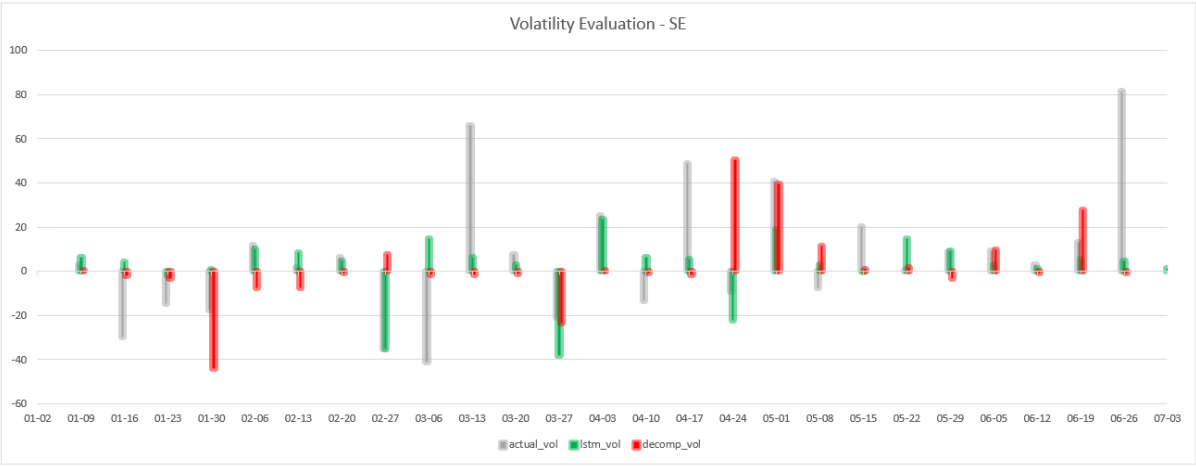


Figure 4-8 *– LSTM vs. DECOMP over volatility accuracy, Southeast/Midwest. Source: Author's elaboration, CCEE historical data [37] and DECOMP results (Annex).*

RMSE LSTM Vol: 26.067

RMSE DECOMP Vol: 30.415

The RMSE (Section 2.2.4.1) was applied to the volatility predictions of both models to assess their performance. The smaller the RMSE, the greater the accuracy of the model in predicting how intensively the price would change from one week to the next week. The results show that the proposed model performed slightly better than the reference (DECOMP).

4.2 Northeast

This section shows the results for the Northeast (NE) submarket.

4.2.1 First week

Below, the model was trained to predict the first weekly Settlement Price ahead based on all input data of 4 previous weeks (check Section 3.2.3 for more detailed information about the training methods).

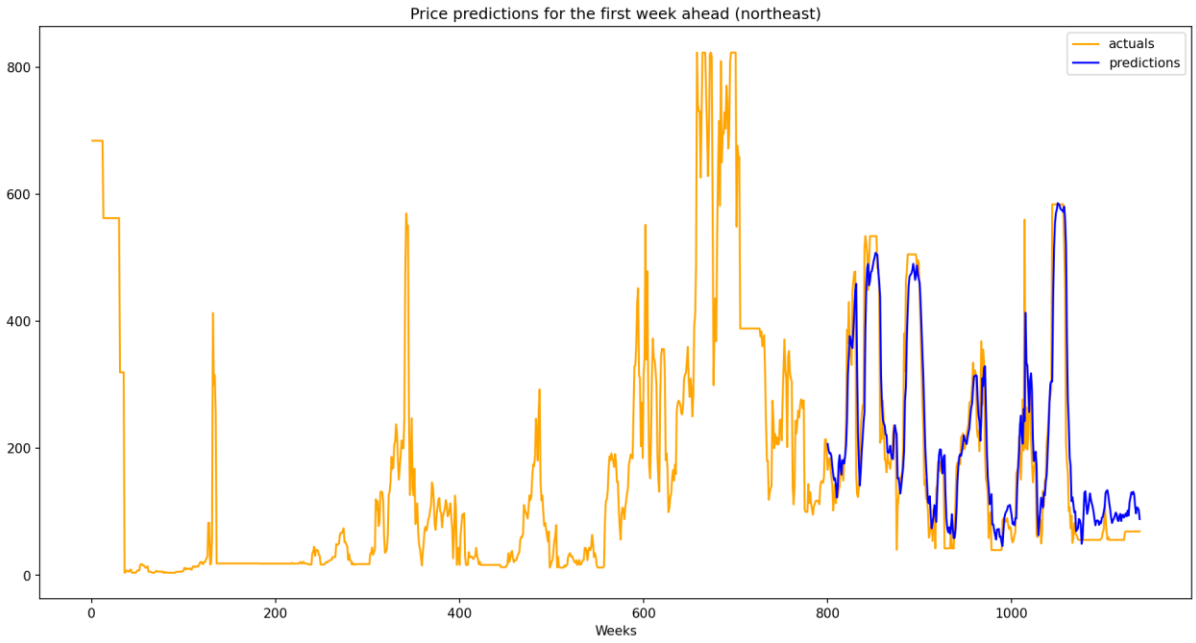


Figure 4-9 *– Simulation test results for predicting settlement price of the first week ahead in Northeast sub-market. Source: Author's elaboration and CCEE historical PLD data [37].*

Predictions RMSE Northeast 1: 65.592

This graph illustrates the historical series of the actual weekly Settlement Prices for the 1039 weeks sampled with the orange line and the prediction results with the blue line. Because of the 4 weeks lookback window and 3 weeks look forward window, the training went over the first 795 weeks and the predictions for the last 340 weeks (check Section 3.2.3 for more detailed information about the training configuration). The Root Mean Squared Error, detailed in Section 2.2.4.1, is $65.592 BRL.

4.2.2 Second week

Below, the model was trained to predict the second weekly Settlement Price ahead based on all input data of 4 previous weeks (check Section 3.2.3 for more detailed information about the training methods).

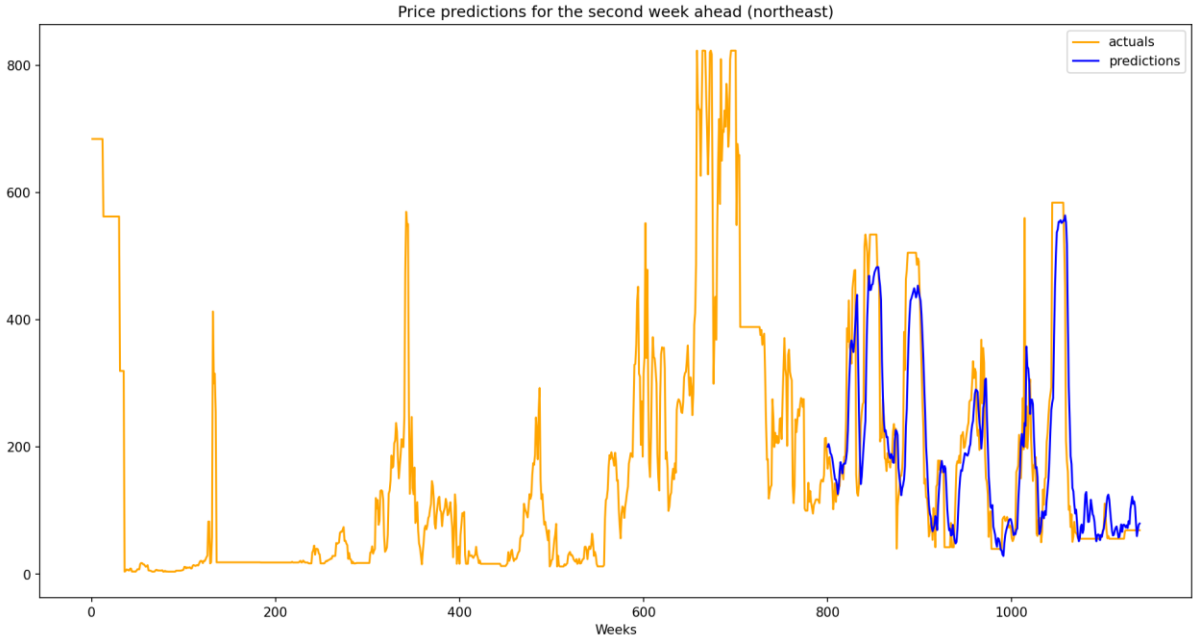


Figure 4-10 *– Simulation test results for predicting settlement price of the second week ahead in Northeast sub-market. Source: Author's elaboration and CCEE historical PLD data [37].*

Predictions RMSE Northeast 2: 85.836

This graph illustrates the historical series of the actual weekly Settlement Prices for the 1039 weeks sampled with the orange line and the prediction results with the blue line. Because of the 4 weeks lookback window and 2 weeks look forward window, the training went over the first 795 weeks and the predictions for the last 339 weeks (check Section 3.2.3 for more detailed information about the training configuration). The Root Mean Squared Error, detailed in Section 2.2.4.1, is $85.836 BRL.

4.2.3 Third week

Below, the model was trained to predict the third weekly Settlement Price ahead based on all input data of four previous weeks (check Section 3.2.3 for more detailed information about the training methods).

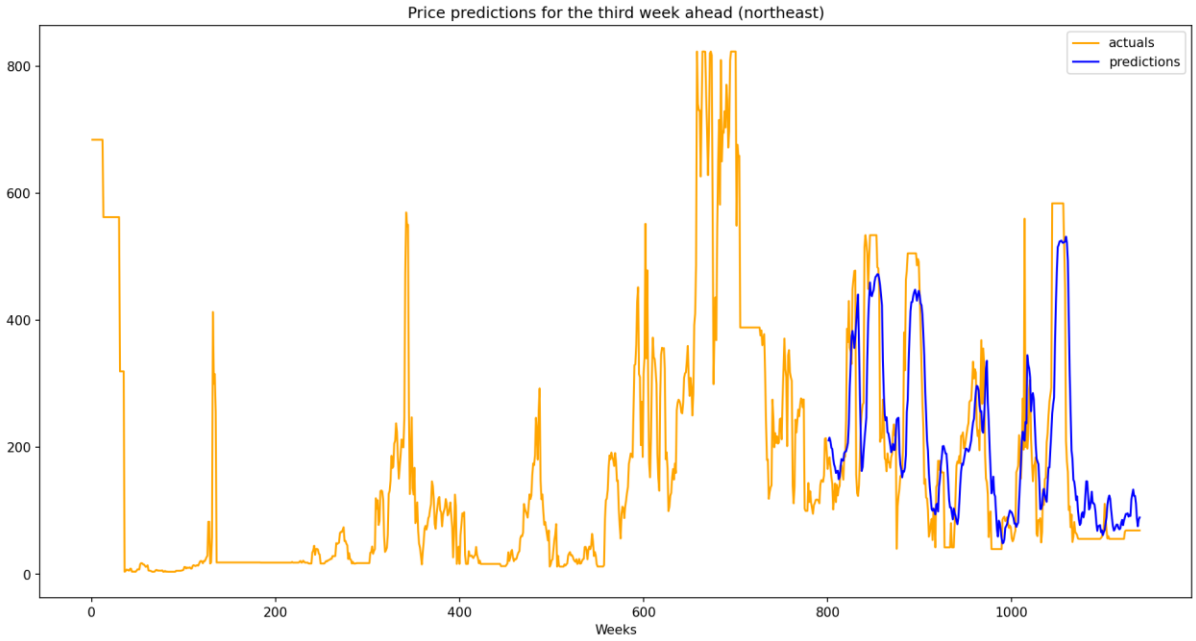


Figure 4-11 *– Simulation test results for predicting settlement price of the third week ahead in Northeast sub-market. Source: Author's elaboration and CCEE historical PLD data [37].*

Predictions RMSE Northeast 3: 101.821

This graph illustrates the historical series of the actual weekly Settlement Prices for the 1039 weeks sampled with the orange line and the prediction results with the blue line. Because of the 4 weeks lookback window and 3 weeks look forward window, the training went over the first 794 weeks and the predictions for the last 339 weeks (check Section 3.2.3 for more detailed information about the training configuration). The Root Mean Squared Error, detailed in Section 2.2.4.1, is $101.821 BRL.

4.2.4 Fourth week

Below, the model was trained to predict the fourth weekly Settlement Price ahead based on all input data of four previous weeks (see Section 3.2.3 for more detailed information about the training methods).

