

UNIVERSIDADE FEDERAL DE MINAS GERAIS  
School of Engineering  
Production Engineering Graduate Program

Tiago Silveira Gontijo

**Essays on Electricity Price Forecasting**

Belo Horizonte

2023

Tiago Silveira Gontijo

## **Essays on Electricity Price Forecasting**

Doctoral Thesis presented as partial requirement for obtaining the title of Doctor in Production Engineering, conceded by the Production Engineering Graduate Program of the Universidade Federal de Minas Gerais (UFMG).

Supervisor: Prof. Dr. Marcelo Azevedo Costa

Belo Horizonte

2023

T551e

Tiago Silveira Gontijo

Essays on Electricity Price Forecasting/ Tiago Silveira Gontijo. – Belo Horizonte, 2023-

152p. : il. (algumas color.) ; 30 cm.

Supervisor: Prof. Dr. Marcelo Azevedo Costa

Tese (Doutorado) – Universidade Federal de Minas Gerais – UFMG

Departamento de Engenharia de Produção

Programa de Pós-Graduação, 2023.

1. Data Science. 2. Dynamic Time Scan Forecasting. 3. Electricity Prices. I. Marcelo Azevedo Costa. II. Universidade Federal de Minas Gerais. III. Escola de Engenharia. IV. Development of a methodology to support the decision system for the future energy market



UNIVERSIDADE FEDERAL DE MINAS GERAIS  
Escola de Engenharia  
Programa de Pós-Graduação em Engenharia de Produção

**ATA DE DEFESA DE TESE**

Realizou-se, no dia 02 de março de 2023, às 14:00 horas, online em encurtador.com.br/hLW16, da Universidade Federal de Minas Gerais, a 66ª defesa de tese, intitulada *Essays on Electricity Price Forecasting*, apresentada por TIAGO SILVEIRA GONTIJO, número de registro 2019692699, graduado no curso de CIÊNCIAS ECONÔMICAS, como requisito parcial para a obtenção do grau de Doutor em ENGENHARIA DE PRODUÇÃO, à seguinte Comissão Examinadora: Prof(a). Marcelo Azevedo Costa - Orientador (DEP/UFMG), Prof(a). Anderson Laécio Galindo Trindade (DEP/UFMG), Prof(a). Bruno de Almeida Vilela (UFES), Prof(a). Gustavo de Souza Groppo (COPASA), Prof(a). Marcos Oliveira Prates (DEST/UFMG).

A Comissão considerou a tese:

( X ) Aprovada

( ) Reprovada

**Belo Horizonte, 02 de março de 2023.**

Assinatura dos membros da banca examinadora:

Prof(a). Marcelo Azevedo Costa ( Doutor )

Prof(a). Anderson Laécio Galindo Trindade ( Doutor )

Prof(a). Bruno de Almeida Vilela ( Doutor )

Prof(a). Gustavo de Souza Groppo ( Doutor )

Prof(a). Marcos Oliveira Prates ( Doutor )



Documento assinado eletronicamente por **Anderson Laécio Galindo Trindade, Professor do Magistério Superior**, em 02/03/2023, às 15:58, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Marcelo Azevedo Costa, Professor do Magistério Superior**, em 02/03/2023, às 18:08, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Marcos Oliveira Prates, Professor do Magistério Superior**, em 03/03/2023, às 15:17, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Gustavo de Souza Groppo, Usuário Externo**, em 03/03/2023, às 15:55, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).

---



Documento assinado eletronicamente por **Bruno de Almeida Vilela, Usuário Externo**, em 06/03/2023, às 19:20, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).

---



A autenticidade deste documento pode ser conferida no site [https://sei.ufmg.br/sei/controlador\\_externo.php?acao=documento\\_conferir&id\\_orgao\\_acesso\\_externo=0](https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0), informando o código verificador **2112447** e o código CRC **8DB5A011**.

---

For my parents.

# Acknowledgements

This journey would not have been possible without the support of my family, professors, mentors, and friends. First and foremost, I wish to thank my advisor, Professor Marcelo Azevedo Costa, head of the Decision Support and Reliability Laboratory (LADEC-UFMG), for his insight, patience and editing skills in helping me to structure and write this thesis. For this I am sincerely grateful. I am grateful to Peter for his collaboration. I especially thank Dr. Rodrigues, one of the key people in my academic career. I would like to thank Prof. Renato, who opened the doors of Industrial Engineering to me. To my old friends Alexandre, Andressa, Derival, Lucas, Márcio, Renan, Renato, and Thiago. I would like to thank all my PhD colleagues, Álvaro, Cassius, Hendrigo, and Rodrigo, with whom I have shared moments of happiness. In particular, I thank Dr. De Santis for high-level discussions throughout this thesis. I am grateful for the cooperation of doctors Anderson Trindade, Bruno Vilela, Gustavo Groppo, Ilka Reis, and Marcos Prates. To my parents, thank you for encouraging me in all my pursuits and inspiring me to follow my dreams. This work was partially developed with resources from National Council for Scientific and Technological Development (CNPq) and "Companhia Energética Integrada (CEI)".

*"Those who preach god need god  
Those who preach paece do not have peace  
Those who preach love do not have love."  
Charles Bukowski*



# Resumo

Desenvolver modelos preditivos é uma tarefa complexa, pois envolve a incerteza e o comportamento estocástico de variáveis. Especificamente no que diz respeito às *commodities*, prever com precisão seus preços futuros permite minimizar riscos e estabelecer mecanismos de suporte à decisão mais confiáveis. A discussão sobre este assunto é extensa, e a atenção acadêmica está sendo dada à construção de modelos não paramétricos para serem aplicados aos mercados de energia. Estes modelos apresentaram resultados preditivos promissores, o que justifica esta pesquisa. Diante do exposto, formula-se o seguinte questionamento: Como é possível prever com precisão os preços de energia no mercado *spot* brasileiro? A presente tese fornece uma revisão sistemática da literatura sobre os principais métodos de previsão aplicados ao setor de energia. No presente estudo, foi possível identificar lacunas de pesquisa e, assim, propor novos modelos preditivos. Esta tese apresenta modelos preditivos baseados na ideia de análogos. Os análogos consistem no processo de escanear uma série temporal e então, identificar padrões (os chamados "matches") semelhantes às últimas observações disponíveis. Além disso, a recente teoria hierárquica de previsão de séries temporais foi incorporada, uma vez que muitos bancos de dados de energia têm padrões de dependência bem definidos entre si.

**Palavras-Chave:** análise exploratória de dados, aprendizado de máquina, *big data*, ciência de dados, estatística aplicada, métodos quantitativos, preço da eletricidade, previsão, revisão sistemática da literatura, séries temporais.

# Abstract

Developing predictive models is a complex task since it deals with the uncertainty and the stochastic behavior of variables. Specifically concerning commodities, accurately predicting their future prices allows for risk minimization and establishment of more reliable decision support mechanisms. Discussion of this issue is extensive, and academic attention is being paid to the construction of nonparametric models to be applied to energy markets. They have presented promising predictive results, which justifies this research. Given the above, the following question is formulated: How is it possible to predict energy prices accurately in the Brazilian spot market? The present thesis provides a systematic literature review of the main forecasting methods applied to the energy sector. In the present study, it was possible to identify research gaps and, thus, propose new predictive models. The present thesis presents predictive models based on the idea of analogs. Analogs consist of scanning a time series and identifying patterns (so-called "matches") that are similar to the last available observations. Additionally, the recent hierarchical time series prediction theory has been incorporated, since many energy databases have well-defined dependency patterns.

**Keywords:** applied statistics, big data, data science, exploratory data analysis, electricity price, forecasting, machine learning, quantitative methods, systematic literature review, time series.

# List of Figures

Figure 1 – Thesis structure. . . . .	22
Figure 2 – Flowchart outlining the protocol adopted in this systematic review based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Four-Phase Flow Diagram. . . . .	30
Figure 3 – Dashboard on Electricity Price Forecasting research publications. . . . .	32
Figure 4 – Wordcloud containing the top 100 keywords used in the sample. . . . .	35
Figure 5 – Boxplot of the number of citations per paper over the years (part a). . . . .	36
Figure 6 – Boxplot of the number of citations per paper over the years (part b). . . . .	36
Figure 7 – Collaboration networks between research institutions . . . . .	37
Figure 8 – Collaboration networks between researchers . . . . .	38
Figure 9 – BruteForce algorithm . . . . .	45
Figure 10 – JustInTime algorithm . . . . .	45
Figure 11 – Mueen’s algorithm for similarity search (modified) . . . . .	48
Figure 12 – Flowchart with the detailed process of simulation and analysis of similarity search algorithms . . . . .	49
Figure 13 – Illustration of the DTSF time series scan procedure. . . . .	57
Figure 14 – Example of DTSF application to forecasting a time series. The three colored lines represent the top three analogs correlated to the queried period. The dashed lines are the subsequent observations of the analogs. The forecast is given by the median of the adjusted forecast from the subsequent observations of the top analogs. . . . .	58
Figure 15 – Forecasting methods average $sMAPE$ for each of the 414 hourly time series, ordered by the accuracy of the DTSF method. The proposed method obtained fewer errors for most of the time series in this particular domain of application. . . . .	63
Figure 16 – Average $sMAPE$ (obtained in the 414 hourly time series . . . . .	64
Figure 17 – Methodological procedure to obtain the analyzed sample. . . . .	70
Figure 18 – Time series of the PLD ( $P_{SE}$ ) and ten best matches found using $DTSF$ . . . . .	72
Figure 19 – Temporal WordCloud for the 50 most repeated keywords (all) in the analyzed sample. . . . .	74
Figure 20 – Analysis of Variance (ANOVA): response of $PLD$ prices for the analyzed groups. . . . .	75
Figure 21 – Tukey’s Honestly Significant Difference (HSD) test. . . . .	76
Figure 22 – Comparison between the $DTSF$ and the analyzed forecasting models (benchmark). . . . .	77
Figure 23 – Comparison between the $DTSF$ and the analyzed forecasting models (benchmark). . . . .	78

Figure 24 – Time series display of PLD (R\$/MWh), considering ACF and lag correlation.	87
Figure 25 – Boxplot of annual, monthly, and daily PLD (R/MWh) variations.	88
Figure 26 – <i>sMAPE</i> (%) – Comparison between the <i>DTSF</i> and the benchmark models.	88
Figure 27 – <i>MASE</i> - Comparison between the <i>DTSF</i> and the benchmark models.	89
Figure 28 – Hierarchical aggregation structure for the energy generation in Brazil.	94
Figure 29 – Hierarchical forecasting for electricity generation based on the <i>ARIMA</i> procedure (MAPE). Note: The performance was indicated into a color scale, where green means better values for calculated accuracy, and red means worse accuracy. The intermediate values are colored yellow.	103
Figure 30 – Hierarchical forecasting for electricity generation based on the <i>ETS</i> procedure. Note: The performance was indicated into a color scale, where green means better values for calculated accuracy, and red means worse accuracy. The intermediate values are colored yellow.	104
Figure 31 – Hierarchical forecasting for power generation: electrical subsystem versus forecast horizon.	105
Figure 32 – Annual scientific production.	130
Figure 33 – Average article citations per year.	131
Figure 34 – Sources dynamics.	132
Figure 35 – Sources dynamics.	133
Figure 36 – <i>sMAPE</i> (%) by step - comparison between the <i>DTSF</i> and the benchmark models.	134
Figure 37 – <i>MASE</i> by step - comparison between the <i>DTSF</i> and the benchmark models.	135
Figure 38 – Hierarchical forecasting for power generation: electrical subsystem versus generating source (ARIMA).	136
Figure 39 – Hierarchical forecasting for power generation: electrical subsystem versus generating source (ETS).	137
Figure 40 – Hierarchical forecasting for electricity generation based on the <i>ARIMA</i> procedure (RMSE, MAE, MASE) Note: The performance was indicated into a color scale, where green means better values for calculated accuracy, and red means worse accuracy. The intermediate values are colored yellow.	138
Figure 41 – Hierarchical forecasting for electricity generation based on the <i>ETS</i> procedure (RMSE, MAE, MASE). Note: The performance was indicated into a color scale, where green means better values for calculated accuracy, and red means worse accuracy. The intermediate values are colored yellow.	139

# List of Tables

Table 1 – Products generated by this thesis - Article Journals . . . . .	24
Table 2 – List of descriptors used in the present study. . . . .	30
Table 3 – Highlights of Electricity Price Forecasting publications. . . . .	33
Table 4 – Most relevant words – Author’s keywords. . . . .	34
Table 5 – Nord pool energy submarkets analyzed . . . . .	49
Table 6 – Sample data length of 2,400 (100 days) [hours]. The query length $m$ varies from 3 to 48 [hours]. Each scenario is repeated 100 times, the mean computational times (in ms) are shown in the table. . . . .	50
Table 7 – Sample data length of 24,000 (1,000 days) [hours]. The query length $m$ varies from 6 to 48 [hours]. Each scenario is repeated 100 times, the mean computational times (in ms) are shown in the table. . . . .	51
Table 8 – Summary of $M4$ competition dataset, including time-frequency, minimum length of time series, and forecast horizon of each time series. . . . .	56
Table 9 – Parameters range adopted for $DTSF$ . . . . .	60
Table 10 – The performance of $DTSF$ compared to $M4$ benchmark statistical methods – $sMAPE$ metric. . . . .	62
Table 11 – The performance of $DTSF$ compared to $M4$ benchmark statistical methods – $OWA$ metric. . . . .	62
Table 12 – Average $sMAPE$ obtained in the 414 hourly time series . . . . .	64
Table 13 – Total and average times necessary for fitting the methods. . . . .	65
Table 14 – $DTSF$ , benchmarks, and standards for comparison of the $M4$ Competition. . . . .	72
Table 15 – Comparison between the $DTSF$ and the analyzed forecasting models. . . . .	89
Table 16 – Amounts of power generation in Brazil (GWh). . . . .	94
Table 17 – TD disaggregation proportions according to the historical proportions of the data. . . . .	97
Table 18 – Hierarchical forecasting for electricity generation based on the $ARIMA$ procedure. . . . .	99