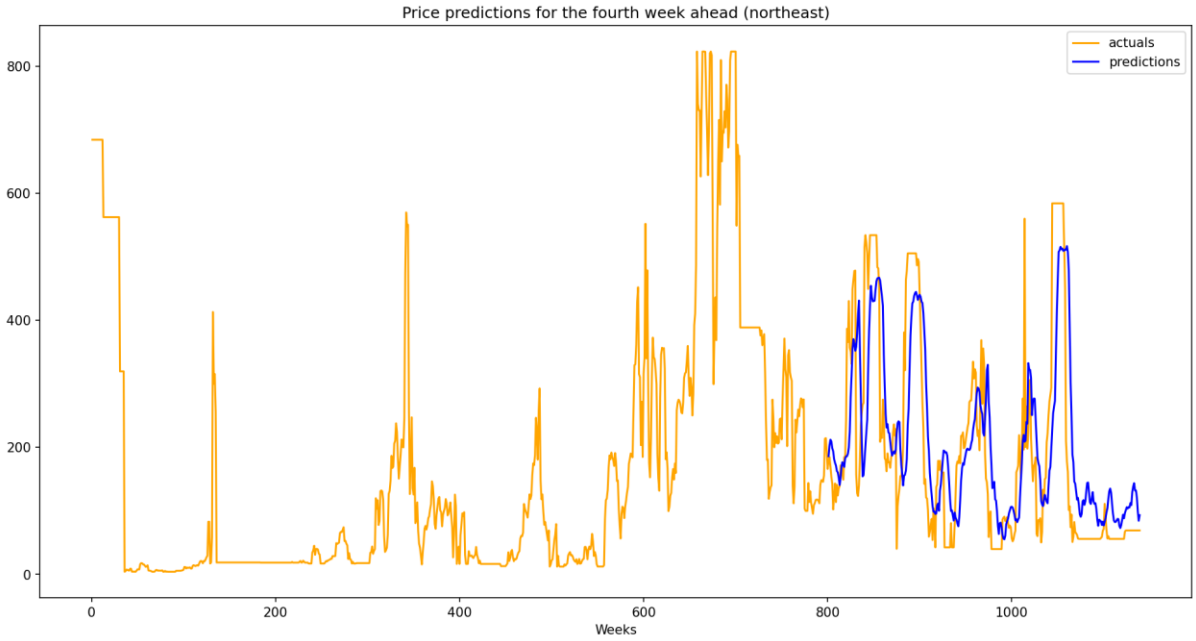


Figure 4-12 *– Simulation test results for predicting settlement price of the fourth week ahead in Northeast sub-market. Source: Author's elaboration and CCEE historical PLD data [37].*

Predictions RMSE Northeast 4: 114.676

This graph illustrates the historical series of the actual weekly Settlement Prices for the 1039 weeks sampled with the orange line and the prediction results with the blue line. Due to the 4-week look-back window and 4-week look-forward window, the training went over the first 793 weeks and the predictions for the last 339 weeks (check Section 3.2.3 for more detailed information about the training configuration). The Root Mean Squared Error, detailed in Section 2.2.4.1, is $114.676 BRL.

Note that the Northeast submarket results also confirm the expected behavior of the RMSE mentioned in Section 4.1.4: for any prediction model, the further away the target, the worse the performance.

4.2.5 Benchmark Analysis

This section compares the results of the proposed LSTM model against the DECOMP model predictions when forecasting the next four weekly prices in the Northeast region. This analysis was done during the first semester of 2021 because that is the closest period from now that the Brazilian electricity prices were showing high volatility. On the other hand, the years 2022 and 2023 have presented floor static prices (see Section 3.1.2.1).

Figure 4-13 *– LSTM vs. DECOMP evaluation on Settlement Price prediction for the next four weeks as of the beginning of every month of the first semester of 2021 in the Northeast. Source: Author's elaboration, CCEE historical data [37] and DECOMP results (Annex).*

The graph below aggregates all seven previous prediction slices for the first semester of 2021.

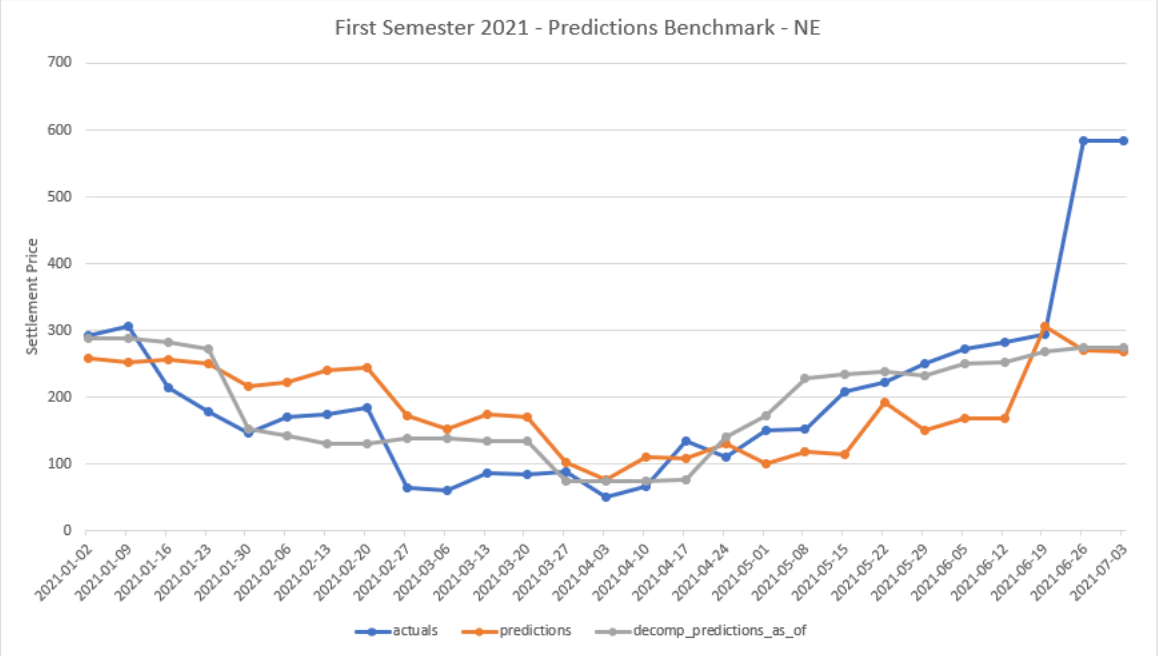


Figure 4-14 *– LSTM vs. DECOMP results evaluation for the first semester of 2021 in the Northeast. Source: Author's elaboration, CCEE historical data [37] and DECOMP results (Annex).*

RMSE LSTM Absolute Values: 107.537

RMSE DECOMP Absolute Values: 94.689

The LSTM results (orange line) present a worse RMSE compared to the DECOMP model results (grey line). It means that the benchmark is closer to the absolute value of the actual prices (blue line).

The following graph shows the Trend Direction Accuracy Measurement for the results of each model. As detailed in Section 2.2.4.2, it indicates how well the model predicts the direction of the price changes.

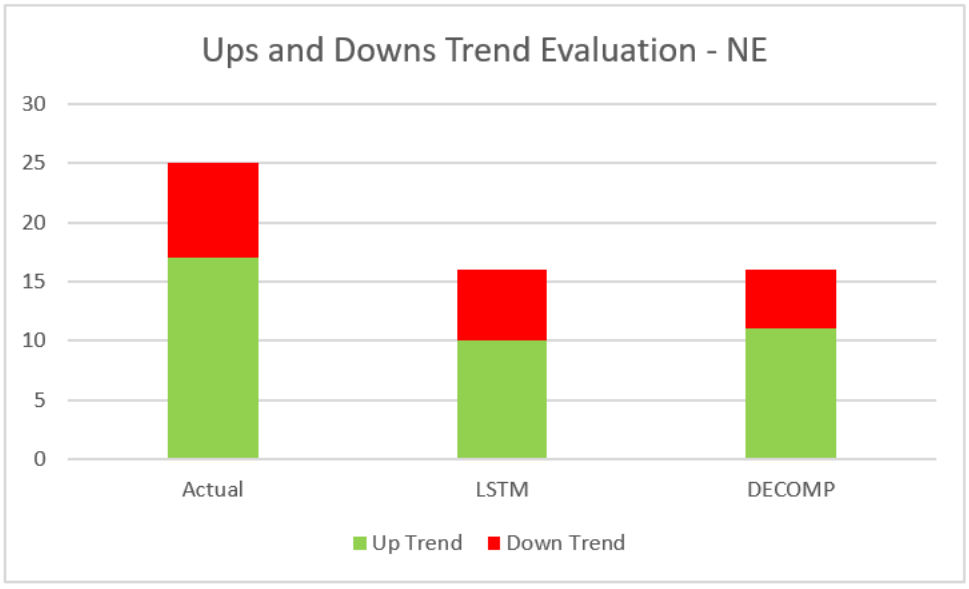


Figure 4-15 *– LSTM vs. DECOMP over trend direction accuracy, Northeast. Source: Author's elaboration, CCEE historical data [37] and DECOMP results (Annex).*

LSTM Trend Direction Accuracy: 64.0 %

DECOMP Trend Direction Accuracy: 64.0 %

The results show that the LSTM model showed equal performance in predicting if the electricity Settlement Price is going up or down, when compared to the DECOMP model. Of 25 trends, the proposed model and the benchmark got 64% of the trends correctly.

The following graph shows the actual weekly price volatility with the grey bars, and the predicted weekly volatility of each model, green bars for LSTM and red bars for DECOMP. Volatility is presented as the percentage of the price rising or falling from one week to the next.

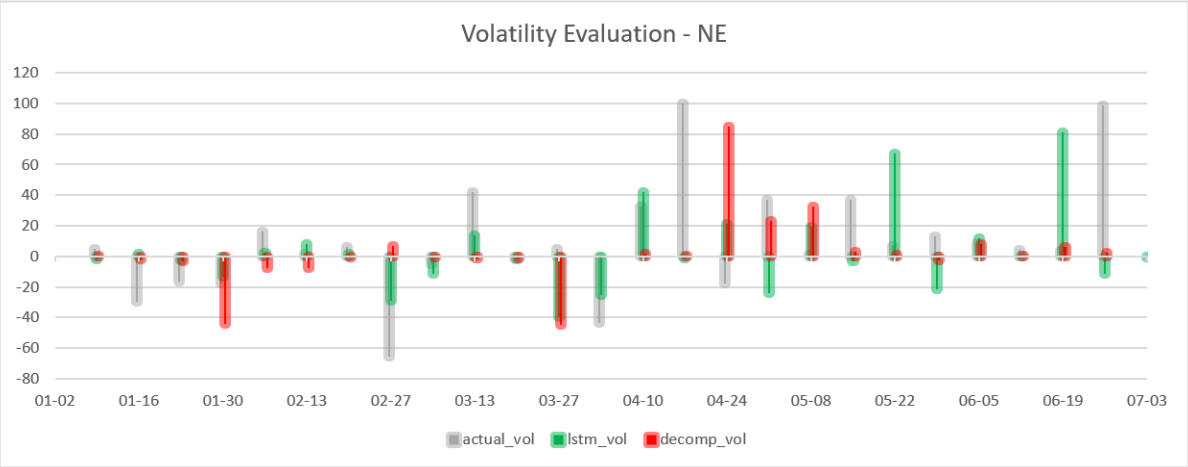


Figure 4-16 *– LSTM vs. DECOMP over volatility accuracy, Northeast. Source: Author's elaboration, CCEE historical data [37] and DECOMP results (Annex).*

RMSE LSTM Vol: 41.591

RMSE DECOMP Vol: 41.837

The RMSE (Section 2.2.4.1) was applied to the volatility predictions of both models to assess their performance. The smaller the RMSE, the greater the accuracy of the model in predicting how intensively the price would change from one week to the next week. The results show that the proposed model performed similarly to the benchmark.

4.3 South

This section shows the results for the South (S) submarket.

4.3.1 First week

Below, the model was trained to predict the first weekly Settlement Price ahead based on all input data of four previous weeks (see Section 3.2.3 for more detailed information about the training methods).

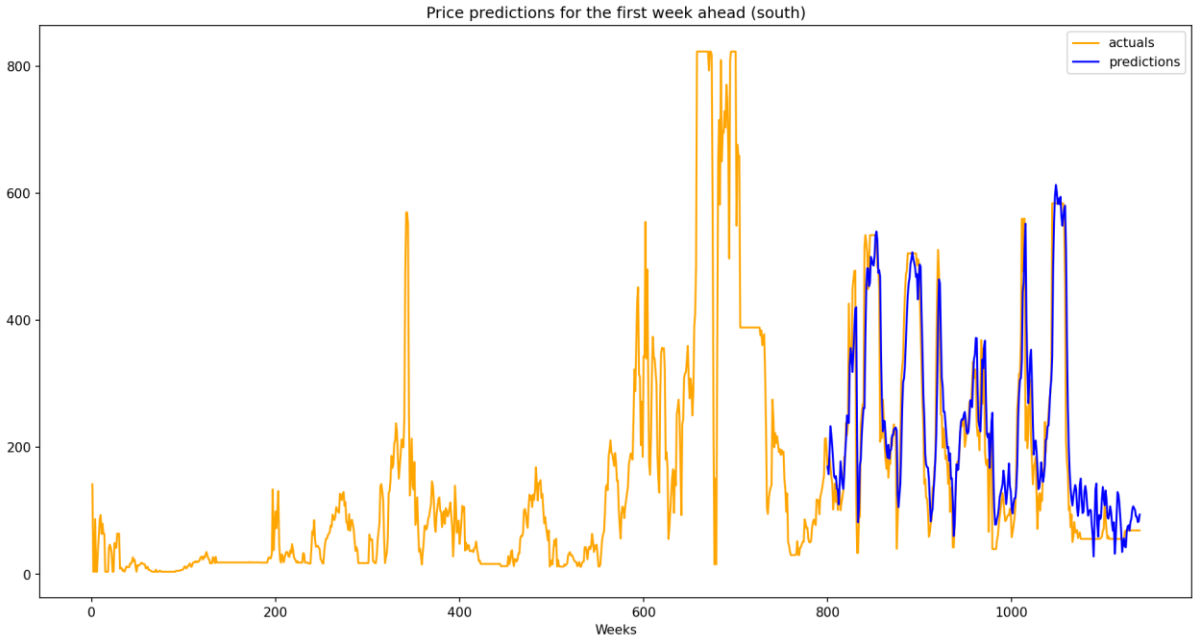


Figure 4-17 *– Simulation test results for predicting settlement price of the first week ahead in South sub-market. Source: Author's elaboration and CCEE historical PLD data [37].*

Predictions RMSE South 1: 70.327

This graph illustrates the historical series of the actual weekly Settlement Prices for the 1039 weeks sampled with the orange line and the prediction results with the blue line. Because of the 4 weeks lookback window and 1 week look forward window, the training went over the first 795 weeks and the predictions for the last 340 weeks (see Section 3.2.3 for more detailed information about the training configuration). The Root Mean Squared Error, detailed in Section 2.2.4.1, is $70.327 BRL.

4.3.2 Second week

Below, the model was trained to predict the second weekly Settlement Price ahead based on all input data of four previous weeks (see Section 3.2.3 for more detailed information about the training methods).

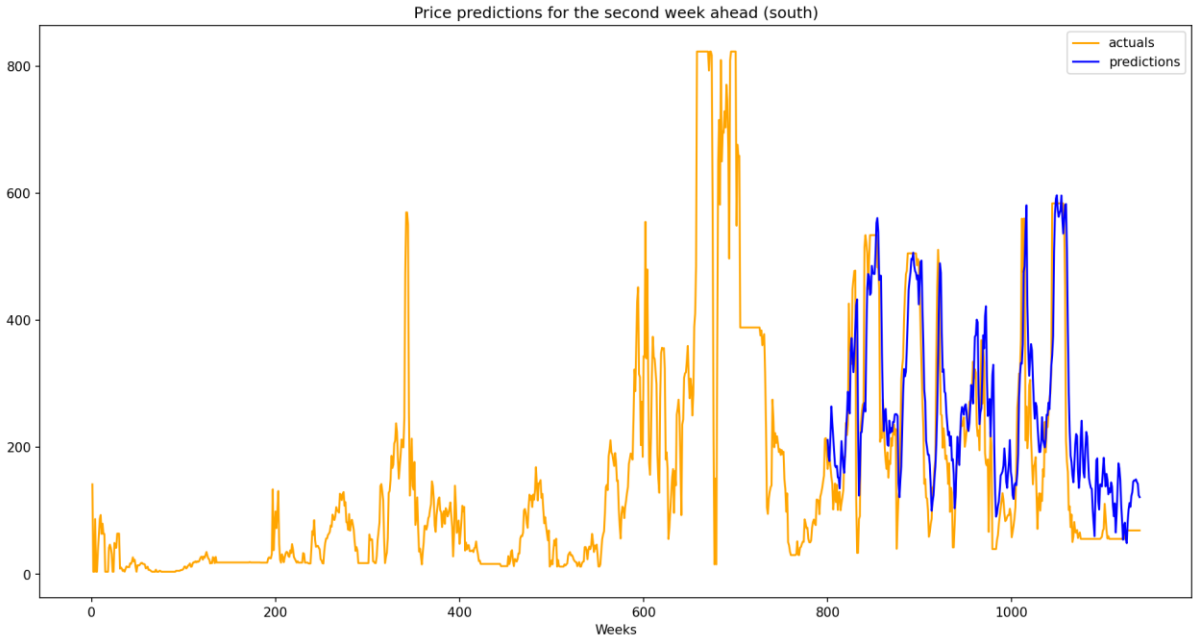


Figure 4-18 *– Simulation test results for predicting settlement price of the second week ahead in South sub-market. Source: Author's elaboration and CCEE historical PLD data [37].*

Predictions RMSE South 2: 104.896

This graph illustrates the historical series of the actual weekly Settlement Prices for the 1039 weeks sampled with the orange line and the prediction results with the blue line. Because of the 4 weeks lookback window and 2 weeks look forward window, the training went over the first 795 weeks and the predictions for the last 339 weeks (check Section 3.2.3 for more detailed information about the training configuration). The Root Mean Squared Error, detailed in Section 2.2.4.1, is $104.896 BRL.

4.3.3 Third week

Below, the model was trained to predict the third weekly Settlement Price ahead based on all input data of four previous weeks (check Section 3.2.3 for more detailed information about the training methods).

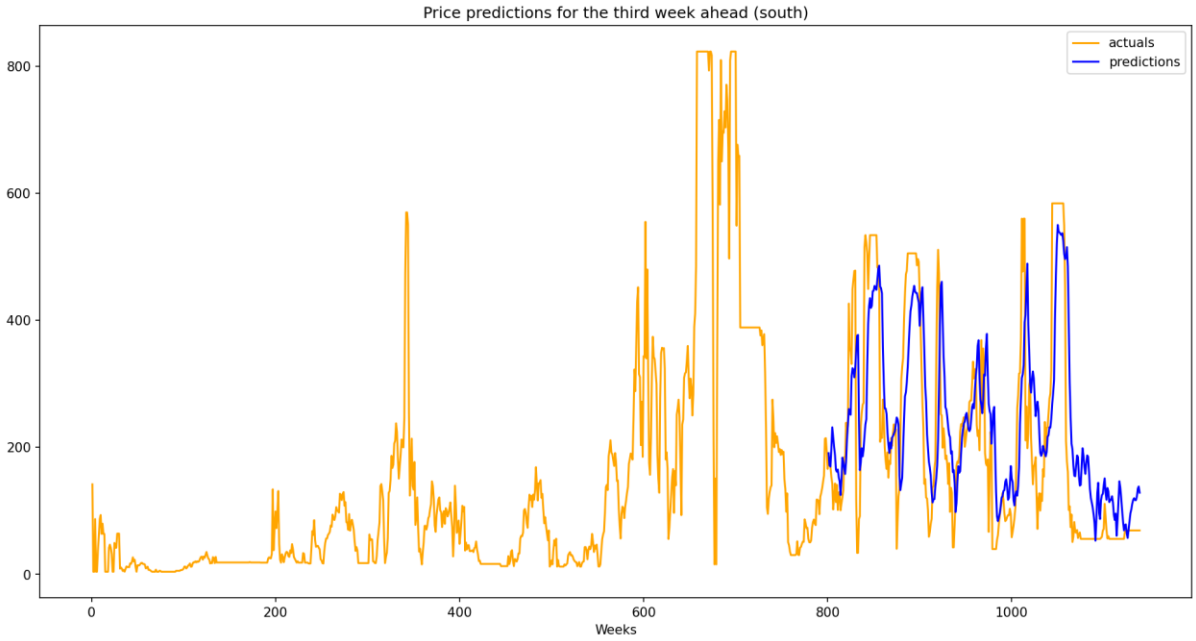


Figure 4-19 *– Simulation test results for predicting settlement price of the third week ahead in South sub-market. Source: Author's elaboration and CCEE historical PLD data [37].*

Predictions RMSE South 3: 115.068

This graph illustrates the historical series of the actual weekly Settlement Prices for the 1039 weeks sampled with the orange line and the prediction results with the blue line. Because of the 4-week lookback window and 3-week look-forward window, the training went over the first 794 weeks and the predictions for the last 339 weeks (check Section 3.2.3 for more detailed information about the training configuration). The Root Mean Squared Error, detailed in Section 2.2.4.1, is $115.068 BRL.

4.3.4 Fourth week

Below, the model was trained to predict the fourth weekly Settlement Price ahead based on all input data of four previous weeks (check Section 3.2.3 for more detailed information about the training methods).

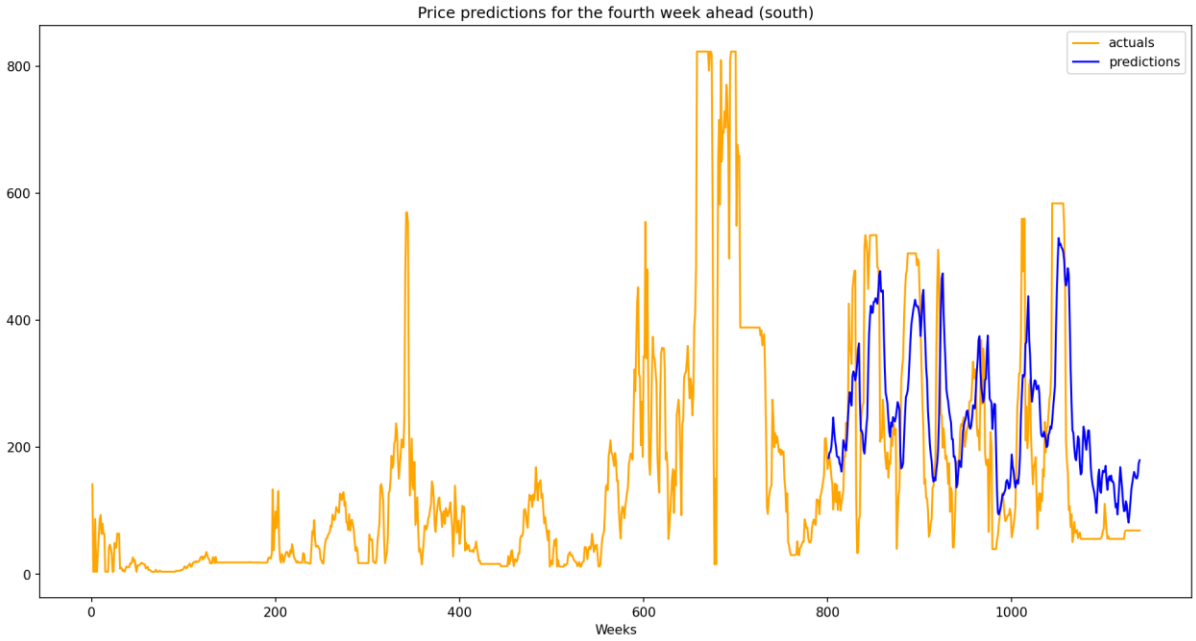


Figure 4-20 *– Simulation test results for predicting settlement price of the fourth week ahead in South sub-market. Source: Author's elaboration and CCEE historical PLD data [37].*

Predictions RMSE South 4: 133.248

This graph illustrates the historical series of the actual weekly Settlement Prices for the 1039 weeks sampled with the orange line and the prediction results with the blue line. Due to the 4-week lookback window and 4-week look-forward window, the training went over the first 793 weeks and the predictions for the last 339 weeks (check Section 3.2.3 for more detailed information about the training configuration). The Root Mean Squared Error, detailed in Section 2.2.4.1, is $133.248 BRL.

Note that the South submarket results also confirm the expected behavior of the RMSE mentioned in Section 4.1.4: for any prediction model, the further away the target, the worse the performance.

4.3.5 Benchmark Analysis

This section compares the results of the proposed LSTM model against the DECOMP model predictions when forecasting the next 4 weekly prices ahead in the South region. This analysis was done during the first semester of 2021 because that is the closest period from now that the Brazilian electricity prices were showing high volatility. On the other hand, the years 2022 and 2023 have presented floor static prices (see Section 3.1.2.1).