# List of abbreviations and acronyms

ACF	Autocorrelation function
ARIMA	Autoregressive integrated moving average model
BEI	Brasil Energia Inteligente
BU	Bottom-up
CCEE	Câmara de Comercialização de Energia Elétrica
CEI	Companhia Energética Integrada
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
DTSF	Dynamic time scan forecasting
EDA	Exploratory data analysis
EPF	Electricity price forecasting
ETS	Error, trend, and seasonality model
FFT	Fourier transformation
GWh	Gigawatt hours
HSD	Honestly significant difference
HTS	Hierarchical time series
ICT	Instituto de Ciência e Tecnologia
IEEE	Institute of Electrical and Electronics Engineers
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MASE	Mean absolute scaled error
MASS	Mueen's algorithm for Similarity Search
MinT	Minimum trace reconciliation
OLS	Ordinary least squares

ONS	Operator of the National System
OPEC	Organization of the Petroleum Exporting Countries
OWA	Overall Weighted Average
PCH	Pequena Central Hidrelétrica small hydropower plant
PLD	Preço de Liquidação de Diferenças difference settlement price
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RMSE	Root mean squared error
SBPO	Symposium on Operational Research
SES	Simple exponential smoothing
SJR	Scientific Journal Rankings
sMAPE	Symmetric Mean Absolute Percentage Error
sNaïve	Seasonal Naïve model
TBATS	Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components model
TD	Top-down
TDFP	Top-down forecast proportions
TDGSA	Top-down Gross-Sohl method A
TDGSF	Top-down Gross-Sohl method F
WLS	Weighted least squares
WOS	Web of Science
WT	Wavelet transform
XGBoost	Extreme Gradient Boosting model

# Contents

<b>1</b>	<b>INTRODUCTION</b>	<b>19</b>
<b>1.1</b>	<b>Context, and problem of the research</b>	<b>19</b>
<b>1.2</b>	<b>Objectives</b>	<b>20</b>
<b>1.3</b>	<b>Methodological path, structure, and contributions of the thesis</b>	<b>20</b>
<b>1.4</b>	<b>Structure of the thesis and chapter contents</b>	<b>22</b>
<b>2</b>	<b>PRELIMINARY RESULTS AND PUBLICATIONS</b>	<b>24</b>
<b>I</b>	<b>SYSTEMATIC LITERATURE REVIEW AND THEORY</b>	<b>26</b>
<b>3</b>	<b>ELECTRICITY PRICE FORECASTING: A SYSTEMATIC REVIEW OF PUBLICATIONS BASED ON TEXT MINING PROCEDURES</b>	<b>27</b>
<b>3.1</b>	<b>Introduction</b>	<b>27</b>
<b>3.2</b>	<b>Materials and Methods</b>	<b>29</b>
3.2.1	Population, sample, and data collection	29
3.2.2	Data treatment and analysis	31
3.2.3	Software and Hardware	31
<b>3.3</b>	<b>Results and Discussion</b>	<b>31</b>
<b>3.4</b>	<b>Conclusions</b>	<b>38</b>
<b>II</b>	<b>EMPIRICAL APPLICATIONS</b>	<b>40</b>
<b>4</b>	<b>SIMILARITY SEARCH IN ELECTRICITY PRICES: AN ULTRA-FAST METHOD FOR FINDING ANALOGS</b>	<b>41</b>
<b>4.1</b>	<b>Introduction</b>	<b>41</b>
<b>4.2</b>	<b>Materials and Methods</b>	<b>43</b>
4.2.1	Similarity profile computation based on the Pearson correlation distance	44
4.2.2	Similarity profile computation based on Euclidean distance	46
4.2.3	Dataset and simulation procedures	48
<b>4.3</b>	<b>Results and discussion</b>	<b>50</b>
<b>4.4</b>	<b>Conclusions</b>	<b>51</b>
<b>5</b>	<b>DYNAMIC TIME SCAN FORECASTING: A BENCHMARK WITH <math>M4</math> COMPETITION DATA</b>	<b>53</b>
<b>5.1</b>	<b>Introduction</b>	<b>53</b>



<b>5.2</b>	<b>Materials and methods</b>	<b>55</b>
5.2.1	<i>M4</i> competition dataset	55
5.2.2	Dynamic time scan forecasting	56
5.2.3	Statistical forecasting methods	59
5.2.4	Model selection procedure	60
5.2.5	Software and hardware	61
<b>5.3</b>	<b>Results and discussion</b>	<b>61</b>
<b>5.4</b>	<b>Conclusions</b>	<b>65</b>
<b>6</b>	<b>APPLICATIONS OF DYNAMIC TIME SCAN FORECASTING ON ELECTRICITY SPOT MARKET: A CASE STUDY BASED ON THE BRAZILIAN DIFFERENCE SETTLEMENT PRICE</b>	<b>67</b>
<b>6.1</b>	<b>Introduction</b>	<b>67</b>
<b>6.2</b>	<b>Materials and Methods</b>	<b>69</b>
<b>6.3</b>	<b>A review of the literature on the Electricity Spot Market</b>	<b>69</b>
6.3.1	Data retrieval	71
6.3.2	Dynamic time scan forecasting methodology and benchmark comparison	71
<b>6.4</b>	<b>Results and Discussion</b>	<b>72</b>
6.4.1	Academic research on electricity spot market	72
6.4.2	Statistical analysis of difference settlement price time series (PLD)	75
6.4.3	Case study: applying <i>DTSF</i> to the heavy southeast energy submarket	76
<b>6.5</b>	<b>Conclusions</b>	<b>79</b>
<b>7</b>	<b>APPLICATION OF A DATA-DRIVEN <i>DTSF</i> AND BENCHMARK MODELS FOR THE PREDICTION OF ELECTRICITY PRICES IN BRAZIL: A TIME-SERIES CASE</b>	<b>80</b>
<b>7.1</b>	<b>Introduction</b>	<b>80</b>
<b>7.2</b>	<b>Materials and Methods</b>	<b>82</b>
7.2.1	Dataset	82
7.2.2	Dynamic Time Scan Forecasting	83
7.2.3	Benchmark Models	84
7.2.4	Forecast Evaluation	85
7.2.5	Hardware and Software	86
<b>7.3</b>	<b>Exploratory data analysis (EDA)</b>	<b>86</b>
7.3.1	Forecasting results	87
<b>7.4</b>	<b>Conclusions</b>	<b>90</b>

<b>III</b>	<b>COMPLEMENTARY STUDIES</b>	<b>91</b>
<b>8</b>	<b>FORECASTING HIERARCHICAL TIME SERIES IN POWER GENERATION</b>	<b>92</b>
<b>8.1</b>	<b>Introduction</b>	<b>92</b>
<b>8.2</b>	<b>Materials and Methods</b>	<b>93</b>
8.2.1	The <i>Bottom – Up</i> (BU) Approach	95
8.2.2	The <i>Top – Down</i> (TD) Approach	96
8.2.3	The Optimal Reconciliation Approaches	97
8.2.4	<i>ARIMA</i> and <i>ETS</i> Formulation	98
8.2.5	Evaluating Forecast Accuracy	100
<b>8.3</b>	<b>Results and Discussion</b>	<b>101</b>
<b>8.4</b>	<b>Conclusions</b>	<b>105</b>
<b>9</b>	<b>CONCLUSIONS</b>	<b>107</b>
	<b>References</b>	<b>109</b>
	<b>APPENDIX</b>	<b>129</b>
	<b>APPENDIX A – COMPLEMENTARY RESULTS</b>	<b>130</b>
<b>A.1</b>	<b>Chapter 3</b>	<b>130</b>
<b>A.2</b>	<b>Chapter 8</b>	<b>134</b>
<b>A.3</b>	<b>Chapter 9</b>	<b>136</b>
	<b>APPENDIX B – APPROVALS</b>	<b>140</b>
	<b>ANNEX</b>	<b>150</b>
	<b>ANNEX A – AUTHOR PROFILE</b>	<b>151</b>
<b>A.1</b>	<b>Brief curriculum vitae</b>	<b>151</b>
<b>A.2</b>	<b>Selected publications</b>	<b>151</b>

# 1 Introduction

## 1.1 Context, and problem of the research

Energy planning policies arouse the interest of regulatory agencies, local governments and the business sector. However, reconciling the interests of all the agents involved is not a simple task (Bhattacharyya, 2019) since, for the management to be fulfilled, it is necessary to achieve simultaneous success in: energy supply, attracting investments, the fiscal balance of the government, and tariff modicity (Rao, 2004). Additionally, investing in renewable energies in the present portends reducing the use of fossil fuels in the future, thus generating a positive externality for society (Tjørring & Gausset, 2015). Therefore, the promotion of energy policies favors regional development and, consequently, an improved standard of living for individuals (Xu et al., 2019).

Due to the complexity of this issue, and the number of variables involved, public policies for energy trading occupy a prominent place in the energy industry since such policies should provide security in the investment environment (Pablo-Romero, Pozo-Barajas, & Yñiguez, 2017). Thus, a safe marketing regime is one that accurately signals the price of electricity to agents, allowing them adequately to remunerate the efficiency, reliability and flexibility of the energy generating sources (Wan, Lin, Wang, Song, & Dong, 2016).

In this context, the Brazilian government defined the attributions of the Electric Energy Trading Chamber (CCEE) with Decree No. 5,177/2004 (Brazil, 2004). One of the *CCEE's* main responsibilities is to account for the amount of electricity sold in the National Interconnected System (SIN), as well as to promote settlement for the operational values of the purchase and sale of electricity in the Short-Term Market (MCP) (Aneel, 2013). The same Decree also establishes that the valuation of the amounts settled in the MCP be used for the Settlement Price of Differences (PLD). This price is calculated weekly by the *CCEE*, considering sub-regional energy markets and load levels to be marketed (Ebert & Sperandio, 2018).

The basis for calculating the *PLD* is the Marginal Operating Cost (CMO), derived from the mathematical methods (Newave and Decomp) used by the National Electric System Operator (ONS) to define the system operation schedule. It should be noted that this arrangement is delimited by a minimum price and a maximum price, established annually by the National Electric Energy Agency (ANEEL) (Aneel, 2013).

Despite its relevance to the free energy market, the Brazilian *PLD* is undergoing reformulations. Accordingly, the Ministry of Mines and Energy (MME) has developed a plan for the modernization of the electrical system with Ordinance No.300/2019 (Brazil, 2019). The proposals include improvements to the existing computational models for the operation of the

national electricity system and adoption of a new method (based on hourly prices) for pricing electricity in the Brazilian spot market. The hourly *PLD* is evolving, and will come into effect completely in 2021. The goal is to bring the price of energy closer to that of the National Electric System (Capeletti, 2019).

The purpose of these methodological arrangements is to stimulate energy pricing in a context of demand response programs (Jordehi, 2019; Kalavani, Mohammadi-Ivatloo, & Zare, 2019); i.e., to assign value to energy according to the moment of production, with higher prices at times of higher demand or lower generation, for example. This should lead to efficiency gains for the electrical system, in the long term. At the same time, the changes in the *PLD* will bring the Brazilian trading system closer to international systems that already adopt hourly prices. These systems include: (i) The Nordic Electricity Market - Nord Pool (Haugom, Molnár, & Tysdahl, 2020); (ii) The Italian Electricity Market - Mercati Energetici Manager (GME) (Ilea, Bovo, et al., 2017); and, (iii) The Iberic Electricity Market – Iberian Electricity Market (MIBEL) (Pastor, Da Silva, Esteves, & Pestana, 2018), among others. The objectives of the present thesis are presented below.

## 1.2 Objectives

To develop a new statistical-computational model that allows an accurate prediction of electricity prices in the spot market, considering the new price structure in Brazil. The specific objectives of the present study are:

- To analyze the state-of-the-art methodology in the field of the electricity spot market.
- To propose an alternative method for finding similar patterns in time series.
- To test dynamic time scan accuracy against the M4 competition dataset
- To investigate the Brazilian electricity spot market.
- To present a new forecasting methodology applied to the Brazilian hourly prices of electricity
- To present a hierarchical model for forecasting power generation.

## 1.3 Methodological path, structure, and contributions of the thesis

This project was developed within the scope of the Academic Doctorate for Innovation Program (DAI), which is a National Council for Scientific and Technological Development (CNPq) initiative that aims to strengthen research, entrepreneurship and innovation in Scientific

and Technological Institutions (ICTs), through the involvement of PhD students in projects of interest to the business sector, through partnerships with companies (CNPq Public Call N° 23/2018). The company participating in this program is: Companhia Energética Integrada (CEI), with headquarters in Belo Horizonte/Minas Gerais, was founded in 2004, with the objective of investing in the segment of electric energy generation through the exploitation of renewable sources (Cei, 2023).