Figure 4-21 *– LSTM vs. DECOMP evaluation on Settlement Price prediction for the next four weeks as of the beginning of every month of the first semester of 2021 in the South. Source: Author's elaboration, CCEE historical data [37] and DECOMP results (Annex).*

The graph below aggregates all seven previous prediction slices for the first semester of 2021.

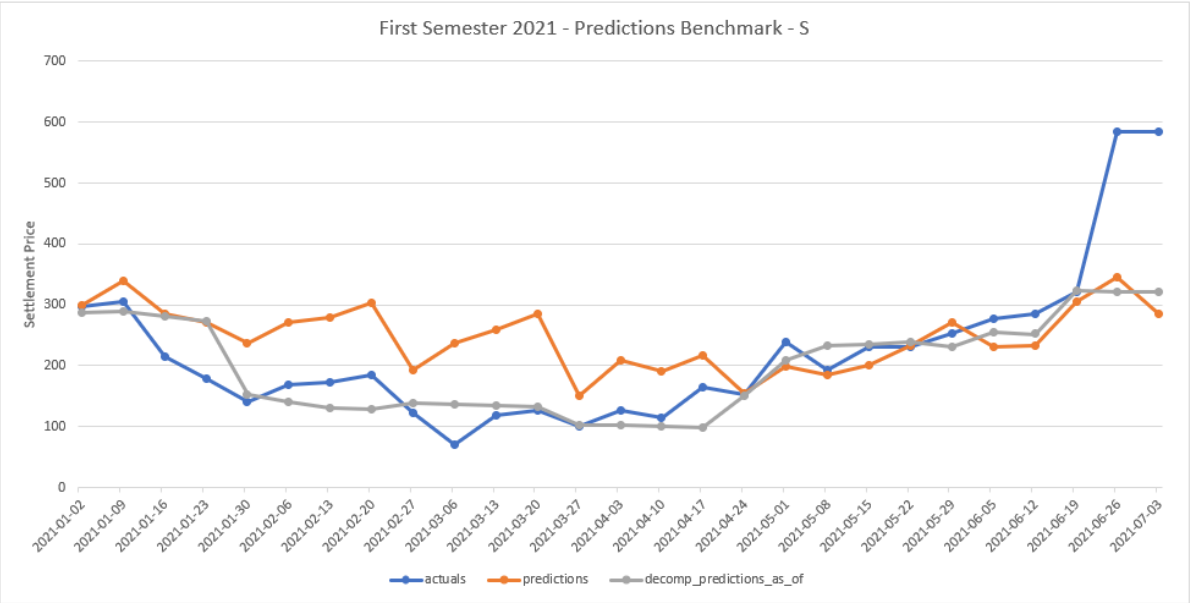


Figure 4-22 *– LSTM vs. DECOMP results evaluation for the first semester of 2021 in the South. Source: Author's elaboration, CCEE historical data [37] and DECOMP results (Annex).*

RMSE LSTM Absolute Values: 106.861

RMSE DECOMP Absolute Values: 79.992

The LSTM results (orange line) present a worse RMSE compared to the DECOMP model results (grey line). It means that the benchmark is closer to the absolute values of the actual prices (blue line).

The following graph shows the Trend Direction Accuracy Measurement for the results of each model. As detailed in Section 2.2.4.2, it indicates how well the model predicts the direction of the price changes.

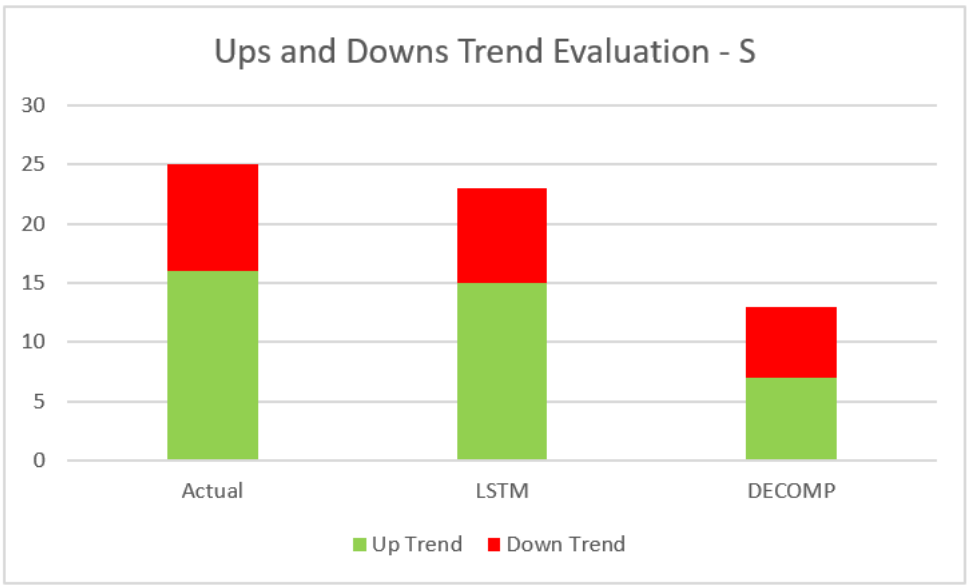


Figure 4-23 *– LSTM vs. DECOMP over trend direction accuracy, South. Source: Author's elaboration, CCEE historical data [37] and DECOMP results (Annex).*

LSTM Trend Direction Accuracy: 92.0 %

DECOMP Trend Direction Accuracy: 52.0 %

The results show that the LSTM model achieved a significantly better performance in predicting whether the electricity settlement price is increasing or decreasing, compared to the DECOMP model. Of 25 trends, the proposed model got 92% of the trends correctly, while the benchmark got 52%.

The following graph shows the actual weekly price volatility with the grey bars, and the predicted weekly volatility of each model, green bars for LSTM and red bars for DECOMP. Volatility is presented as the percentage of the price rising or falling from one week to the next.

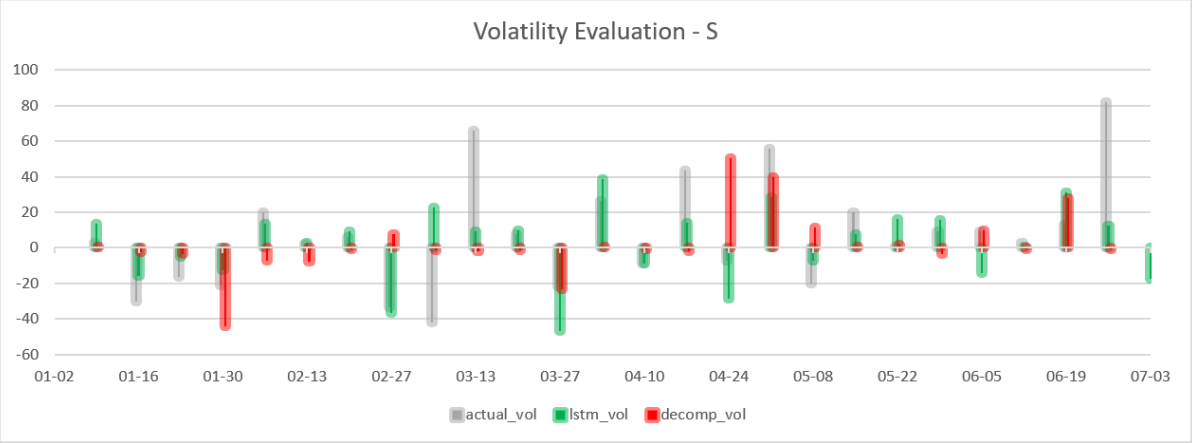


Figure 4-24 *– LSTM vs. DECOMP over volatility accuracy, South. Source: Author's elaboration, CCEE historical data [37] and DECOMP results (Annex).*

RMSE LSTM Vol: 25.380

RMSE DECOMP Vol: 30.582

The RMSE (Section 2.2.4.1) was applied to the volatility predictions of both models to assess their performance. The smaller the RMSE, the greater the accuracy of the model in predicting how intensively the price would change from one week to the next week. The results show that the proposed model performed slightly better than the reference (benchmark).

4.4 North

This section shows the results for the North (N) submarket.

4.4.1 First week

Below, the model was trained to predict the first weekly Settlement Price ahead based on all input data of four previous weeks (see Section 3.2.3 for more detailed information about the training methods).

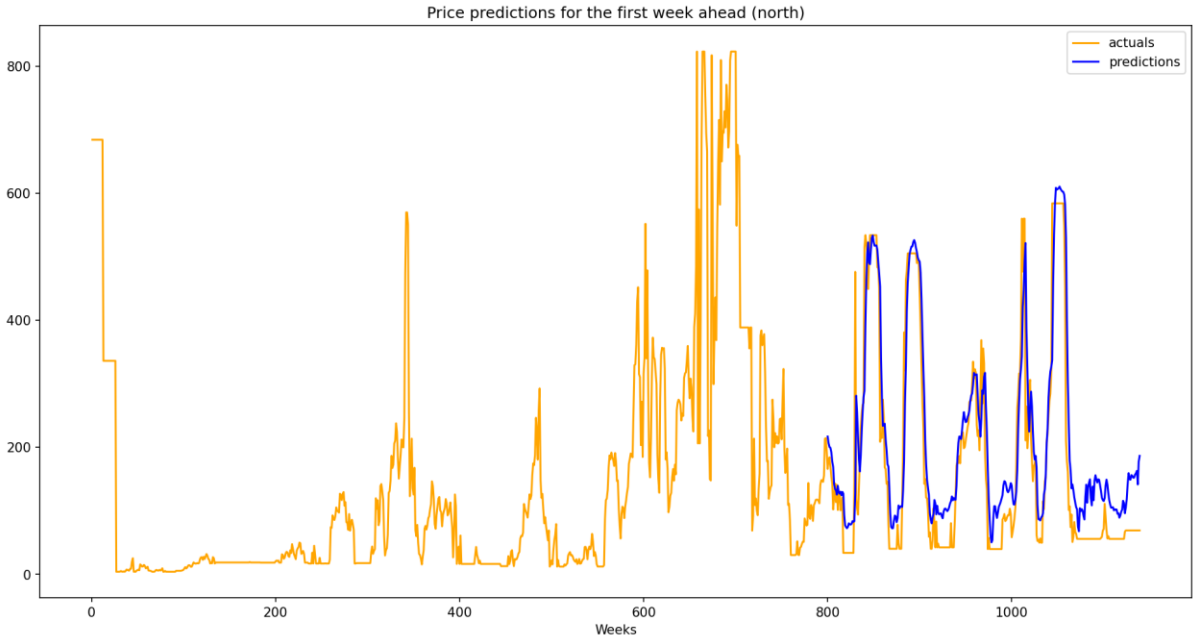


Figure 4-25 *– Simulation test results for predicting settlement price of the first week ahead in North sub-market. Source: Author's elaboration and CCEE historical PLD data [37].*

Predictions RMSE North 1: 72.705

This graph illustrates the historical series of the actual weekly Settlement Prices for the 1039 weeks sampled with the orange line and the prediction results with the blue line. Because of the 4 weeks lookback window and 1 week look forward window, the training went over the first 795 weeks and the predictions for the last 340 weeks (see Section 3.2.3 for more detailed information about the training configuration). The Root Mean Squared Error, detailed in Section 2.2.4.1, is $72.705 BRL.

4.4.2 Second week

Below, the model was trained to predict the second weekly Settlement Price ahead based on all input data of four previous weeks (see Section 3.2.3 for more detailed information about the training methods).