In 2006, the Company purchased its first generation asset, CGH Caquende, installed on Rio Macaúbas, a river in the municipality of Bonfim/MG. Between 2006 and 2012, confirming its vocation for purchasing mature assets, CEI started operating 8 other hydropower plants, both CGH (Central Geradora Hidrelétrica – micro hydropower plant – producing up to 1 MW) and PCH (Pequena Central Hidrelétrica – small hydropower plant – producing from 1 to 30 MW), located in the regions of Rio Casca/MG and south of the state of Minas Gerais. Subsequently, in 2015, CEI, consolidating its position in the electric energy generation market, entered into one of the largest operations in the electric energy industry in Minas Gerais, purchasing 6 PCHs in the Rio Doce basin, in the region of Ouro Preto/MG and Mariana/MG (Companhia Energética Integrada, 2021).

Currently, *CEI* operates 16 hydropower plants in the state of Minas Gerais, with a total of 42.82MW of installed power. All of the plants are fully automated and remotely operated by *BEI* – Brasil Energia Inteligente, a company from *CEI*'s economic group, specialized in providing O&M services for hydropower plants. *CEI*, in addition to generating electric energy, develops greenfield projects in the segments of photovoltaic and hydropower generation. In its quest to open up new frontiers in the Brazilian electric energy sector, *CEI* founded *ATMO* Comercializadora de Energia Elétrica, a company strategically located in São Paulo, in order to handle the sales of electric energy and provide consultancy and advisory services in this field (Companhia Energética Integrada, 2021).

This research presents two innovative points. The first refers to the methodological terms since it will address the construction of new predictive models for electricity prices. Due to the new arrangement of the Brazilian electricity sector, with the recent disclosure of hourly energy prices, this thesis innovates in empirical terms, as it will present an unprecedented forecasting model for the new structure of electricity prices in Brazil (Brasil, 2004b). The introduced models were designed to deal with a high number of observations, which will be relevant in the current Brazilian hourly pricing system.

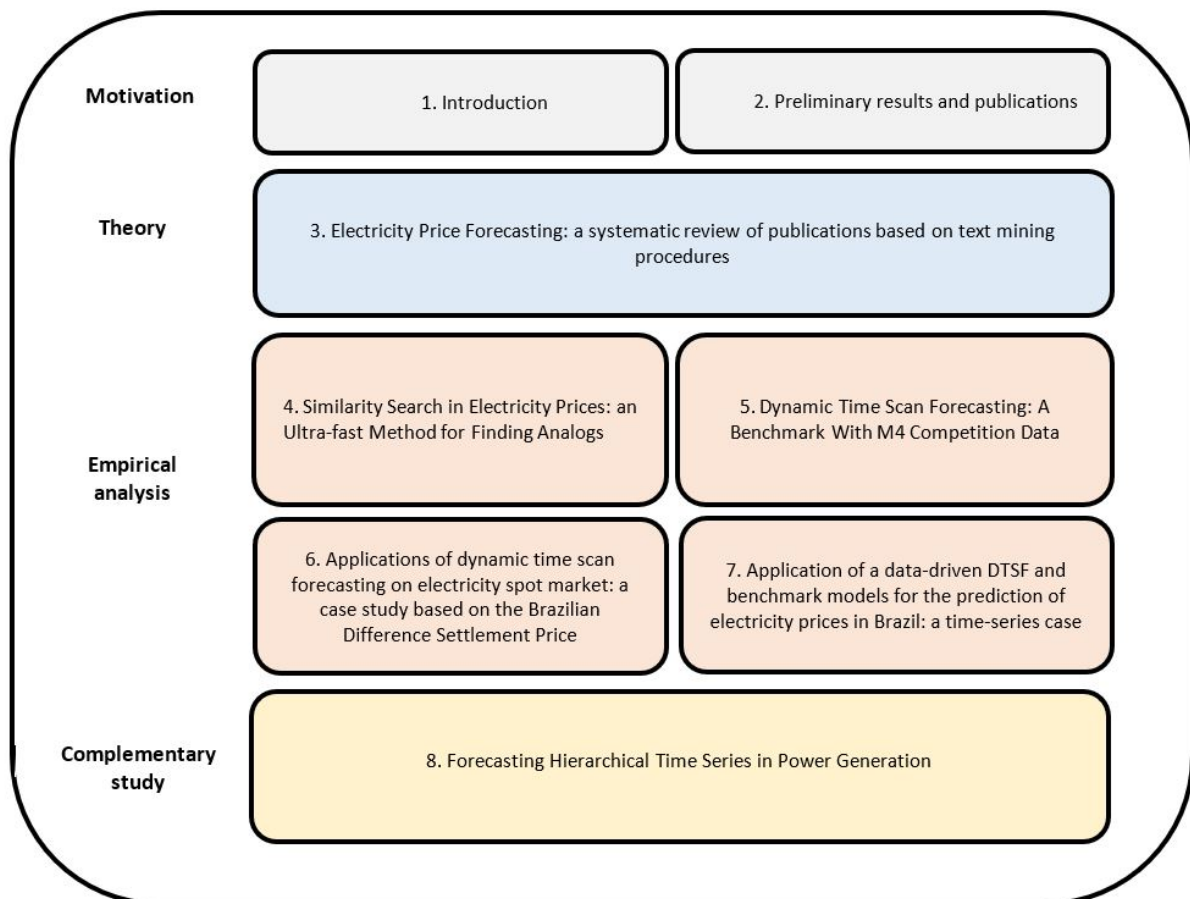
From a practical point of view, this research will generate internal and external impacts to the institution. As an internal impact, the aim is to improve transparency in the decision-making process of energy purchase and sale operators in the future market. As an external impact, it is intended to foster the local development of the University's intellectual capital, encouraging partnerships with the productive sector. Finally, a practical result that will be attempted is the creation of a computational tool. From the academic point of view, the expected result is the

development of a theoretical and methodological framework on how to analyze the decision-making context of energy purchase and sale operators in the future market.

## 1.4 Structure of the thesis and chapter contents

The present thesis is designed as a set of independent essays, each presented in a separate chapter (Figure 1). The studies presented in the essays are complementary, because they address the topic of energy commercialization from different perspectives. The essays are presented in chronological order, to give a clearer reading and understanding of the topic. The general purpose of each chapter is explained, in detail, in the following paragraphs.

Figure 1 – Thesis structure.



Source: Research results.

Chapter 3 presents a systematic review of the literature, providing information about the main authors, countries, and researchers that address the theme of electricity price forecasting. From this review, it was possible to verify the research gaps to be filled by the present thesis.

Chapter 4 addresses the issue of searching for similar patterns and presents a fast algorithm search for long-time series. The search for similarity profiles in time series, especially when based on Pearson's correlation coefficient, connects with the research presented in Chapter

5, which addresses a new methodological tool to be applied to the problem of electricity price forecasting, namely, dynamic time scan forecasting.

Chapter 6 presents a forecasting model for energy prices in the Brazilian spot market, based on weekly prices. It is a preliminary model, based on the search for similar patterns in a time series. Due to the reformulation of the electricity commercialization system in Brazil, Chapter 7 is an update of the previous Chapter, now taking into account prices published on an hourly scale.

Chapter 8 addresses the issue of hierarchical time series and presents a predictive model based on aggregation and disaggregation factors of forecasts. This is a complementary study, which due to methodological limitations, was not deepened.

## 2 Preliminary results and publications

The preliminary publications of this research are presented below, as established in the schedule of the CNPq Public Call. At the end of this project, publication certificates are presented in the attachment section (Table 1).

Table 1 – Products generated by this thesis - Article Journals

<b>Chapter</b>	<b>JCR</b>	<b>Journal</b>	<b>Title</b>
3	-	Revista de Administração, Contabilidade e Economia da Fundace (USP)	Electricity Price Forecasting: a systematic review of publications based on text mining procedures.
4	2.847	Journal of Renewable and Sustainable Energy	Similarity Search in Electricity Prices: an Ultra-fast Method for Finding Analogs.
5	0.967	IEEE Latin America Transactions	Dynamic Time Scan Forecasting: A Benchmark with M4 Competition Data
6	-	E3S Web of Conferences	Electricity price forecasting on electricity spot market: a case study based on the Brazilian Difference Settlement Price
7	-	-	-
8	3.252	Energies	Forecasting Hierarchical Time Series in Power Generation
*	2.021	Proceedings of the Institution of Mechanical Engineers Part O-Journal of Risk and Reliability	Condition-based maintenance in hydroelectric plants: A systematic literature review.
*	3.847	Sensors	A Data-Driven Framework for Small Hydroelectric Plant Prognosis Using Tsfresh and Machine Learning Survival Models.

Source: Research results. Paper \* consists of a publication directly related to this thesis.



Next, the other products generated by this thesis are presented, namely, articles in congresses, books, and awards.

- Chapter 3 was originally featured in the "Encontro de Gestão e Negócios" (EGEN - 2021) at the Federal University of Uberlândia and won the prize for best article in the area of "Production and Logistics". Due to the award, he obtained the fast-track of the Journal RACEF (USP).
- The first version of Chapter 5 was approved and presented at the traditional International Symposium on Forecasting (ISF), in the year 2020.
- Chapter 7 was presented at the 3rd International Conference on Renewable Energy (ICREN - Italy). This article was approved in the congress fast-track and was published in the journal shown in Table 1.
- Chapter 8 is in the submission process.
- The initial version of Chapter 9 was presented at the LII Brazilian Symposium on Operational Research (SBPO) in 2020. This article was later submitted to the Special Issue "Energy Economics and Policy in Developed Countries" in book format by the MDPI group, becoming a chapter.

During the course of my Doctorate, other indirect research was carried out by me, resulting in multidisciplinary publications. To find additional information about it, please verify the Annexes of this thesis.

# Part I

## Systematic literature review and theory

# 3 Electricity Price Forecasting: a systematic review of publications based on text mining procedures

## Abstract

Developing forecasting models is a difficult task. Particularly concerning electricity prices, accurately predicting their forthcoming values makes it possible to minimize planning risks. This fact becomes even more relevant in the current geopolitical scenario, represented by the war between Russia and Ukraine. Given the above, this paper presents a systematic review of the literature on electricity price forecasting (EPF) models. It presents a methodology that does a robust search of the literature, obtaining the most relevant papers ( $n = 554$ ) that addressed this theme. Following that search, we: (i) constructed an attribute matrix of the publications, and (ii) presented a descriptive analysis based on bibliographic data, and network relationships. The sample period comprises the years 1991 to 2019, with an annual growth rate equal to 23.13% and an annual publication rate of 19 papers. Despite the increase in the number of studies on electricity price forecasting, the predominance of papers is produced in only a few countries. This fact reinforces the need to encourage research and development projects related to the energy market. It was also found that research collaboration networks are still weak, highlighting the need for new partnerships between countries, and research institutions. Thus, stimulating global energy security, as well as encouraging cooperation and technology transfer between countries, becomes relevant.

**Keywords:** Electricity price forecasting. Day-ahead market. Systematic Literature Review. Bibliometrix.

## 3.1 Introduction

Commodity prices, in general, exhibit stochastic behavior, meaning that future prices are uncertain and difficult to predict (R. V. Gomes, 2015). The renewable energy market is no different. Due to this complexity, understanding the dynamics of future energy prices in both the short- and long-term markets is of academic, business, and social relevance (Princ & Slabe-Erker, 2020). This fact becomes even more relevant in the current scenario of global energy insecurity, derived from factors such as the war between Russia and Ukraine (Steffen

& Patt, 2022) and the repeated interventions of the Organization of the Petroleum Exporting Countries (OPEC) in oil prices over the last few decades (Zuhaira, Li, & Mohammed, 2022).

Successful electricity price forecasting models are based on different techniques such as: (i) classic time series procedures like the autoregressive moving average, autoregressive integrated moving average, generalized autoregressive conditional heteroscedastic, among others (H. Liu & Shi, 2013; Mišnić, Pejović, Jovović, Rogić, & Đurišić, 2022); (ii) pre-processing techniques like spectrum analysis, wavelets, and Fourier analysis (Miranian, Abdollahzade, & Hassani, 2013; Iwabuchi et al., 2022); and, (iii) machine learning approaches like neural networks, fuzzy systems, and support vector machines (Bui, Tuan, Klempe, Pradhan, & Revhaug, 2016). Additionally, an alternative class of hybrid models (J. Zhang, Tan, & Wei, 2020) aims to combine machine learning (W. Yang, Sun, Hao, & Wang, 2022) representations with different methods. Instances of these methods are focused time-delay neural networks (Y. Chen et al., 2019), neural networks with fuzzy inputs (H. Liu, Tian, Liang, & Li, 2015), finite-impulse response neural networks (Pir, Shah, & Asger, 2017), local feedback dynamic fuzzy neural networks (Nagaraja, Devaraju, Kumar, & Madichetty, 2016), type recurrent fuzzy networks (Jain, Seera, Lim, & Balasubramaniam, 2014; Li, Woo, & Cox, 2021), and neuro-fuzzy inference systems (Moreno & Santos Coelho, 2018), among others.

Due to the economic relevance of the energy market, and the growing interest in renewable sources, the recent development of predictive techniques has attracted the attention of the electricity community (Soeiro & Dias, 2020). This is important because it allows analysis of the behavioral pattern of the prices, as well as comprehension of the evolution of the predictive models used. Thus, a forecasting community has emerged worldwide. To investigate how expert collaboration could be enhanced, it is essential to answer some questions, namely:

- what are the leading countries, authors, and theoretical approaches related to electricity price forecasting?
- what are the gaps in the literature that should be explored, given what has been published so far?

Some papers have begun to address this issue (Weron, 2014; Antonopoulos et al., 2020). The present study approaches the problem using text mining tools such as the Bag-of-words model for language processing and attributes matrices. Benefits include simplifying the representation of substantial textual information. This method has been used in recent literature reviews on innovation (N. J. Van Eck & Waltman, 2017), medicine (Nafade et al., 2018), physics (van Raan, 2017), among others. Based on this framework, we conducted a robust systematic review focusing on the main statistical methods used to predict electricity prices. We also evaluated how the leading agents in this community articulate among themselves.

The present study contributes to the electricity price forecasting literature. The first contribution is methodological, presenting a method for selecting the core publications in the area. It is innovative in that we selected papers from different scientific bases and developed an automatic way to remove duplicates and establish a unified metadata base. Hence, it is possible to establish the flow of knowledge. The second contribution is identifying the most prominent authors, countries, and publications through objective criteria, highlighting the leading research groups.

Finally, our main findings may assist researchers in better understanding the main forecasting techniques and the most important upcoming research topics, arising from this issue. In practical terms, this research will serve as a guide for those interested in the subject, including researchers, policymakers, companies, and other interested parties, showing the leading publications in the area.

The present paper is structured in four sections, as follows. Section 2 explains the methodology and scope of the systematic review. Section 3 presents the papers considered for this review and discusses their main features. Section 4 presents the main findings of the present study, and potential pathways for future studies to explore models, for predicting electricity prices.

## 3.2 Materials and Methods

### 3.2.1 Population, sample, and data collection

An extensive survey of publications, indexed in both the Web of Science (WoS) and the Scopus databases, was conducted. Papers related to electricity price forecasting were evaluated. Table 2 shows the list of descriptors used in this research. This research utilized boolean operators. It was used as conjunctions to combine or exclude keywords in a search: "AND" and "OR". We selected journal publications because they had already gone through a peer-reviewed process.

It is noteworthy that WoS and Scopus are the academic citation databases most used to define a study (Weron, 2014). Data extraction from Scopus and WoS (2020-08-30) considered the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement (Moher, Liberati, Tetzlaff, Altman, & Group, 2009) (Figure 2).

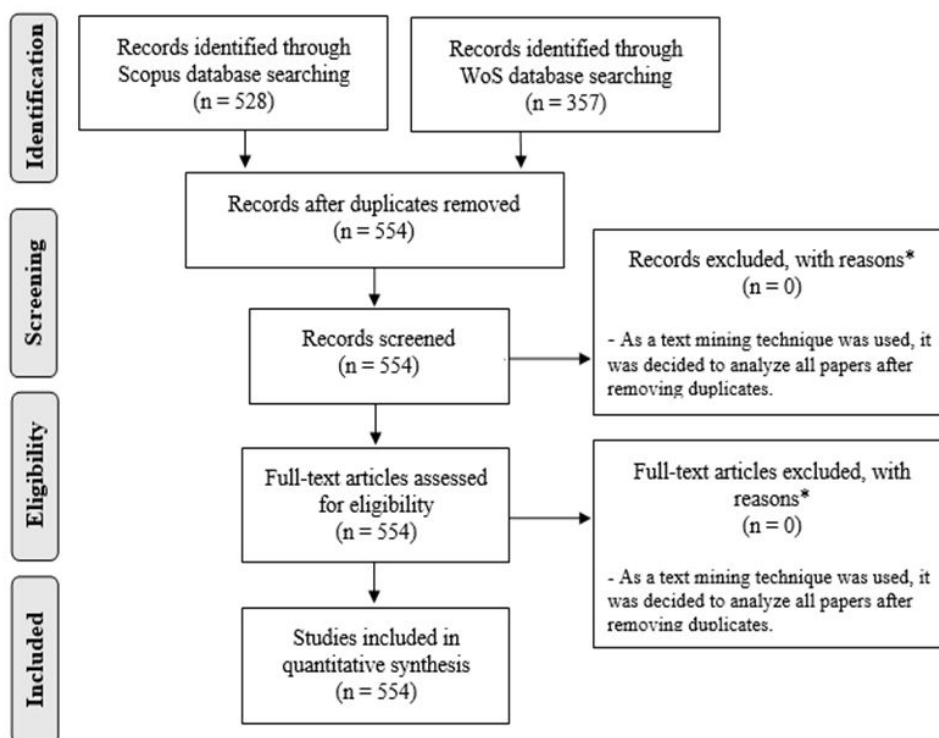
Following the merging of the Scopus ( $n = 528$ ) and Web of Science ( $n = 357$ ) metadata, the duplicates ( $n = 331$ ) were removed. This resulted in a population encompassing all publications in English between 1991 and 2019 ( $n = 554$ ) for the present study. The choice of the research scenario is justified as follows: the first scientific paper on this subject was published in 1991; and 2019 is the year of the most recent publication having complete information available. To filter bibliographic records, we searched for papers that included some of the descriptors presented in Table 2 in their titles, abstracts or keywords.

Table 2 – List of descriptors used in the present study.

Database	Descriptors
WoS	((TS=((("forecasting electricity" OR "predicting electricity") AND ("electricity spot" OR "electricity day-ahead" OR "electricity price" ) OR ("price forecasting" OR "price prediction" OR "forecasting price" OR "predicting price" OR "forecasting spikes" OR "forecasting VAR") AND ("electricity spot price" OR "electricity price" OR "electricity market" OR "day-ahead market" OR "power market"))))))).
Scopus	((TITLE-ABS-KEY(((("forecasting electricity" OR "predicting electricity") AND ("electricity spot" OR "electricity day-ahead" OR "electricity price" )) OR ("price forecasting" OR "price prediction" OR "forecasting price" OR "predicting price" OR "forecasting spikes" OR "forecasting VAR") AND ("electricity spot price" OR "electricity price" OR "electricity market" OR "day-ahead market" OR "power market"))))) AND ( LIMIT-TO ( DOC-TYPE,"ar")))

Source: Research results.

Figure 2 – Flowchart outlining the protocol adopted in this systematic review based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Four-Phase Flow Diagram.



Source: Research results.

### 3.2.2 Data treatment and analysis

After compiling the data, we created a list of publications according to the following attributes: authors and affiliations, paper titles, abstracts, keywords, and the complete references for the analyzed registers. We divided the investigation into sub-stages: first, a descriptive analysis of a bibliographic data frame.

In this sub-stage, we analyzed the annual publication of predictive methods used for electricity price forecasting in the most relevant sources (journals), as well as the most productive countries based on corresponding authors. We also described seminal papers according to the total number of citations, and the most relevant sources utilized in each of them.

Next, we presented the scientific publications on electricity price forecasting as network matrices. These networks displayed meaningful properties of the underlying research, and the influence of bibliometric units, such as scholars, and journals (Waltman & Van Eck, 2012; Aria & Cuccurullo, 2017).

### 3.2.3 Software and Hardware

The systematic review of the literature presented in the present study was developed using both the *R* (v.3.5.2) software and the *Bibliometrix R – package* proposed by (Aria & Cuccurullo, 2017), available at <http://www.bibliometrix.org>. This package utilizes a machine learning framework with data reduction techniques for dealing with substantial textual information, classified here as a classic “Bag of words” problem. To construct the temporal evolution of keywords, and network relationships, we used the free bibliometric software, *VOSviewer*, proposed by (N. Van Eck & Waltman, 2010), available at <http://www.vosviewer.com/>. Hardware specifications of the system used to perform the procedures are CPU Intel Core i5-7200U, 2.70 GHz, 16 GB RAM installed, and the Windows 10 operating system.

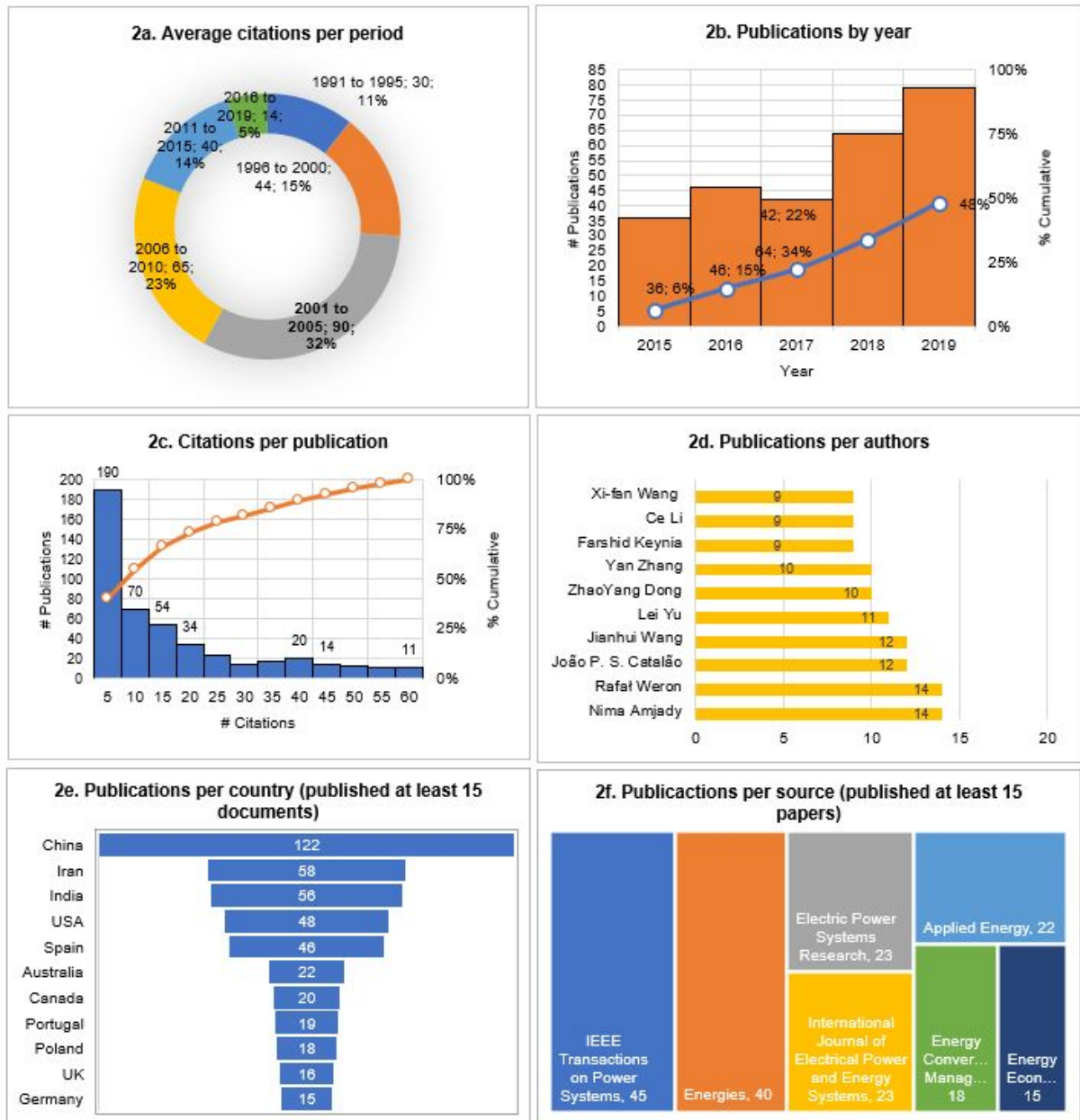
## 3.3 Results and Discussion

The papers used for the present study were published between 1991 and 2019. It is noteworthy that the 554 publications analyzed were written by 1115 different authors and published in 206 journals. The publications analyzed comprised 15099 bibliographic references. It was also observed that the 554 publications of the present study used 1416 distinct keywords.

Figure 3 (2a) shows that approximately 32% of the bibliographic citations were from 2001 to 2005, indicating that this was the period during which the main work in this area was carried out. Figure 3 (2b) shows an exponential growth in the number of papers published on *EPF*, indicating its academic relevance. Figure 3 (2c) shows a concentration of papers having a maximum of 5 citations. Figure 3 (2d) shows empirical evidence of Lotka’s Law, which describes

the frequency of publication by authors in any given field (Lotka, 1926). The importance of professors Nima Amjady and Rafal Weron is highlighted, each with 14 papers on EPF. With 122 publications (22%), China was the country with the greatest number of papers Figure 3 (2e). With 45 papers (8%), IEEE Transactions on Power Systems (impact factor equal to 6.62 in 2019) was the most productive source (Figure 3 (2f)).

Figure 3 – Dashboard on Electricity Price Forecasting research publications.



Source: Research results. Note: In the appendix section complete graphs exploring other information from the analyzed sample are presented (Figures 32,33,34,35).

The publication with the highest number of citations was Mohsenian-Rad & Leon-Garcia (2010). The most recent work was Gellert et al. (2019) (Table 3).

Following the descriptive analyses of these publications, the investigation of the keywords



Table 3 – Highlights of Electricity Price Forecasting publications.

	Year	Citations	Title	Contribution
Pioneers	(1991)	90	ESTIA: A real-time consumer control scheme for space conditioning usage under spot electricity pricing	Utilizes a decision modeling approach developed for prescribing consumer response to varying electricity price. The case of space conditioning usage is analyzed in detail and a real-time control scheme is proposed.
Most cited	(2010)	1288	Optimal residential load control with price prediction in real-time electricity pricing environments	Proposes an optimal and automatic residential energy consumption scheduling framework which attempts to achieve a trade-off between minimizing the electricity payment and minimizing the waiting time for the operation of each appliance in household, in the presence of a real-time pricing tariff combined with inclining block rates.
Most Recent	(2019)	11	A study on forecasting electricity production and consumption in smart cities and factories	A method for forecasting energy demand and production is proposed. Predictions contribute to balancing and smoothing the electricity intake from the power grid. Experimental evaluation is performed on data recorded in a real energy-management system.