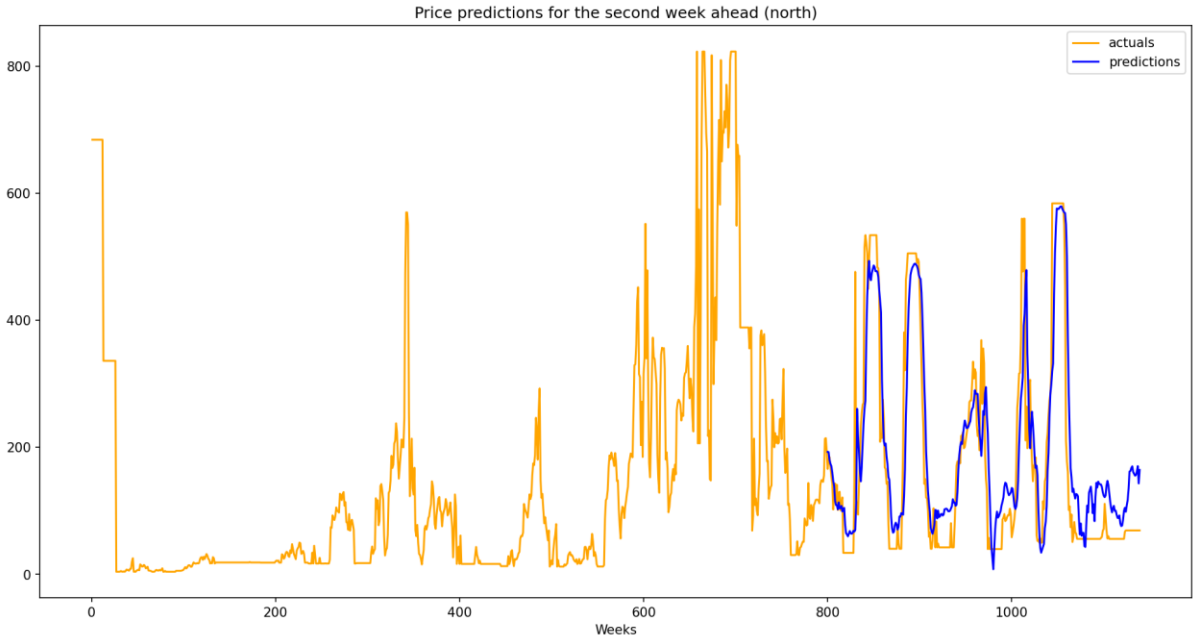


Figure 4-26 *– Simulation test results for predicting settlement price of the second week ahead in North sub-market. Source: Author's elaboration and CCEE historical PLD data [37].*

Predictions RMSE North 2: 85.053

This graph illustrates the historical series of the actual weekly Settlement Prices for the 1039 weeks sampled with the orange line and the prediction results with the blue line. Because of the 4 weeks lookback window and 2 weeks look forward window, the training went over the first 795 weeks and the predictions for the last 339 weeks (check Section 3.2.3 for more detailed information about the training configuration). The Root Mean Squared Error, detailed in Section 2.2.4.1, is $85.053 BRL.

4.4.3 Third week

Below, the model was trained to predict the third weekly Settlement Price ahead based on all input data of four previous weeks (see Section 3.2.3 for more detailed information about the training methods).

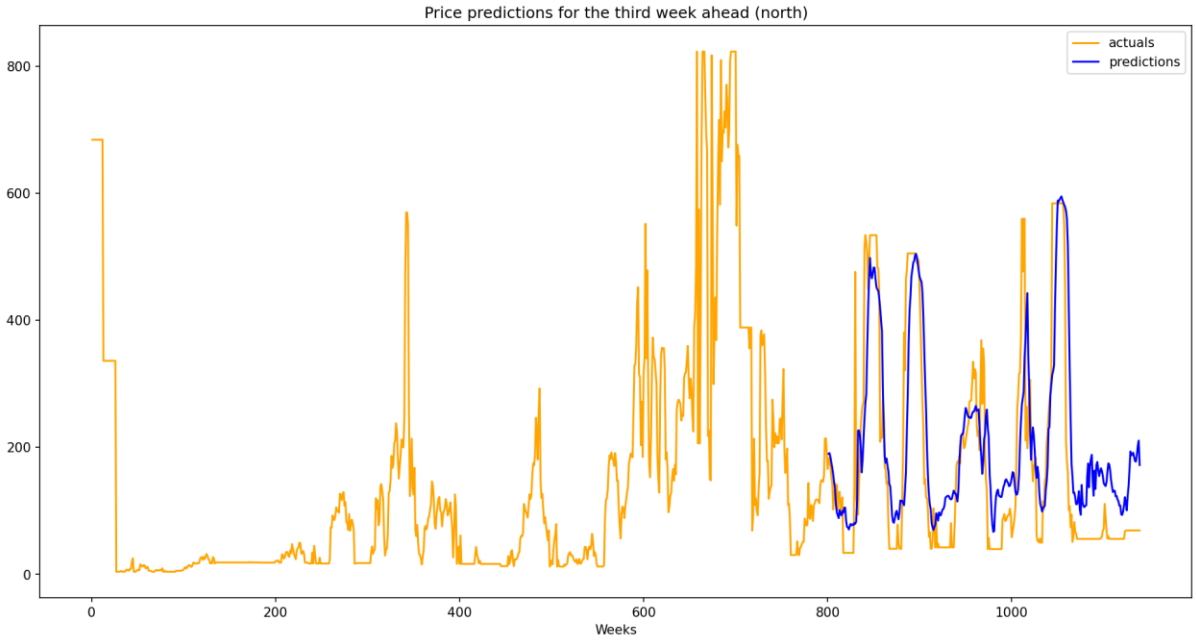


Figure 4-27 *– Simulation test results for predicting settlement price of the third week ahead in North sub-market. Source: Author's elaboration and CCEE historical PLD data [37].*

Predictions RMSE North 3: 103.779

This graph illustrates the historical series of the actual weekly Settlement Prices for the 1039 weeks sampled with the orange line and the prediction results with the blue line. Because of the 4 weeks lookback window and 3 weeks look forward window, the training went over the first 794 weeks and the predictions for the last 339 weeks (check Section 3.2.3 for more detailed information about the training configuration). The Root Mean Squared Error, detailed in Section 2.2.4.1, is $103.779 BRL.

4.4.4 Fourth week

Below, the model was trained to predict the fourth weekly Settlement Price ahead based on all input data of four previous weeks (see Section 3.2.3 for more detailed information about the training methods).

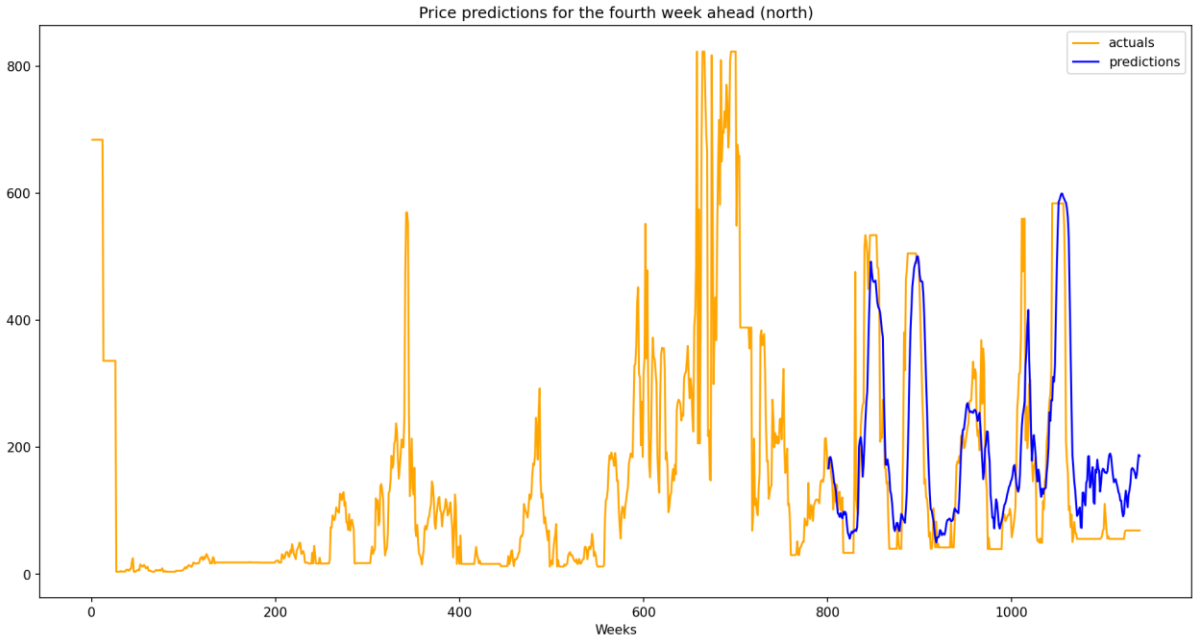


Figure 4-28 *– Simulation test results for predicting settlement price of the fourth week ahead in North sub-market. Source: Author's elaboration and CCEE historical PLD data [37].*

Predictions RMSE North 4: 115.267

This graph illustrates the historical series of the actual weekly Settlement Prices for the 1039 weeks sampled with the orange line and the prediction results with the blue line. Due to the 4-week lookback window and 4-week look-forward window, the training went over the first 793 weeks and the predictions for the last 339 weeks (check Section 3.2.3 for more detailed information about the training configuration). The Root Mean Squared Error, detailed in Section 2.2.4.1, is $115.267 BRL.

Note that the North submarket results also confirm the expected behavior of the RMSE mentioned in Section 4.1.4: for any prediction model, the further away the target, the worse the performance.

4.4.5 Benchmark Analysis

This section compares the results of the proposed LSTM model against the DECOMP model predictions when forecasting the next four weekly prices in the North region. This analysis was done during the first semester of 2021 because that is the closest period from now that the Brazilian electricity prices were showing high volatility. On the other hand, the years 2022 and 2023 have presented floor static prices (see Section 3.1.2.1).

Figure 4-29 *– LSTM vs. DECOMP evaluation on Settlement Price prediction for the next four weeks as of the beginning of every month of the first semester of 2021 in the North. Source: Author's elaboration, CCEE historical data [37] and DECOMP results (Annex).*

The graph below aggregates all seven previous prediction slices for the first semester of 2021.

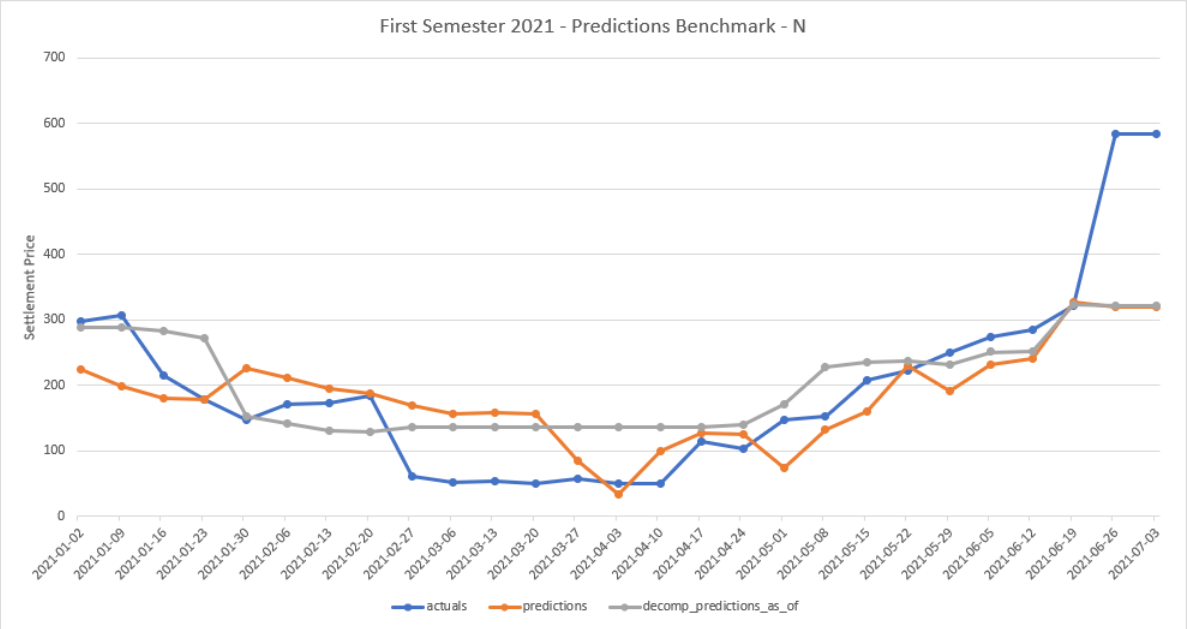


Figure 4-30 *– LSTM vs. DECOMP results evaluation for the first semester of 2021 in the North. Source: Author's elaboration, CCEE historical data [37] and DECOMP results (Annex).*

RMSE LSTM Absolute Values: 92.258

RMSE DECOMP Absolute Values: 89.607

The LSTM results (orange line) present a worse RMSE compared to the DECOMP model results (grey line). It means that the benchmark is closer to the absolute value of the actual prices (blue line).

The following graph shows the Trend Direction Accuracy Measurement for the results of each model. As detailed in Section 2.2.4.2, it indicates how well the model predicts the direction of the price changes.