Source: Research results.

used was undertaken. Table 4 shows the number of times that each of the 50 main keywords was used. As expected, the most used keywords were derivations of the expression electricity price forecasting. Regarding the predictive models used, the expression artificial neural networks (and its derivations) was present in at least 76 papers in the sample.

In addition, models based on the wavelet transform were cited in at least 25 different documents. This is supported by the fact that many energy trading markets operate on hourly frequency basis; therefore, the EPF forecasts have time series with many observations (Y. Zhang, Li, & Li, 2018; Chang, Zhang, & Chen, 2019). Thus, with a lower number of occurrences, there is a greater diversity of techniques used, with emphasis on the classic models of time series, such as those of the Arima class (Bandyopadhyay, Roy, & Ghosh, 2013).

Also, as Table 4 shows, several methodologies were used, such as those based on: support vector machine (Yuan, 2013; Ma, Zhong, Xie, Xia, & Kang, 2018; Zahid et al., 2019), probabilistic forecasting (Uniejewski, Marcjasz, & Weron, 2019), fuzzy logic (Pousinho, Mendes,

& Catalão, 2012), particle swarm optimization (Hannah Jessie Rani & Aruldoss Albert Victoire, 2019), lasso (Steinert & Ziel, 2019), and hybrid models (de Marcos, Bello, & Reneses, 2019), among others. Thus, it is highlighted that *EPF* is a research segment that uses different forecasting methods, and the development of research to investigate new models is relevant.

Table 4 – Most relevant words – Author’s keywords.

Rank	Terms	Freq.	Rank	Terms	Freq.
1	electricity price forecasting	112	26	market clearing price	9
2	price forecasting	86	27	short-term forecasting	9
3	electricity market	67	28	deregulation	8
4	forecasting	59	29	electricity price forecast	8
5	electricity price	33	30	prediction intervals	8
6	artificial neural networks	28	31	ann	7
7	neural networks	28	32	artificial intelligence	7
8	electricity markets	26	33	bidding strategy	7
9	neural network	26	34	electricity price prediction	7
10	wavelet transform	25	35	genetic algorithm	7
11	electricity prices	21	36	price spikes	7
12	artificial neural network	20	37	smart grid	7
13	price forecast	19	38	correlation analysis	6
14	time series analysis	15	39	demand response	6
15	day-ahead market	13	40	differential evolution	6
16	power market	13	41	electricity	6
17	support vector machine	13	42	forecast combination	6
18	arima	12	43	hybrid model	6
19	feature selection	12	44	lasso	6
20	probabilistic forecasting	12	45	load forecasting	6
21	particle swarm optimization	11	46	locational marginal price	6
22	data mining	10	47	particle swarm optimization (pso)	6
23	fuzzy logic	10	48	power markets	6
24	time series	10	49	prediction	6
25	electricity spot price	9	50	price prediction	6

Source: Research results. Note: Freq. it is equal to "frequency".

The results of Table 4 can be visualized in Figure 4 which shows the wordcloud for the main keywords used. Since the number of occurrences of keywords varies widely for the sample analyzed, the square root of the number of occurrences was taken to improve the visualization of Figure 4. Thus, the wordcloud considered the 100 words with the highest number of occurrences in the sample studies.

Figure 4 shows additional details of predictive models. The relevance of models based

Figure 4 – Wordcloud containing the top 100 keywords used in the sample.



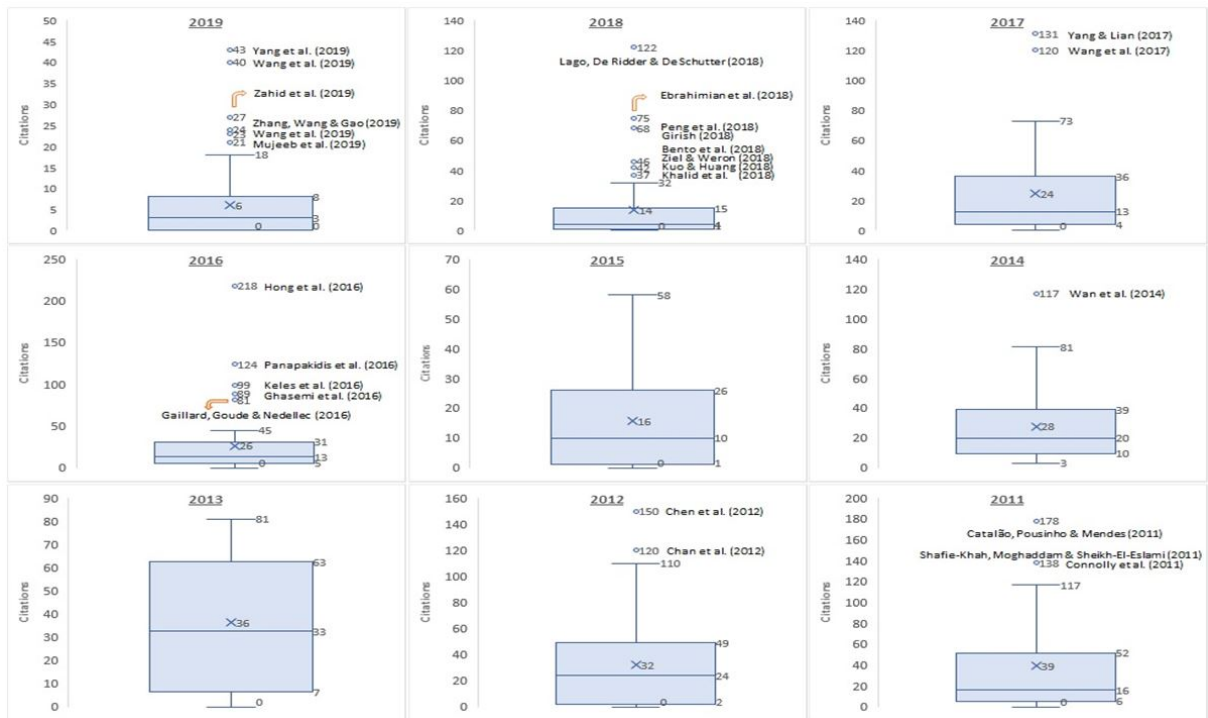
Source: Research results.

on spot prices, such as the day-ahead-market, is highlighted. These models use a wide range of techniques, among which the following stand out: ann (Windler, Busse, & Rieck, 2019), armax (J.-L. Zhang, Zhang, Li, Tan, & Ji, 2019), big data (W. Wang, Chen, Yan, & Geng, 2019), bootstrap (Tahmasebifar, Sheikh-El-Eslami, & Kheirollahi, 2017), calibration window (Hubicka, Marcjasz, & Weron, 2018), classification (Shrivastava, Panigrahi, & Lim, 2016), clustering analysis (Jin, Pok, Paik, & Ryu, 2015), correlation analysis (Peng, Liu, & Xiang, 2013), data mining (Ghayekhloo, Azimi, Ghofrani, Menhaj, & Shekari, 2019), garch (L. Zhang, Wu, Ma, & Wang, 2019), and genetic algorithm (Alamaniotis, Bargiotas, Bourbakis, & Tsoukalas, 2015), among others.

Figures 5 and 6 show the evolution of research published over the years, based on the number of citations for each paper. Although the sample analyzed dates from 1991, it is only after 2002 that it is possible to build a boxplot. Publications located beyond the upper limit of the interquartile distance are highlighted. Due to their academic impact, they can be considered references in *EPF*.

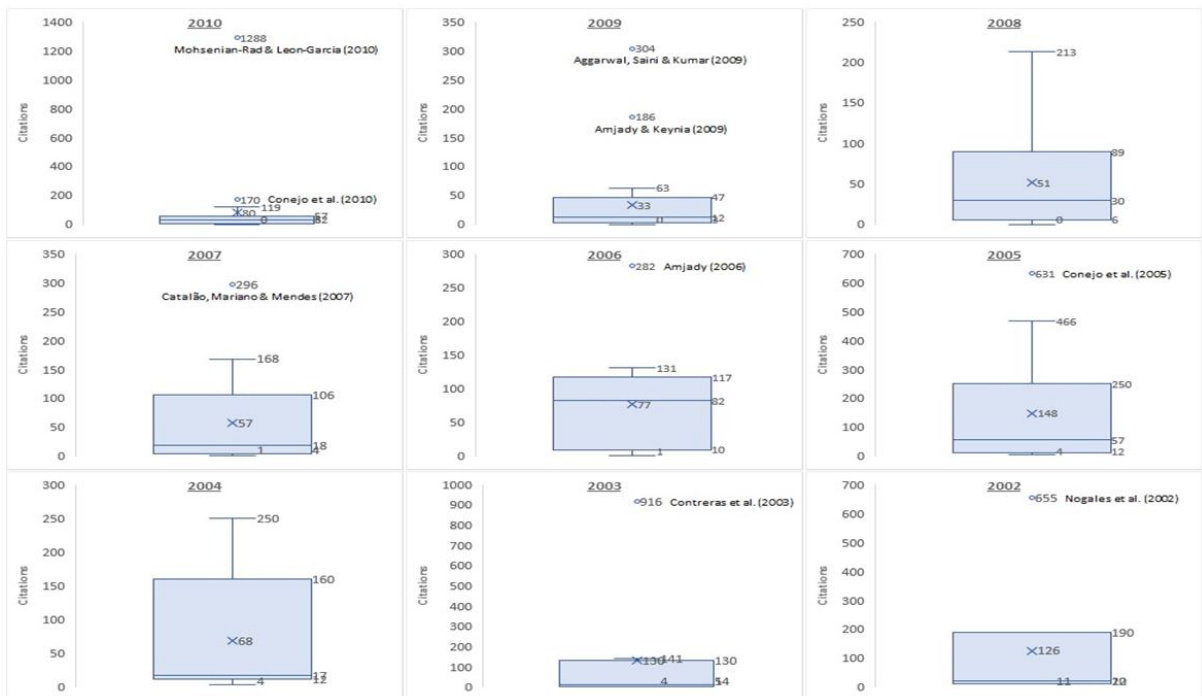
Equally as important as the descriptive analyses of publications on *EPF*, are the investigations by the existing *EPF* collaboration networks. Figure 7 shows the research relationships between the main research institutions that have published on the topic. This network was designed based on information from each of the published papers and the respective teaching and research institutions of the co-authors. The font size for each institution varies, depending on the number of papers published by the institution. The lines that connect the institutions are a

Figure 5 – Boxplot of the number of citations per paper over the years (part a).



Source: Research results.

Figure 6 – Boxplot of the number of citations per paper over the years (part b).

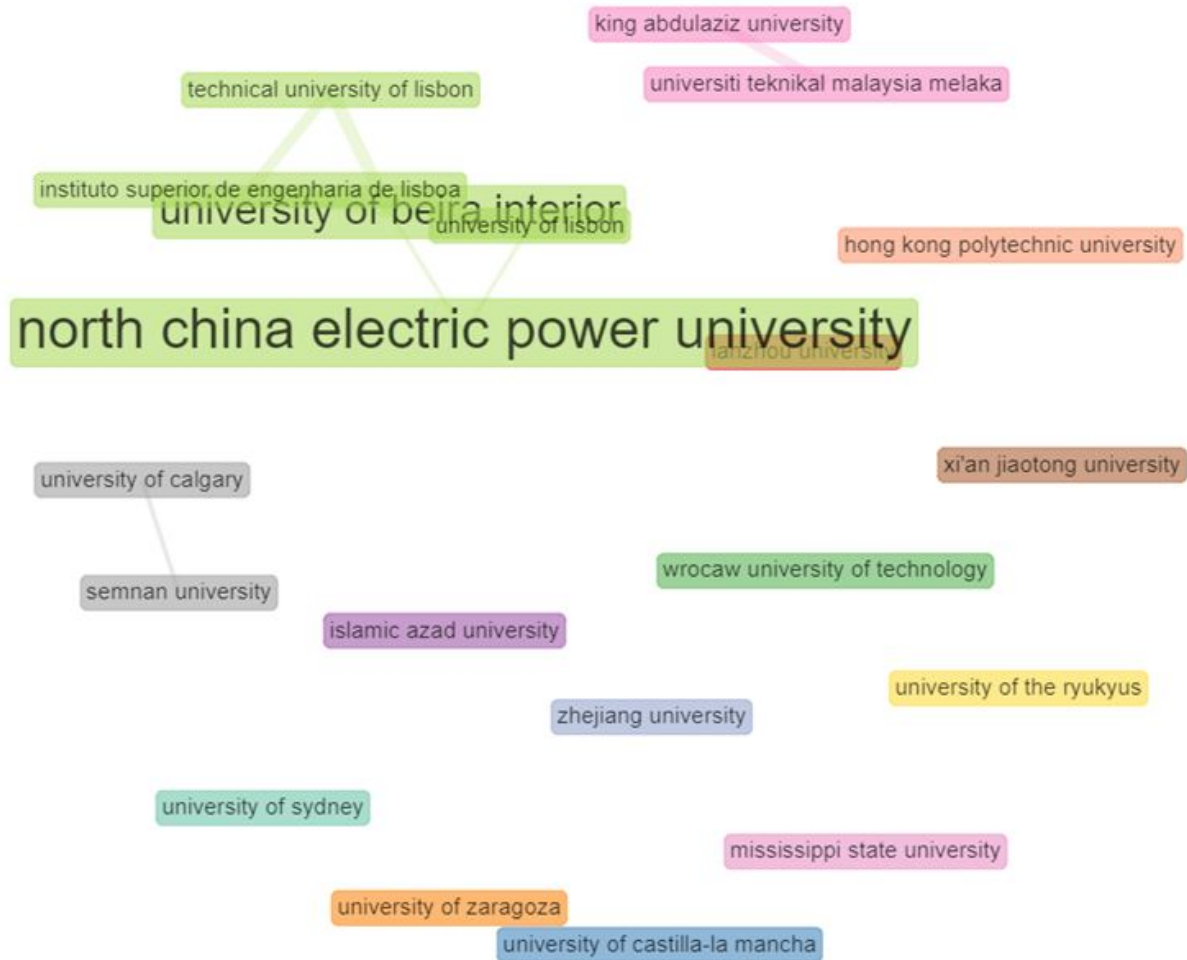


Source: Research results.

visual representation of the strength of the relationship between two institutions, where thicker lines denote stronger relationships. The network shows the top 20 organizations that published

on EPF.