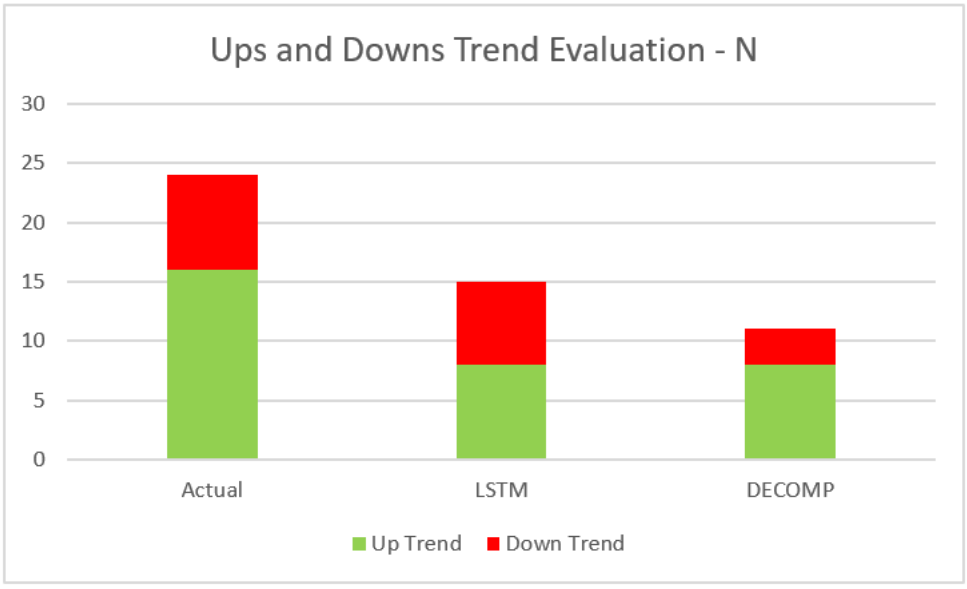


Figure 4-31 *– LSTM vs. DECOMP over trend direction accuracy, North. Source: Author's elaboration, CCEE historical data [37] and DECOMP results (Annex).*

LSTM Trend Direction Accuracy: 62.5 %

DECOMP Trend Direction Accuracy: 45.8333 %

The results show that the LSTM model presented a better performance in predicting whether the electricity settlement price is going up or down, when compared to the DECOMP model. Of 25 trends, the proposed model correctly got 62.5% of them, while the benchmark got 45.83%.

The following graph shows the actual weekly price volatility with the grey bars, and the predicted weekly volatility of each model, green bars for LSTM and red bars for DECOMP. Volatility is presented as the percentage of the price rising or falling from one week to the next.

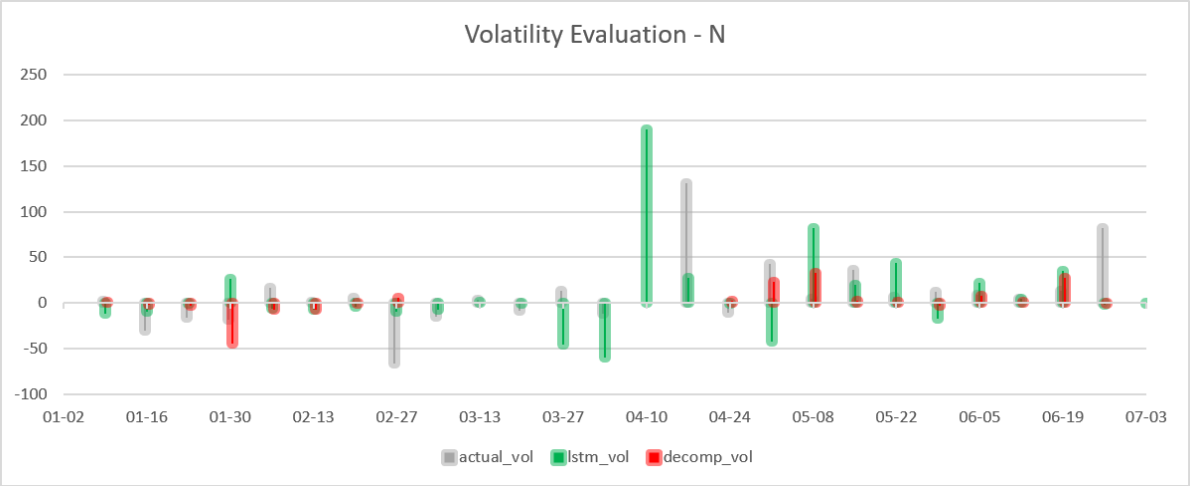


Figure 4-32 *– LSTM vs. DECOMP over volatility accuracy, North. Source: Author's elaboration, CCEE historical data [37] and DECOMP results (Annex).*

RMSE LSTM Vol: 55.534

RMSE DECOMP Vol: 36.078

The RMSE (Section 2.2.4.1) was applied to the volatility predictions of both models to assess their performance. The smaller the RMSE, the greater the accuracy of the model in predicting how intensively the price would change from one week to the next week. The results show that the proposed model performed worse than the benchmark.

4.5 Comparative Numerical Analysis

This section is focused on a numerical comparative analysis of the results presented in this chapter.

4.5.1 Look-Forward Window

The LSTM model training led to different Root Mean Squared Errors (RMSE) for the different look-forward window configurations for each submarket. Below, Table 4-1 brings these results side by side.

Submarket/RMSE Target Week	RMSE Week 1	RMSE Week 2	RMSE Week 3	RMSE Week 4
Southeast-Midwest	70.04	101.093	126.74	141.511
Northeast	65.592	85.836	101.821	114.676
South	70.327	104.896	115.068	133.248
North	72.705	85.053	103.779	115.267

Table 4-1 – Root Mean Squared Error of training results for each submarket by targeted week ahead. Source: Author's elaboration.

The differences among the submarkets are not significant. However, it is important to notice that the indicator increases as the training setting changes. More specifically, as the look-forward window gets further away, the prediction error increases. See in Figure 4-33 how the RMSE behaves.

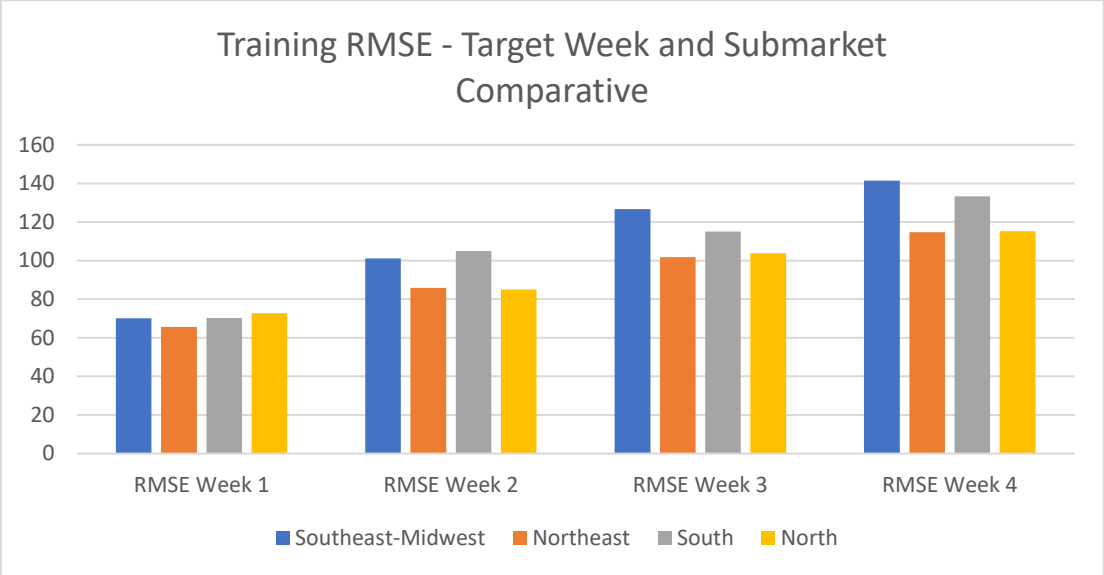


Figure 4-33 *– Graphic view of the Root Mean Squared Error of training results for each submarket by targeted week ahead. Source: Author's elaboration.*

As mentioned in Section 4.1.4, it is expected that the further away the predictive target, the worse would be the predictive performance.

4.5.2 Absolute Values

Throughout the Benchmark Analyses it was clear that DECOMP and LSTM models had different performances regarding the assertiveness between predicted and actual prices.

Following, Table 4-2 shows the RMSE for both models by submarket.

Submarket/RMSE Model	RMSE LSTM	RMSE DECOMP
Southeast-Midwest	107.766	79.395
Northeast	107.537	94.689
South	106.861	79.992
North	92.258	89.607

Table 4-2 – Root Mean Squared Error of settlement price absolute values for each submarket by predictive model. Source: Author's elaboration.

DECOMP model presented smaller error values than LSTM for all submarkets. However, it is relevant to notice that both models performed very close for the Northeast and North regions, while for Southeast-Midwest and South DECOMP was significantly better.

Below, these results are shown graphically in Figure 4-34:

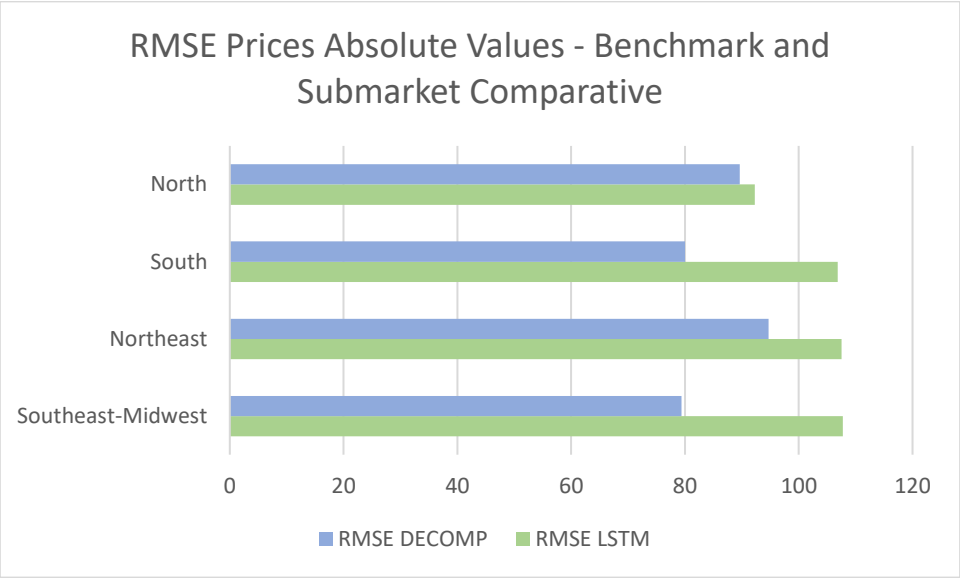


Figure 4-34 *– Graphic view of Root Mean Squared Error of settlement price absolute values for each submarket by predictive model. Source: Author's elaboration.*

4.5.3 Trend Direction

The Benchmark Analyses also showed that DECOMP and LSTM models had different performances regarding the assertiveness of the price trend direction, that is, to predict if the price would rise or fall.

Following, Table 4-3 shows the TDAM (refer to Section 2.2.4.2) for both models by submarket.

Submarket	TDAM LSTM	TDAM DECOMP
Southeast/Midwest	76	52
Northeast	64	64
South	92	52
North	62.5	45.83

Table 4-3 – Trend Direction Accuracy Measurement (TDAM) for each submarket by predictive model. Source: Author's elaboration.

In this regard, LSTM results were very promising. The proposed model presented better results than DECOMP, especially for the Southeast-Midwest and South submarkets. For the Northeast both models had similar assertiveness, and for the North, LSTM was slightly better.