Figure 7 – Collaboration networks between research institutions



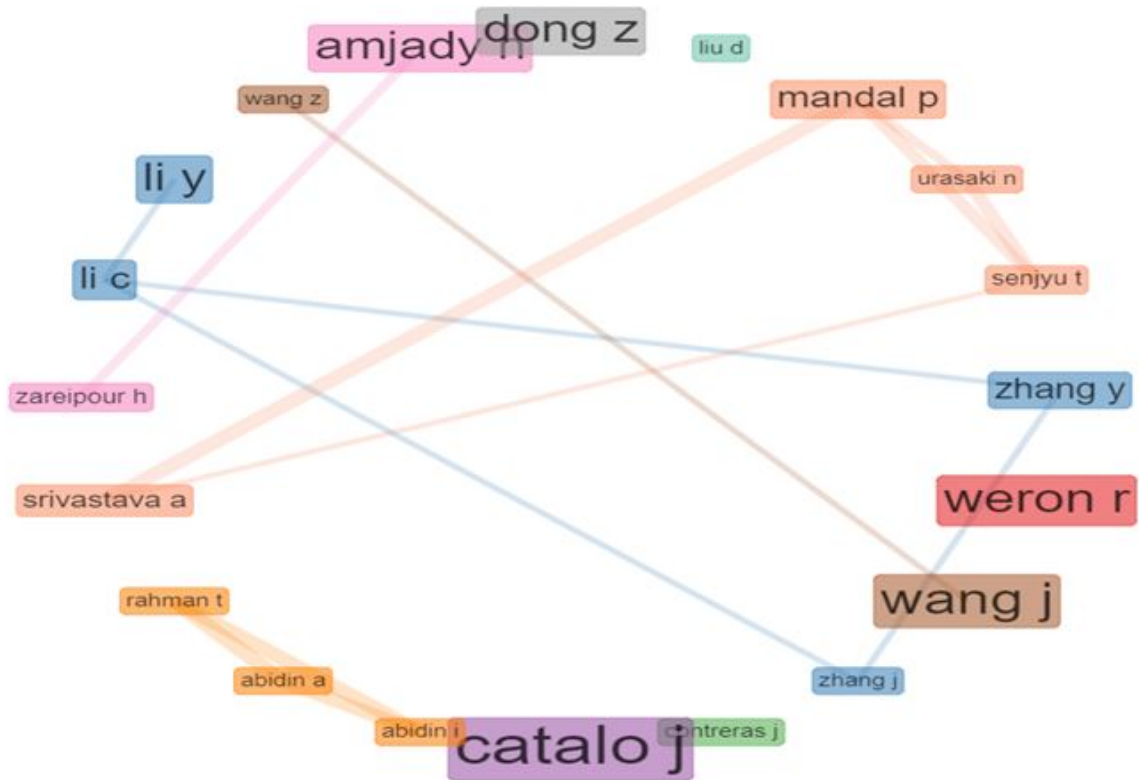
Source: Research results.

As shown in Figure 7, the institutions are separated by color, according to the proximity of their relationships. However, most of the papers on *EPF* are published in the single country publication format since, as this is a strategic and local issue, collaborations between institutions are still incipient. The main cluster, green, is formed by four Portuguese educational institutions (Technical University of Lisbon; Instituto Superior de Engenharia de Lisboa; University of Veira Interior and University of Lisbon) and one Chinese institution (North China Electric Power University). Complementary to Figure 7, Figure 8 presents a similar analysis. However, it focuses on the partnership networks between the main authors.

It is reinforced that the degree of collaboration between authors is limited. In most cases, it is restricted to collaboration among authors of the same nation.



Figure 8 – Collaboration networks between researchers



Source: Research results.

### 3.4 Conclusions

Forecasting electricity prices attracts the attention of different agents, since it is a central issue for good planning in the energy production chain. Due to its importance, the growth rate of publications on *EPF*, in the main journals around the world, is approximately 23% per year. The present study investigated 554 publications on this topic, published between 1991 and 2019, that were indexed simultaneously in the Scopus and Web of Science databases. The present paper used the *PRISMA* research protocol, which provides a high degree of reliability in the face of the analyses carried out. In addition, data science procedures, such as text mining, were used to describe the main attributes of the bibliographic information collected.

It is noteworthy that the present paper empirically verified the validity of some bibliometric laws, namely, Lotka's Law and Bradford's Law. That is, it provides evidence of a predominance of a few authors (Nima Amjady, Rafal Weron) who publish more, as well as a small set of journals (IEEE Transactions on Power Systems, Energies) that focus scientific production, on the topic. It was found that the years from 2001 to 2005 form the period during which the greatest volume of citations was concentrated.

The present paper also analyzed the main keywords and quantitative methods used in studies on *EPF*. It was found that, despite the great recurrence of studies based on artificial

neural networks, and wavelet transform, there is also a wide range of research being developed based on classic statistical approaches, regression models, hybrid methods, and recent machine learning procedures, for example. Finally, despite the increasing number of recent studies, the predominance of papers is still being produced in only a few countries. This inequality highlights both the importance of international cooperation to close the gap, and the need for more connected research clusters. Future development of predictive models depends heavily on the collection, and availability of reliable databases, which is a notable research obstacle in developing countries.

## Part II

### Empirical applications



# 4 Similarity Search in Electricity Prices: an Ultra-fast Method for Finding Analog

## Abstract

Accurately predicting electricity prices allows us to minimize risks and establish more reliable decision support mechanisms. In particular, the theory of analogs has gained increasing prominence in this area. The analog approach is constructed from the similarity measurement, using fast search methods in time series. The present paper introduces a rapid method for finding analogs. Specifically, we intend to: (i) simplify the leading algorithms for similarity searching, and (ii) present a case study with data from electricity prices in the Nordic market (there were not enough observations available in Brazil at the time of writing this chapter). To do so, Pearson's distance correlation coefficient was rewritten in simplified notation. This new metric was implemented in the main similarity search algorithms, namely: *BruteForce*, *JustInTime*, and Mueen's Algorithm for Similarity Search (*MASS*). Next, the results were compared to the Euclidean distance approach. Pearson's correlation, as an instrument for detecting similarity patterns in time series, has shown promising results. The present study provides innovation in that Pearson's distance correlation notation could reduce the computational time of similarity profiles by an average of 17.5%. It is noteworthy that computational time was reduced in both short and long time series. For future research, we suggest testing the impact of other distance measurements, e.g., Cosine correlation distance and Manhattan distances.

**Keywords:** Analog. Ensemble forecasting. Similarity search. Electricity prices.

## 4.1 Introduction

The construction of predictive models is gaining prominence in the literature (Geisser, 2017), since economic agents deal with uncertainty and aim to achieve the best results using available resources (Choi, 1993). Therefore, developing models with acceptable accuracy presents a meaningful challenge to researchers. George Box stated, "All models are wrong, but some are useful" (G. Box, 1976). In other words, prediction is a technique that deals with risk, and there will always be a fundamental error associated with it. The best model is the one that most adequately represents the phenomenon of interest.

In relation to the object of our study, electricity prices, there are several forecasting applications: (i) classical time series models like the autoregressive moving average, autoregressive integrated moving average, generalized autoregressive conditional heteroscedastic, among others (H. Liu & Shi, 2013); (ii) pre-processing techniques like spectrum analysis, wavelets and Fourier analysis (Miranian et al., 2013); and, (iii) machine learning approaches such as neural networks, fuzzy systems and support vector machine (Bui et al., 2016). Additionally, an alternative class known as hybrid models aims to combine machine learning representations with different methods. Instances of these methods are focused time-delay neural networks (Y. Chen et al., 2019), neural networks with fuzzy inputs (H. Liu et al., 2015), finite-impulse response neural networks (Pir et al., 2017), local feedback dynamic fuzzy neural networks (Nagaraja et al., 2016), type recurrent fuzzy networks (Jain et al., 2014), neuro-fuzzy inference systems (Moreno & Santos Coelho, 2018), among others.

The energy market is known for being an industry with high-frequency data (Madadi, Nazari-Heris, Mohammadi-Ivatloo, & Tohidi, 2018), for several reasons. First, sensor usage is widespread in energy (Jaradat, Jarrah, Bouselham, Jararweh, & Al-Ayyoub, 2015). Second, high-frequency data can better represent specific weather conditions, enabling the improvement of energy modeling (Aigner, Miksch, Müller, Schumann, & Tominski, 2007). Examples are diverse, such as: (i) solar radiation, which can be collected in minutes (Assuncao, Escobedo, & Oliveira, 2003); and, (ii) air humidity, atmospheric pressure, temperature and wind speed, which can also be measured in minutes (Longman et al., 2018).

In particular, the pricing of electricity also has significant volumes of information, in most cases, arranged on an hourly scale (Voronin & Partanen, 2014). Although the literature on this question is extensive, there is academic interest in the construction of nonparametric models applied to electricity prices, as they have presented promising predictive results. In general, these models are designed to deal with long-time series and are chiefly based on analog ensemble (AnEn) searches (D. Yang & Alessandrini, 2019; D. Yang, Kleissl, Gueymard, Pedro, & Coimbra, 2018) and scan-clustering methodologies (Costa, Ruiz-Cárdenas, Mineti, & Prates, 2021).

Due to both the complexity and the high volume of information, finding patterns in time series is a data science challenge. Given that, similarity analysis has been studied since the 1960s (Lorenz, 1969). In addition to the complexity of creating highly accurate models, significant volumes of information lead to developing algorithms with low computational time. As a result, the literature reflects efforts in mathematical and computational solutions to this problem (Mueen et al., 2017; J. Yang, Astitha, Delle Monache, & Alessandrini, 2018).

In general, similarity and analog studies are based on searches of similarity patterns between the latest available observations and the old observations through a scanning process on data (Gensler, Sick, & Pankraz, 2016). This methodology is widely used in climatology studies, where an AnEn is developed by first matching up the actual prediction from a numerical weather prediction (NWP) model with similar past projections (Eckel & Delle Monache, 2016).

As an example, some research in this area deserves special mention. Yang et al. (2018) presented a dual NWP model approach, by jointing the AnEn and the bias-corrected analog ensemble (BCAnEn) procedure and demonstrated that by combining different NWP models, it is possible to improve the storm wind speed prediction. Another critical study was carried out by Yang (D. Yang, 2019), which pointed out that using the kd-tree in AnEn, it could be possible to save computational time when necessary to test different model adjustments. Still, in this context, research on the forecast of solar irradiation is frequent, and (J. Yang et al., 2018) presented a substantive review of this area's main procedures.

Although relevant, previous work on the similarity search is mainly aimed at climatological research. This article innovates, as it addresses this methodology in the energy commercialization sector. Also, it is highlighted that previous analog forecasting studies are based on Euclidean distance as a metric of similarity (Mueen et al., 2017; D. Yang, 2019). McDermott & Wikle (2016) show that this procedure may present trouble. Since searches of analogs rely on embedding vectors being spatially similar over time, it is not certain that Euclidean distance ever leads to first-rate analogs, particularly for the spatiotemporal state processes. Pearson distance has mathematical similarities to the Euclidean approach (Immink & Weber, 2015), and could be a simplified way of rewriting its notation.

A research gap still needs to be addressed: finding alternative measures for the similarity pattern to reduce the computation time of analog searches. The research question is formulated: how it is possible to rewrite the classical analog ensemble models, based on the Euclidean distance profile, into simplified Pearson distance notation to obtain computational gains in the main analog algorithms? Therefore, our goal is to simplify the notation of the analog procedure to achieve the same distance profile with less computational time. The present paper contributes to the debate about electricity since it introduces a new predictive instrument based on the analog procedure, using the Nord Pool prices of electricity as a case study.

This paper is structured as follows: Section 1 outlines the objectives of this paper. Section 2 presents the materials and methods employed in preparing this paper. Section 3 presents the results obtained. Finally, section 4 discusses the implications of this research as well as possibilities for future research.

## 4.2 Materials and Methods

The algorithms used in this paper are: (i) *BruteForce*, (ii) *JustInTime*, and (iii) *MASS*. Usually, the Euclidean formula is presented in the literature on analogs to calculate the distance between length- $m$  query ( $X_i$ ) and each length- $m$  subsequence ( $Y_i$ ) in a given time series (Radack & Badler, 1989; D. Yang & Alessandrini, 2019; Zhu, Imamura, Nikovski, & Keogh, 2019). Generally, this approach calculates Euclidean distance  $d=(Y, X)$ , based on the normalized values of  $Y_i^*$  and  $X_i^*$ , as  $d = \sqrt{(Y_i^* - X_i^*)^2}$ . If we perform z-score normalization

on each object, the Euclidean Distance behaves similarly to the Pearson correlation coefficient (Höppner & Klawonn, 2009). Finally, the use of the Pearson correlation can produce simplified mathematical expressions, as shown in Equation (5.1).

#### 4.2.1 Similarity profile computation based on the Pearson correlation distance

The Pearson coefficient,  $\rho$ , measures the degree of correlation and the direction of this correlation, positive or negative, between two random variables. The Pearson correlation coefficient is defined as follows (Pearson, 1895):

$$\rho_{xy} = \frac{\sum_{i=1}^m (X_i - \mu_X) \cdot (Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^m (X_i - \mu_X)^2 \cdot \sum_{i=1}^m (Y_i - \mu_Y)^2}} \quad (4.1)$$

Equation (5.1) represents a single-pass algorithm for calculating the Pearson correlation. However, depending on the amount of data, it can demand considerable computational time. Using a little algebra, we can rearrange Equation (5.1) as follows, obtaining the Pearson product-moment correlation coefficient (Kelley, 1925):

$$\rho_{xy} = \frac{1}{m} \sum_{i=1}^m \left( \frac{X_i - \mu_X}{\sigma_X} \right) \left( \frac{Y_i - \mu_Y}{\sigma_Y} \right) \quad (4.2)$$

where  $\sigma_X = \sqrt{\sum_{i=1}^m (X_i - \mu_X)^2 / m}$  and  $\sigma_Y = \sqrt{\sum_{i=1}^m (Y_i - \mu_Y)^2 / m}$ .

Note that Equation (5.2) presents a simplified way for calculating the correlation between two sets, which reduces the computational time. It is noteworthy that this approach will be tested in the calculation algorithm called *BruteForce*. Finally, Equation (5.2) can be written in the abbreviated notation, illustrated below:

$$\rho_{x^*y^*} = \frac{1}{m} \sum_{i=1}^m (X_i^* \cdot Y_i^*) \quad (4.3)$$

where  $X_i^* = (X_i - \mu_X) / \sigma_X$  and  $Y_i^* = (Y_i - \mu_Y) / \sigma_Y$ .

The *BruteForce* algorithm was based on the Pearson formulation product-moment correlation coefficient. Figure 9 details the steps of this procedure, which consists of calculating the normalized values of the last observations (query) and the rest of the data.

An essential principle of a given similarity measure should be the invariance, under some specific conditions, e.g., data manipulation without changing the scale (Strehl, Ghosh, & Mooney, 2000). Thus, we highlight that the Pearson correlation coefficient is invariant to scaling (Orang & Shiri, 2012). This means that, when multiplying all elements by a non-zero constant,

Figure 9 – BruteForce algorithm

```

1: procedure BruteForce(data, query)
2:    $n \leftarrow \text{len}(\textit{data})$ 
3:    $m \leftarrow \text{len}(\textit{query})$ 
4:    $l \leftarrow n - m + 1$ 
5:    $CP[1:l] \leftarrow 0$ 
6:    $Q \leftarrow \text{zNorm}(\textit{query})$ 
7:   for  $i = 1:l$  do
8:      $CP[i] \leftarrow \text{sum}(\text{zNorm}(\textit{data}[i:i+m-1] * Q)) / m$ 
9:   end for
10:  return  $CP$ 
11: end procedure

```

Source: adapted by authors from: Yang & Alessandrini (2019).

the correlation remains the same. The same is valid when adding any constant to all the elements. This is a fundamental property, since the main goal of correlation is not to verify if two vectors are similar in absolute terms, but if they vary in the same direction:

$$\rho_{XY} = \rho_{X^*Y^*} = \rho(X^*Y) = \rho(XY^*) \quad (4.4)$$

According to D. Yang and Alessandrini (2019), a valid research strategy is to measure the degree of association between one normalized variable and one without normalization. This procedure assists in simplifying notations and will be utilized in the *JustInTime* algorithm. However, the main difference from the *BruteForce* procedure is that *JustInTime* only normalizes the latest information (query), leaving the rest of the series without any transformation (Figure 10).

Figure 10 – JustInTime algorithm

```

1: procedure JustInTime(data, query)
2:    $n \leftarrow \text{len}(\textit{data})$ 
3:    $m \leftarrow \text{len}(\textit{query})$ 
4:    $l \leftarrow n - m + 1$ 
5:    $CP[1:l] \leftarrow 0$ 
6:    $Q \leftarrow \text{zNorm}(\textit{query})$ 
7:    $\vec{\sigma} \leftarrow \text{mvstd}(\textit{data})$ 
8:   for  $i = 1:l$  do
9:      $CP[i] \leftarrow \text{sum}(\textit{data}[i:i+m-1] * Q) / (m * \vec{\sigma}[i])$ 
10:  end for
11:  return  $CP$ 
12: end procedure

```

Source: adapted by authors from: Yang & Alessandrini (2019).

Considering the correlation between  $X_i^*$  and  $Y_i$ , where  $X_i^* \approx N(0, 1)$ , Equation (5.2) can be rewritten as:

$$\rho_{X^*Y} = \frac{1}{m} \sum_{i=1}^m \left( \frac{X_i^* \cdot Y_i}{\sigma_Y} \right) - \frac{1}{m} \sum_{i=1}^m \left( \frac{X_i^* \cdot \mu_Y}{\sigma_Y} \right) \quad (4.5)$$

since  $\frac{1}{m} \sum_{i=1}^m \left( \frac{X_i^* \cdot \mu_Y}{\sigma_Y} \right) = \frac{m \cdot \bar{Y}}{m \cdot \sigma_Y} \sum_{i=1}^m \left( \frac{X_i^*}{m} \right) = \frac{m^2 \cdot \mu_Y \cdot \mu_{X_i^*}}{m \cdot \sigma_Y} = 0$ , then:

$$\rho_{X^*Y} = \frac{1}{m} \sum_{i=1}^m \left( \frac{X_i^* \cdot Y_i}{\sigma_Y} \right) \quad (4.6)$$

Additionally, adopting the correlation coefficient as a distance metric has other advantages. For example, it is possible to develop a regression model relating query  $X_i$  to the last observations  $Y_i$ . Assuming that the joint distribution of  $X_i$  and  $Y_i$  is the bivariate normal distribution, that  $\mu_Y$  and  $\sigma_Y^2$  are the mean and variance of  $Y$ , that  $\mu_X$  and  $\sigma_X^2$  are the mean and variance of  $X$ , and that  $\rho$  is the correlation coefficient between  $Y$  and  $X$  (Montgomery, Peck, & Vining, 2012). The conditional distribution of  $Y$  for a given value of  $X = x$  is:

$$f_{Y|x}(Y) = \frac{1}{\sqrt{2\pi\sigma_{Y|x}}} \exp \left[ -\frac{1}{2} \left( \frac{y - (\beta_0 + \beta_1 x)}{\sigma_{Y|x}} \right)^2 \right] \quad (4.7)$$

where

$$\beta_0 = \mu_Y - \mu_X \rho \frac{\sigma_Y}{\sigma_X} \quad (4.8)$$

and

$$\beta_1 = \frac{\sigma_Y}{\sigma_X} \rho \quad (4.9)$$

and the variance of the conditional distribution of  $Y$  given  $X = x$  is

$$\sigma_{Y|x}^2 = \sigma_Y^2 (1 - \rho^2) \quad (4.10)$$

For additional details on computational procedures, see the Attachment section.

#### 4.2.2 Similarity profile computation based on Euclidean distance

Equation (5.11) presents the mathematical formulation of the Euclidean distance between the elements of two vectors. Note that the formula below illustrates the case where the two vectors have previously been normalized. This is the formulation used in the *BruteForce* method.

$$d(X, Y) = \sqrt{\sum_{i=1}^m \left( \frac{X_i - \mu_X}{\sigma_X} - \frac{Y_i - \mu_Y}{\sigma_Y} \right)^2} = \sqrt{\sum_{i=1}^m (X_i^* - Y_i^*)^2} \quad (4.11)$$

where  $X_i^* = (X_i - \mu_X / \sigma_X)$  and  $Y_i^* = (Y_i - \mu_Y / \sigma_Y)$ .

The *JustInTime* method can be considered a rewrite of Equation (5.11), above. However, it takes one normalized variable and one without normalization. Equation (5.12) uses some algebra steps to demonstrate how to determine the adjusted equation for Euclidean distance.

Assuming the normalization of variable  $X$  (query), we can simplify Equation (5.12):

$$d(X, Y) = \sqrt{2 \left( m - \frac{\sum_{i=1}^m X_i Y_i}{\sigma_Y} \right)} \quad (4.13)$$

Equation (5.14) uses some algebra to demonstrate the existence of a relationship between Pearson's correlation coefficient and the Euclidean distance formula. Thus, since the correlation coefficient values vary between minus one and one, the smaller the distance between the vectors, the greater the force (correlation) between them:

$$\rho_{X*Y} = 1 - \frac{d(X, Y)^2}{2m} \quad (4.14)$$

The Euclidean distance incorporates the Pearson correlation function. Thus, the present paper will search for similarity patterns considering the Pearson coefficient, since this approach will return similar results but with lower computational costs. To illustrate this advantage, these results will be compared with those obtained using the Euclidean distance method.

According to [D. Yang and Alessandrini \(2019\)](#), there are alternative ways to calculate the correlation between a pair of vectors using of convolution procedure. Suppose  $A$  is the set of six data points,  $A = A_1, A_2, A_3, A_4, A_5, A_6$ , and  $F$  represents 4 forecasts to be matched,  $F = F_1, F_2, F_3, F_4$ . The full convolution between these two vectors is given by:

$$(A) \otimes (F) = \begin{bmatrix} A_1 F_1 \\ A_1 F_3 + A_2 F_4 \\ A_1 F_2 + A_2 F_3 + A_3 F_4 \\ A_1 F_1 + A_2 F_2 + A_3 F_3 + A_4 F_4 \\ A_2 F_1 + A_3 F_2 + A_4 F_3 + A_5 F_4 \\ A_3 F_1 + A_4 F_2 + A_5 F_3 + A_6 F_4 \\ A_4 F_1 + A_5 F_2 + A_6 F_3 \\ A_5 F_1 + A_6 F_2 \\ A_6 F_1 \\ 0 \\ 0 \end{bmatrix} \quad (4.15)$$

Thus, the convolution formulation for the correlation coefficient is presented by Equation (5.16).

$$\rho_{X*Y} = \frac{(X^*) \otimes (Y)}{m \cdot \sigma_Y} \quad (4.16)$$

Finally, we present the *MASS* algorithm proposed by Mueen et al. (2017) (Figure 11). This procedure uses the concept of ‘‘Convolution’’, i.e., a mathematical method between two sets that produces a third one, expressing how the shape of one is modified by the other. Convolution refers to both the resulting function and the computing of it (Burrus & Parks, 1985).

Figure 11 – Mueen’s algorithm for similarity search (modified)

```

1: procedure Mass(data, query)
2:   n ← len(data)
3:   m ← len(query)
4:   Q ← zNorm(query)
5:   σ ← mvstd(data)
6:   Q ← rev(Q)
7:   dots ← conv(data, Q)
8:   CP ← dots[m:n] / (m * σ)
9:   return CP
10: end procedure

```

Source: adapted by authors from: Mueen et al. (2017).

The next section presents the dataset used, refers to the Nordic electricity market in the short term (Nord Pool), and outlines the simulation procedures employed.

### 4.2.3 Dataset and simulation procedures

The data used in the present study were obtained from the Nord Pool, the leading power market in Europe (Janke et al., 2020). The dataset includes the hourly average electricity price for seven different countries, segregated into market areas (Pool, 2020).

The period of data analysis ranges from January 1<sup>st</sup>, 2014, 00 : 00, to September 2<sup>nd</sup>, 2019 00 : 00, totaling 49, 709 registers for each time series. There were six missing data points, from hour 02 : 00 to 03 : 00, at the end of March of each year. Missing data were computed using the average price of the preceding and subsequent hours. The time series utilized, including the number of time series per country and their acronyms, are presented in Table 8.

From each of the time series, 30 samples of size  $n$  equal to 720, 2, 400, 7, 200, 12, 000 and 24, 000 are randomly drawn. These values are associated with time series lengths of 30, 100, 300, 500 and 1000 days. The values of  $m$  adopted for this simulation were, 6, 9, 24 and 48 hours.

Figure 12 shows the flowchart with the detailed simulation process used to compare the similarity search algorithms based on Pearson’s correlation and those based on normalized distance. The validation of the analysis is obtained by comparing the computational times calculated for the different methods in carrying out the same task, building the similarity profile.