Below, these results are shown in Figure 4-35:

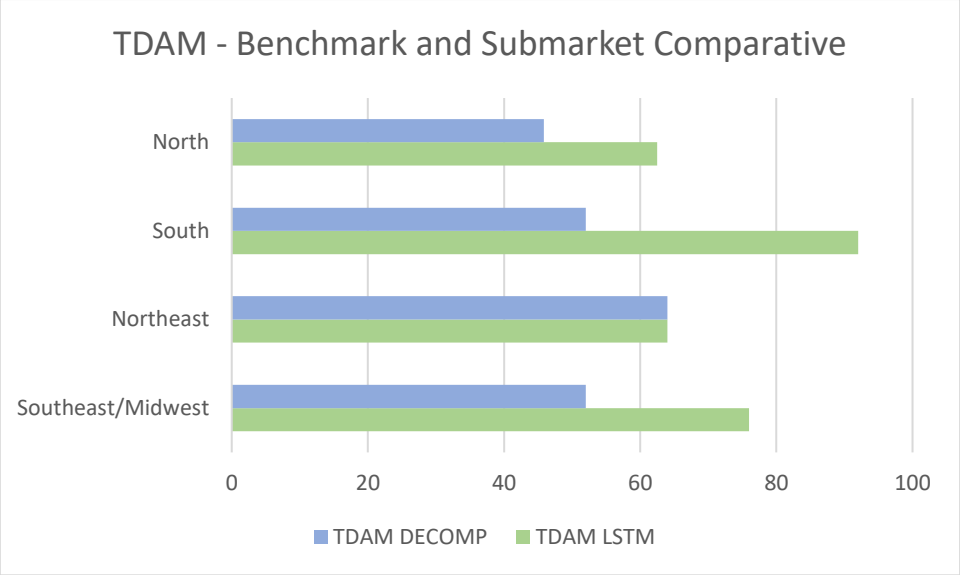


Figure 4-35 *– Graphic view of Trend Direction Accuracy Measurement (TDAM) for each submarket by predictive model. Source: Author's elaboration.*

4.5.4 Volatility

Moreover, the Benchmark Analyses showed that DECOMP and LSTM models had different performances regarding the assertiveness of the price volatility, that is, to predict how much the price would rise or fall in relation to its previous value.

Following, Table 4-4 shows the Volatility RMSE for both models by submarket.

Submarket	VOL RMSE LSTM	VOL RMSE DECOMP
Southeast/Midwest	26.067	30.415
Northeast	41.591	41.837
South	25.38	30.582
North	55.534	36.078

Table 4-4 - Root Mean Squared Error of settlement price volatility for each submarket by predictive model. Source: Author's elaboration.

These results show that the LSTM model had a better performance in predicting the price volatility for the Southeast-Midwest and South regions; a very similar performance for the Northeast; and a worse significantly worse performance for the North submarket.

Following, these results are shown in Figure 4-36:

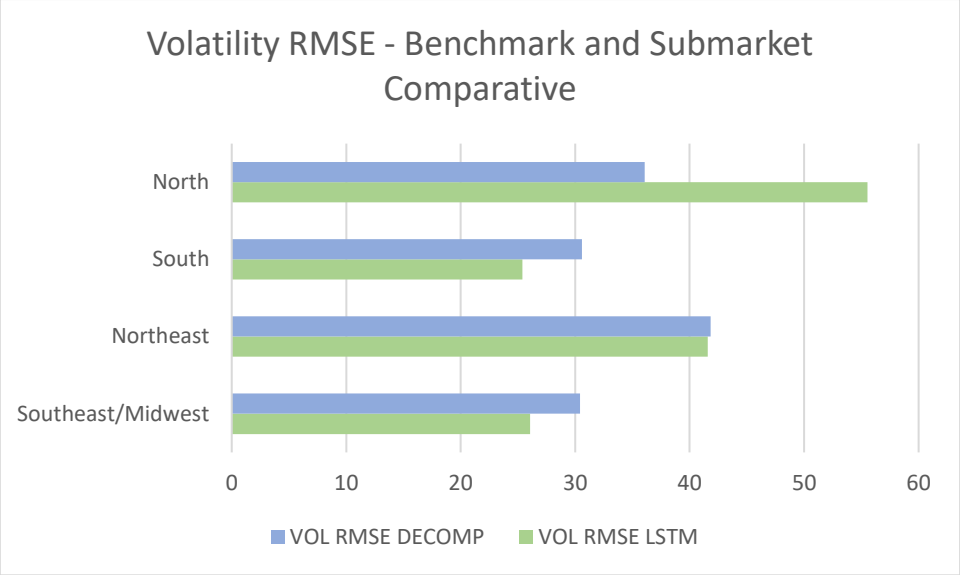


Figure 4-36 *– Graphic view of Root Mean Squared Error of settlement price volatility for each submarket by predictive model. Source: Author's elaboration.*

5 General Discussion

According to the results shown in Chapter 4, the proposed model showed a better performance in predicting price direction and volatility and a worse assertiveness of the absolute price values, compared to the results of the DECOMP model.

Regarding model training, regardless of the region, it was clear that there is a gradual worsening of RMSE as the look-forward window increases. In other words, the farther away the week is to be the target of the forecast, the less accurate the results.

In the comparative analysis, the benchmark demonstrated a better performance regarding the RMSE of the price levels predictions - except for the North region, where both models had a very similar performance in this aspect. This may be connected to the fact that the North is the only region that presented an exclusive strong correlation between the Settlement Price and Thermoelectrical Power Generation. It is possible that the LSTM model was able to make a better predictive learning approach due to its simpler variable correlations.

Regarding the assertiveness in predicting the direction of price change from week to week, the LSTM model demonstrated a significantly better performance than DECOMP. Except for the Northeast, where both models performed equally, for all other submarkets, the developed model showed surprisingly better results, especially in the South. It means that the proposed model is better able to predict whether the price will rise or fall in the forecast of 4 weeks ahead.

Regarding forecast volatility versus actual volatility, the LSTM model was better in the South and Southeast/Midwest regions. For the Northeast region, both performed very closely, and for the North region, the DECOMP model was better. This means that, in general, both models are very close in this regard.

It is important to highlight that the DECOMP model deals with granularized historical data at the plant level. Furthermore, the information consumed by the model is on a daily basis. In contrast, the LSTM model in this research deals with average data by subregion and on a weekly basis. This is a relevant fact to explain the better predictive learning capacity of the DECOMP model in relation to absolute values.

6 Conclusions

Due to the relevance of the electricity markets for the proper functioning and development of the economy of any country, this work evidenced the complex formation of prices in the short-term market established in Brazil. From computational models managed by the ONS, spot electricity prices are calculated and disclosed for the four different subsystems belonging to the National Interconnected System. As exposed, such prices do not necessarily have a direct relationship between energy supply and demand for electricity, as agreed for many energy markets around the world. Such a shape of operation, combined with the high volatility characteristic of energy prices, brings many risks to be managed by the agents of this market. Whether generators, transmitters, traders, distributors, or consumers of electricity, all need to manage their contract portfolios with exposure to each operative week's Settlement Price level.

This study, using Artificial Neural Network techniques, sought to build a model that would be able to bring some level of predictability to electricity spot prices to the Brazilian market. For this purpose, an LSTM model was developed considering historical variables disclosed by the operating planner. The proposed model was trained with weekly data collected since June of 2001 and carried out projections four weeks ahead for the average Settlement Price levels disclosed and for the four existing subsystems in the Brazilian energy context.

To assess the predictive ability of the model, different performance metrics were considered, and the generated projections were opposed to the actual ONS model, DECOMP, which projects the expected prices considering a huge amount of input variables in a very detailed treatment and is used to set actual prices in the Brazilian electricity market.

From a backtest considering 30% of the total sample collected and analyzing the projections made according to the performance metrics considered, the LSTM model achieved significantly superior results when compared to DECOMP with respect to price trend and volatility. On the other hand, it showed inferior results regarding price absolute values. These results showed minor differences between the subsystems. Therefore, the LSTM model proved to be an accurate technique in forecasting the movement and volatility of the electricity price, but less precise in relation to predicting the price level.

In that regard, the LSTM model developed in this work complied with its main objective of bringing predictability, minimizing the uncertainty, to the spot prices practiced in the different subsystems of the Brazilian electricity market.

7 Future Work

The model proposed in this work is a specific architecture of the many possibilities of Artificial Neural Networks existing in the literature. An alternative for future research would be to compare LSTM Recurrent Networks with other neural network architectures, such as Multilayer Perceptron (MLP) or Non-Linear Auto-Regressive with Exogenous Inputs (NARX), among many other possibilities. The application of hybrid network models, known as Ensembles, could also be applied to the same sample evaluated in this work, to check the possibility of obtaining improved results.

Within the proposed LSTM model, only historical variables were considered, with weekly periodicity. The model could be replicated for variables with smaller periodicity, such as daily or even hourly levels of data, so that the training set would be more robust, which could generate more accurate results for the projections of price levels. This could also open doors to verify the behavior of the model for projections of shorter or longer horizons with different time steps. For example, predict the prices for the next 30 days instead of the next 4 weeks and compare the results.

Towards the input layer, the model could have not only historical data as inputs but using future predicted data could improve its predictive skills. One way to do that would be to use the entire output of the DECOMP model as input for the LSTM model. As DECOMP model is used to build future scenarios for operation planning, it does not only predict the following electricity prices but also the levels of all other variables, such as Load Energy, Hydroelectric Power Generation, Maximum Demand, etc. Therefore, using its output as input for the proposed model could result in a new module to the DECOMP architecture to improve price predictions.

Still considering the input data, it could be helpful to compile other variables that can be strongly correlated with electricity prices, such as inflation rates, US dollar quotes against Brazilian real and the market forward curve for the electricity Settlement Price. Additionally, future researchers could also perform techniques like Random Forest Regression to achieve different forms of assessing the importance of features.

Finally, it could be upstanding for the training results to manipulate the data to make it cleaner for the model to interpret it. For example, excluding periods like 2022 and 2023 in which the prices were flat to a certain level due to regulatory reasons could help the training results. This happens because normally the time series data bring anomalies that the model will read during its training, and it will exert influence on the learning process to predict something that is unusual, losing part of its skill to predict under normal conditions. This kind of data manipulation is known as Feature Engineering in the Data Science field.

8 References

[1] L. P. Rosa, N. F. da Silva, M. G. Pereira, L. D. Losekann, “The Evolution of Brazilian Electricity Market,” in Evolution of Global Electricity Markets, Ed. San Francisco, CA, USA: Sioshansi, 2013, pp. 435-459.

[2] F. C. Munhoz. (May 2021). “Two-settlement system for the Brazilian electricity market”. Presented at Brazilian Electricity Regulatory Agency, SGAN 603 I e J, Brasília, DF, Brazil. [Online]. doi: 10.1016/j.enpol.2021.112234.

[3] M. E. P. Maceiral *et al*., "Ten Years of Application of Stochastic Dual Dynamic Programming in Official and Agent Studies in Brazil – Description of the NEWAVE Program," in 2008 Power Systems Computation Conference (PSCC), Glasgow, Scotland, 2008, pp. 14-18. Accessed: Feb. 25, 2023. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d5a35ca26e8cd2c1875084c1ef77476aa5c17a05

[4] M. E. P. Maceiral *et al*., "Twenty Years of Application of Stochastic Dual Dynamic Programming in Official and Agent Studies in Brazil - Main Features and Improvements on the NEWAVE Model," in 2018 Power Systems Computation Conference (PSCC), Dublin, Ireland, 2018, pp. 1-7, doi: 10.23919/PSCC.2018.8442754.

[5] A. Helseth, A. Cordeiro, G. Melo. “Scheduling Toolchains in Hydro-Dominated Systems - Evolution, Current Status and Future Challenges”. Aug. 2020. Accessed: Feb. 25, 2023. [Online]. Available: https://www.researchgate.net/publication/344025387_Scheduling_Toolchains_in_Hydro-Dominated_Systems_-_Evolution_Current_Status_and_Future_Challenges

[6] L. D. Losekann, “The second reform of the Brazilian electric sector,”. In International Journal of Global Energy Issues. Vol. 29, pp. 75-87, Feb. 2008. doi: 10.1504/IJGEI.2008.016342.

[7] OECD (2021). “Driving Performance at Brazil’s Electricity Regulatory Agency,” in The Governance of Regulators. Organization for Economic Co-operation and Development (OECD). OECD Publishing, Paris, France. Oct. 2020. Accessed: Mar. 10, 2023. [Online]. Available: https://www.oecd-ilibrary.org/sites/11824ef6-en/index.html?itemId=/content/publication/11824ef6-en doi: 10.1787/11824ef6-en

[8] ONS (2023). “Sobre o ONS – Atuação,” Operador Nacional do Sistema Elétrico (ONS) (in Portuguese). ONS Publishing. 2023. Brazil. Accessed: Mar. 11, 2023. [Online]. Available: https://www.ons.org.br/paginas/sobre-o-ons/atuacao

[9] M. Hochberg, R. Poudineh, “The Brazilian electricity market architecture: An analysis of instruments and misalignments,” in Utilities Policy. Vol. 72. Oct. 2021. doi: 10.1016/j.jup.2021.101267

[10] L. Rohter. “Electricity Rationing in Brazil Inflames Regional Animosities,” New York Times Publishing, Nov. 4, 2001. Accessed: Mar. 17, 2023. [Online]. Available: https://www.nytimes.com/2001/11/04/world/electricity-rationing-in-brazil-inflames-regional-animosities.html

[11] EPE (2021). “BEN 2021 Summary Report 2021, Reference year 2020,” Energy Research Office (EPE in Portuguese) Publishing, 2021. Accessed: Mar. 18, 2023. [Online]. Available:

https://www.epe.gov.br/sites-en/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-231/BEN_S%C3%ADntese_2020_EN.pdf

[12] EPE (2021). “2. Energy Consumption,” in 2031 Ten-Year Energy Expansion Plan, Energy Research Office (EPE in Portuguese) Publishing, Brazil, 2021. Accessed: Mar. 25, 2023. [Online]. Available: https://www.epe.gov.br/sites-en/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-245/Relatorio_PDE2031_Cap02_EUS.pdf

[13] EPE (2021). “3. Central Power Generation,” in 2031 Ten-Year Energy Expansion Plan, Energy Research Office (EPE in Portuguese) Publishing, Brazil, 2021. Accessed: Mar. 25, 2023. [Online]. Available: https://www.epe.gov.br/sites-en/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-245/Relatorio_PDE2031_Cap03_EUS.pdf

[14] EPE (2021). “10. Socio-Environmental Analysis,” in 2031 Ten-Year Energy Expansion Plan, Energy Research Office (EPE in Portuguese) Publishing, Brazil, 2021. Accessed: Mar. 26, 2023. [Online]. Available: https://www.epe.gov.br/sites-en/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-245/Relatorio_PDE2031_Cap10_EUS.pdf

[15] ANEEL (May 2023) “Brazilian Energy Matrix,” Dashboard Tool, ANEEL Publishing, Brazil, 2023. Accessed: May 24, 2023. [Online]. Available: https://app.powerbi.com/view?r=eyJrIjoiNjc4OGYyYjQtYWM2ZC00YjllLWJlYmEtYzdkNTQ1MTc1NjM2IiwidCI6IjQwZDZmOWI4LWVjYTctNDZhMi05MmQ0LWVhNGU5YzAxNzBlMSIsImMiOjR9

[16] ONS (2023) “Sobre o SIN,” Operador Nacional do Sistema Elétrico (ONS) (in Portuguese). ONS Publishing. 2023. Brazil. Accessed: Mar. 26, 2023. [Online]. Available: https://www.ons.org.br

[17] ONS, EPE, CCEE (2023) “Previsão de carga para o Planejamento Anual da Operação Energética do Sistema Interligado Nacional 2023-2027,” Operador Nacional do Sistema Elétrico (ONS) (in Portuguese). ONS Publishing. Jan. 2023. Brazil. Accessed: Mar. 27, 2023. [Online]. Available: https://www.ons.org.br/AcervoDigitalDocumentosEPublicacoes/NT_PLAN_2023_2027_EPE_ONS_CCEE_Jan2023.pdf

[18] ONS (2022) “Plano da Operação Energética 2022-2026,” Operador Nacional do Sistema Elétrico (ONS) (in Portuguese). ONS Publishing. 2022. Brazil. Accessed: Mar. 30, 2023. [Online]. Available: https://www.ons.org.br/AcervoDigitalDocumentosEPublicacoes/ONS_PEN_2022_Revisao_05092022.pdf

[19] ONS (2023) “PMO de Janeiro 2023 | Semana Operativa de 07/01/2023 a 13/01/2023,”, Operador Nacional do Sistema Elétrico (ONS) (in Portuguese). ONS Publishing. 2023. Brazil. Accessed: Mar. 30, 2023. [Online]. Available: https://www.ons.org.br/AcervoDigitalDocumentosEPublicacoes/Informe%20do%20PMO%20-%20JAN_2023%20-%20RV1.pdf

[20] P. G. C. Ferreira, F. L. C. Oliveira, R. C. Souza “The stochastic effects on the Brazilian Electrical Sector,” in Energy Economics, Vol. 49, May 2015, pp. 328-335. Accessed: Mar. 30, 2023. doi: 10.1016/j.eneco.2015.03.004

[21] N. G. V. Kampen, “Chapter IV - Markov Processes,” in Stochastic Processes in Physics and Chemistry (Third Edition), North-Holland Personal Library, 2007, pp. 73-95. Accessed: Mar. 30, 2023. doi: 10.1016/B978-044452965-7/50007-6

[22] R. Bellman, R. Kalaba, "On adaptive control processes," in IRE Transactions on Automatic Control, vol. 4, no. 2, pp. 1-9, November 1959, doi: 10.1109/TAC.1959.1104847.

[23] J. F. Benders, "Partitioning procedures for solving mixed-variables programming problems." in Numerische Mathematik (in German), Vol. 4, 1962/63, pp. 238-252. Accessed: Apr. 1, 2023. [Online]. Available: http://eudml.org/doc/131533

[24] CEPEL (2018), “Modelo DECOMP, Manual de Referência,” v. 28, Apr. 2018. Accessed: Apr. 8, 2023. [Online]. Available: http://www.cepel.br/wp-content/uploads/2022/03/DECOMP_ManualMetodologia_2020-01_v30.1.pdf

[25] CCEE (2023), “Regras de Comercialização, Preço de Liquidação das Diferenças,” CCEE, Mar. 2023. Accessed: Apr. 8, 2023. [Online]. Available: https://www.ccee.org.br/documents/80415/919404/00%20-%20Pre%C3%A7o%20de%20Liquida%C3%A7%C3%A3o%20das%20Diferen%C3%A7as_2023.3.0_2023-JAN.pdf/06cbf27f-d61a-8fc4-22fd-e1b9e936260f

[26] S. Russell, P. Norvig, “Artificial intelligence” in Artificial Intelligence a Modern Approach, Pearson Education, Inc., Third Edition, 2010, pp. 1-59.

[27] E. Alpaydin, “Introduction” in Introduction to Machine Learning, Massachusetts Institute of Technology (MIT), Fourth Edition, 2020, pp. 1-20. Accessed: Apr. 15, 2023. [Online]. Available: https://books.google.se/books?hl=en&lr=&id=tZnSDwAAQBAJ&oi=fnd&pg=PR7&ots=F3ZU90auwf&sig=8DVySVYRIf2y68uvTKFNpcwD14A&redir_esc=y#v=onepage&q&f=false

[28] A. Krenker, J. Bešter and A. Kos, “Introduction to the Artificial Neural Networks” in Artificial Neural Networks – Methodological Advances and Biomedical Applications” Faculty of Electrical Engineering, University of Ljubljana Slovenia. Published by InTech, Croatia, Mar. 2011, pp. 3-19. Accessed: Apr. 15, 2023. [Online]. Available: https://www.researchgate.net/profile/Kenji-Suzuki-2/publication/319316102_Artificial_Neural_Networks_-_Methodological_Advances_and_Biomedical_Applications/links/59a42f16aca272a6461bb35e/Artificial-Neural-Networks-Methodological-Advances-and-Biomedical-Applications.pdf#page=15

[29] R. May, G. Dandy and Ho. Maier, “Review of Input Variable Selection Methods for Artificial Neural Networks” in Artificial Neural Networks – Methodological Advances and Biomedical Applications” University of Adelaide, Australia. Published by InTech, Croatia, Mar. 2011, pp. 19-45. Accessed: Apr. 15, 2023. [Online]. Available: https://www.researchgate.net/profile/Kenji-Suzuki-2/publication/319316102_Artificial_Neural_Networks_-_Methodological_Advances_and_Biomedical_Applications/links/59a42f16aca272a6461bb35e/Artificial-Neural-Networks-Methodological-Advances-and-Biomedical-Applications.pdf#page=15

[30] S. Hochreiter, J. Schmidhuber “Long Short-Term Memory” in Neural computation. Dec. 1997. doi: 10.1162/neco.1997.9.8.1735.

[31] K. Rink, “Time Series Forecast Error Metrics You Should Know” in Hands-on Tutorials. Published in Towards Data Science, Oct. 2021. Accessed: Apr. 15, 2023. [Online]. Available: https://towardsdatascience.com/time-series-forecast-error-metrics-you-should-know-cc88b8c67f27

[32] R. Watson, “Quantitative research,” in Nursing standard: official newspaper of the Royal College of Nursing, University of Hull, 2015, pp. 44-48. Accessed: Apr. 15, 2023. [Online]. Available: https://hull-repository.worktribe.com/output/374637/quantitative-research doi: 10.7748/ns.29.31.44.e8681

[33] Computer Science Inc., (Dec. 22, 2019). Accessed: Apr. 20, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=QIUxPv5PJOY

[34] Greg Hogg, University of Waterloo in Ontario, Canada (Oct. 5, 2021). Accessed: Apr. 20, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=c0k-YLQGKjY

[35] Greg Hogg, University of Waterloo in Ontario, Canada (Oct. 8, 2021). Accessed: Apr. 20, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=kGdbPnMCdOg&t=1290s

[36] DigitalSreeni, United States, (Dec. 8, 2020). Accessed: Apr. 22, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=tepxdcepTbY&t=976s

[37] CCEE (2023), "Painel de Preços" CCEE, May. 2023. Accessed: May. 10, 2023. [Online]. Available: https://www.ccee.org.br/precos/painel-precos

[38] H. Ritchie, M. Roser and P. Rosado (2022), "Energy," in Our World in Data. Published online at OurWorldInData.org, 2022. Accessed: May. 10, 2023. [Online]. Available: https://ourworldindata.org/electricity-mix

[39] ONS (2023), "Histórico da Operação". Published by ONS. Accessed: Mar. 11, 2023. [Online]. Available: https://www.ons.org.br/paginas/resultados-da-operacao/historico-da-operacao/dados-gerais

[40] I. K. M. Jais, A. R. Ismail and S. Q. Nisa (2019), "Adam Optimization Algorithm for Wide and Deep Neural Network" in Knowledge Engineering and Data Science, Vol. 2, No 1. Published by Department of Electrical Engineering and Informatics, Negeri Malang University, 2019. Accessed: May. 15, 2023. [Online]. Available: http://journal2.um.ac.id/index.php/keds/article/view/6775 doi: 10.17977/um018v2i12019p41-46

[41] CEPEL (2003), "Manual de Referência do Modelo DESSEM," V. 8.2a. Published by CEPEL, Jul. 2003. Accessed: Jun. 3, 2023. [Online]. Available:https://simsee.org/simsee/biblioteca/Brasil/DC201203/Dessem_comentado.pdf

[42] H. Liu, C. Chen, Y. Li, Z. Duan and Y. Li, "Chapter 9 - Characteristic and correlation analysis of metro loads," in Smart Metro Station Systems, Data Science and Engineering, 2022, pp. 237-267. Accessed: May. 3, 2023. [Online]. doi: 10.1016/B978-0-323-90588-6.00009-3

[43] Chris Chatfield, Andreas S. Weigend, "Time series prediction: Forecasting the future and understanding the past: Neil A. Gershenfeld and Andreas S. Weigend, 1994, 'The future of time series', in A.S. Weigend and N.A. Gershenfeld, eds., (Addison-Wesley, Reading, MA), 1-70.," in International Journal of Forecasting, Volume 10, Issue 1, 1994, pp. 161-163. Accessed: May. 3, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0169207094900582 doi: 10.1016/0169-2070(94)90058-2

[44] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization". Published at the third International Conference for Learning Representations, San Diego, 2015. Accessed: Jun. 3, 2023. [Online]. doi: 10.48550/arXiv.1412.6980

9 Annex

The annex files are available in the GitHub link below. Use the following table for reference.

https://github.com/henrique836/master_thesis_lstm_model.git

File name	Description
FINAL_INPUTS_v2.xls	Entire sample of inputs used in the LSTM model, by submarket.
def_lstm.ipynb	Python Notebook code that defines the LSTM model proposed in this research.
DECOMP_Results.pdf	Results of DECOMP model during the first semester of 2021 (Benchmark). Source:
benchmark_consolidated.xlsx	LSTM and DECOMP results during the first semester of 2021 and analysis by submarket.
graphic_inputs.ipynb	Python Notebook containing code that generated the Independent Variables (inputs) graphs for Section 3.1.2 of this paper.
master_thesis_data_treatment.ipynb	Python Notebook containing code used to manipulate data to build the FINAL_INPUTS_v2.xls file, and other minor manipulations.
Research Proposal Final.pdf	Research Proposal presented to KTH and Aalto universities before starting of the Thesis.
PLD_junho_2001_abril_2023.xls	Entire Settlement Price (PLD) historical data from June/2001 to April/2023. Source: CCEE website.
Numerical_Analysis.xlsx	Excel sheets used for building the tables and figures presented in Section 4.5.