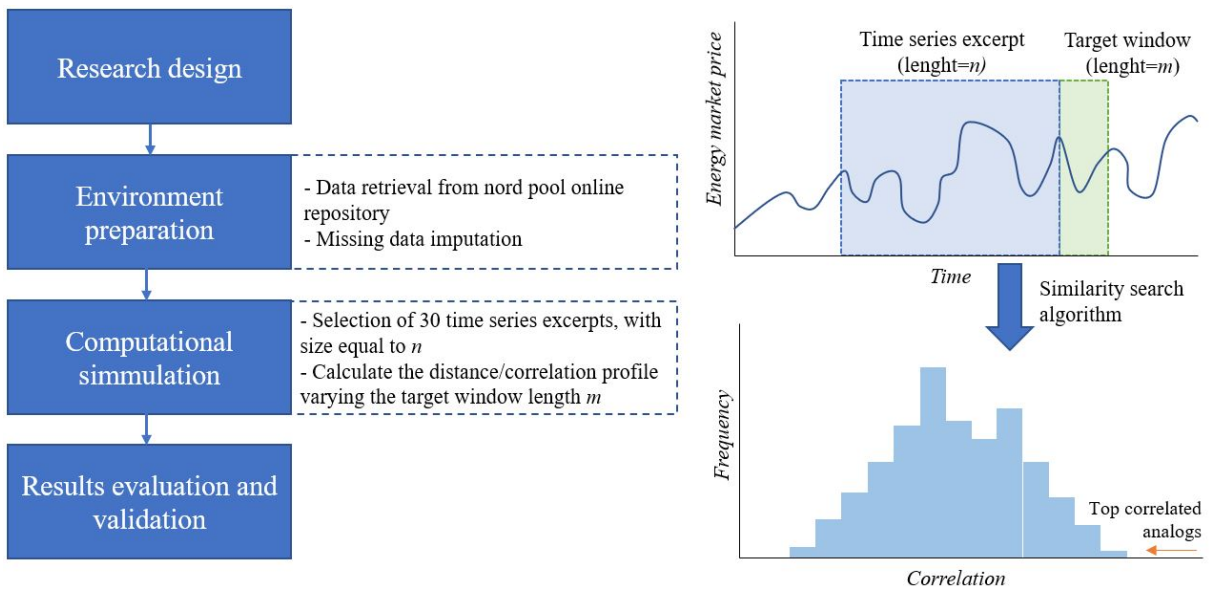
Table 5 – Nord pool energy submarkets analyzed

<i>Time series</i>	<i>Acronyms</i>	<i># Time Series</i>
System reference	SYS	1
Sweden	SE1, SE2, SE3, SE4	4
Finland	FI	1
Denmark	DK1, DK2	2
Norway	Oslo, Krsand, Bergen, Molde, Trhein, Tromso	6
Estonia	EE	1
Latvia	LV	1
Lithuania	LT	1

Source: Research results.

As the similarity profile is deterministic, the accuracy is the same as long as the model reaches its objective.

Figure 12 – Flowchart with the detailed process of simulation and analysis of similarity search algorithms



Source: Research results.

Routines were implemented using the *R* 3.6.0 programming language, adapting algorithms from (Mueen et al., 2017) and (D. Yang & Alessandrini, 2019). The *R* package *RollingWindow* was used to calculate the standard deviation of the data, considering fixed-width subsets of observations, called windows. This package is available from the *GitHub* repository at: <https://github.com/andrewuhl/RollingWindow>.

The computer used to execute the algorithms and to calculate the correlation and distance profiles had: *CPU* Intel Core *i5-4570* 3.20 GHz, 16 GB of *RAM* and operating system Windows 10x64. Computational time was calculated from the system's time delta before and after each execution of the methods.

### 4.3 Results and discussion

Each of the search algorithms (*BruteForce*, *JustInTime*, and *MASS*) were properly calculated using both Pearson correlation metrics and Euclidian distance formulation. To make the simulation more robust, results were obtained considering different samples, namely 2400 and 24000 observations. The search criteria used variable windows for data regarding the last observations (query) of sizes equal to 6, 9, 24 and 48 hours.

Table 9 presents the results of the 100 simulations performed, considering the sample formed from the last 2400 observations. The best methods, in terms of computational time, are highlighted in bold. The standard deviation of the simulations is shown in parentheses. Note that the Pearson's distance-based similarity search methods, especially *JustInTime* and *MASS*, respectively, had shorter computational times.

Table 6 – Sample data length of 2,400 (100 days) [hours]. The query length  $m$  varies from 3 to 48 [hours]. Each scenario is repeated 100 times, the mean computational times (in ms) are shown in the table.

m	Avg. time (ms)					
	Correlation similarity profile			Euclidean distance profile		
	<i>BruteForce</i>	<i>JustInTime</i>	<i>MASS</i>	<i>BruteForce</i>	<i>JustInTime</i>	<i>MASS</i>
m = 6	37.488 (7.582)	1.658 (4.232)	0.423 (2.138)	38.348 (7.587)	1.920 (4.487)	0.450 (2.230)
m = 9	38.401 (7.526)	1.704 (4.228)	0.445 (2.198)	38.218 (7.266)	2.022 (4.442)	0.533 (2.470)
m = 24	39.293 (7.647)	1.780 (4.138)	6.642 (6.740)	39.437 (7.635)	2.198 (4.603)	6.726 (6.704)
m = 48	39.869 (7.548)	2.136 (4.684)	6.765 (6.781)	40.220 (7.695)	2.387 (4.802)	7.141 (6.861)

Source: Research results.

There is a notable difference between the performance of the *BruteForce* algorithm and the others. Computational time is about 20 times greater than the *JustInTime* model, and 10 to 100 times greater than the *MASS* model. The computational time of *MASS* especially draws attention, as it reaches optimal values when  $m$  is equal to 6 and 9. Then, it presents computational times up to 4 times less than the *JustInTime* model.

However, when  $m$  is greater than 24, the computational time tends to be 3 times greater than the time calculated using *JustInTime*. This greater time variation found for the *MASS* model is intrinsic to the nature of the *MASS*, which computational time is mainly affected by the convolution procedure. In many cases, this is an advantage over the *JustInTime* algorithm. In other cases, however, the former is more efficient. It is up to the user of the algorithm to assess which model best suits their specific situation and data types.

Here, a similar analysis is presented (Table 10). However, the sample universe was substantially increased ( $n = 24,000$  hours). Again, the Pearson correlation-based models stood out concerning computational time, with the *JustInTime* algorithm as the most promising method for computing long time series (the most abundant sample universe).

By increasing the sample size of the available period by ten times, we obtained proportional increases in computational times for almost all models. The standard deviation increase, however, was limited to twice its original value. Thus, the computational advantage of models

Table 7 – Sample data length of 24,000 (1,000 days) [hours]. The query length  $m$  varies from 6 to 48 [hours]. Each scenario is repeated 100 times, the mean computational times (in ms) are shown in the table.

m	Correlation similarity profile			Euclidean distance profile		
	BruteForce	JustInTime	MASS	BruteForce	JustInTime	MASS
m =6	377.402 (12.924)	15.678 (4.842)	136.023 (8.606)	379.560 (14.351)	19.249 (6.436)	135.817 (8.815)
m =9	381.169 (15.298)	16.321 (5.009)	84.343 (8.274)	384.333 (19.447)	20.051 (6.644)	84.518 (7.871)
m =24	394.380 (14.623)	18.377 (5.699)	1087.231 (29.833)	396.693 (14.785)	22.121 (7.312)	1086.358 (28.683)
m =48	406.765 (14.676)	20.495 (6.714)	10.859 (6.644)	408.716 (14.300)	24.404 (7.587)	11.140 (6.748)

Source: Research results.

based on the correlation similarity profile becomes more evident.

The *JustInTime* algorithm, using the similarity profile, showed a 17.5% reduction in computational time. With the *BruteForce* and *MASS* algorithms, the gains were more discrete due to the greater variability of computational times. Again, the computational time of the *MASS* algorithm showed high sensitivity to the parameter  $m$ , assuming values 0.5 to 60 times the average time value of the *JustInTime* algorithm.

Finally, conclusions of the present paper are presented, emphasizing the time saving of the proposed formulation as well as suggesting potential studies to be developed in the future.

## 4.4 Conclusions

The electricity energy market is known for having high-frequency data. The examples are numerous, as the large-scale use of sensors across a wide range of processes provides a robust set of data. Thus, as the amount of information stored continuously increases over time, the search for statistical solutions that model this data is remarkable. Regarding predictive models, the range of approaches is broad. In particular, the literature has highlighted the relevance of predictive methods based on similarity or analogous searches. These methods scan a time series and, from the most recent observations, define moments where there is a high degree of affinity.

The main work on the methodology of analogs ensemble (AnEn) has made use of the Euclidean distance function. Our methodology revealed a high degree of similarity between the Euclidean formulation and Pearson's method. Thus, the present study is innovative in that, by rewriting Pearson's correlation equation, it was able to obtain the same results as the traditional approach but using less computational time. Therefore, the results of the present study are expected to provide a fast and robust tool for finding patterns in long time series, contributing to different actors in the energy planning sector.

The present study contributes to the energy planning processes of different agents, given that understanding price patterns has singular importance for minimizing risks and supporting reliable production planning. Good forecasts for future energy pricing can support operational arrangements, e.g., when the energy price is high, it may be more valuable for an industry to delay part of its production temporarily, trade the surplus electricity, and carry out preventive

maintenance on machines and accessories.

There are no disadvantages in applying Pearson's correlation in the search for analogs, as the correlation profile is a mathematical simplification of the normalized distance: the temporal analog with the shortest normalized distance is also the one with the most significant correlation with the search. The proposition is valid for the other windows: the analog with the second shortest distance has the second-largest correlation, and so on. The same is not observed; however, for the search algorithms: the *JustInTime* algorithm presented the lowest computational times in most excerpts of the series; however, the *MASS* algorithm obtained the best efficiency in others.

Future research should test the effect of different probability distributions on the data standardization process. A study of other measurement functions, such as distance from Manhattan, is recommended. Finally, yet no less importantly, we suggest the analysis of the impact of using different coefficient approaches such as entropy, Kendall, and Spearman.

# 5 Dynamic Time Scan Forecasting: A Benchmark With $M4$ Competition Data

## Abstract

Univariate forecasting methods are fundamental for many different application areas.  $M$  – competitions provide important benchmarks for scientists, researchers, statisticians, and engineers in the field, for evaluating and guiding the development of new forecasting techniques. In this paper, the Dynamic Time Scan Forecasting (DTSF), a new univariate forecasting method based on scan statistics, is presented. *DTSF* scans an entire time series, identifies past patterns which are similar to the last available observations and forecasts based on the median of the subsequent observations of the most similar windows in past. In order to evaluate the performance of this method, a comparison with other statistical forecasting methods, applied in the  $M4$  competition, is provided. In the hourly time domain, an average  $sMAPE$  of 12.9% was achieved using the method with the default parameters, while the baseline competition the simple average of the forecasts of Holt, Damped, and Theta methods was 22.1%. The method proved to be competitive in longer time series, with high repeatability.

**Keywords:** Univariate methods.  $M4$  competition. Benchmarking. Dynamic time scan forecasting.

## 5.1 Introduction

The development of predictive models is widely debated in the literature (Hill, Marquez, O'Connor, & Remus, 1994; Pai & Lin, 2005; Dudek, 2016; Shanmugam, 2006), since it assists the control of associated uncertainty intrinsic to random variables. Given the above, there are several categories of predictive models based on this physical knowledge (such as spectral analysis (Tchrakian, Basu, & O'Mahony, 2011)) of intensive machine learning and statistical approaches (Voyant et al., 2017). Forecasting models associated with a single random variable as a function of time support univariate forecasting, which is a very important area given its application in various sectors such as (Hassani & Silva, 2018; Cai, Chen, Hong, & Jiang, 2017; Bernardini & Cubadda, 2015), business (Khan Jaffur, Sookia, Nunkoo Gonpot, & Seetanah, 2017; Y. Zhang, Zhong, Geng, & Jiang, 2017; Tularam & Saeed, 2016), energy (Girish, Tiwari, et al., 2016; Rana, Koprinska, & Agelidis, 2016; Raviv, Bouwman, & Van Dijk, 2015), among others. In this context, it is fundamentally valuable to develop meticulous criteria for selecting the models (Billah, Hyndman, & Koehler, 2005).

The *M – competition* (Makridakis, Spiliotis, & Assimakopoulos, 2018; Makridakis & Hibon, 2000; Makridakis et al., 1993; Makridakis & Hibon, 1979) is the most important forecasting competition in academia, in which researchers from all around the world test their methods on real-life, anonymous time series from distinct areas of industry. The 4th edition took place in 2018 (Makridakis et al., 2018), and 17 methods based on combinations of statistical- and machine-learning or hybrids were tested on 100,000-time series. Outputs from these events are registered in review articles, pointing out the directions of development and refinement of the most promising forecasting techniques (Flores et al., 2019). The 5th edition took place in 2020, and focused on a retail sales application with 42,850 unit sales hierarchical series, with the objective to produce the most accurate point forecast as well as the most accurate estimation of the uncertainty of these forecasts (Makridakis, Spiliotis, & Assimakopoulos, 2021). The 6th competition focused on predicting the overall market returns of individual stocks (*The M6 financial forecasting competition*, n.d.).

Whereas most well-known forecasting methods are based on identifying intrinsic components of the time series, such as level, trend, or seasonality, a particular group of methods based on similarity searches have been arousing interest in the areas of meteorology and renewable energy (D. Yang & Alessandrini, 2019; Hoeltgebaum, Dias, & Costa, 2021). These methods consist of identifying past weather patterns ("analogs") that closely resemble the current state. These methods are capable of handling lengthy historical time series in order to produce accurate and interpretive forecasts.

Among these methods DTSE consists of a new and simple analog-based forecasting technique (Costa et al., 2021). It generates forecasts based on similar patterns, those with the highest  $R^2$  scores, calculated from the last available window.

The accuracy of analog-based methods is scarcely reported in areas other than energy prediction and is mostly limited to wind and solar energy forecasting applications (Gontijo, Costa, & de Santis, 2020, 2021), which begs the question: "are analog-search-based models competitive compared to classical statistical prediction methods?". Additionally, no research was found that compared analog search methods and statistical methods.

To fill this gap, the current paper describes the *DTSE* forecasting method and discloses its performance on the *M4* competition time series. We compare *DTSE* with eight classical statistical methods (Naive, Seasonal Naive, Simple Exponential Smoothing, Holt, Damped, Theta, AutoRegressive Integrated Moving Average (ARIMA), and Exponential Smoothing state space model (ETS)) and a combination of the outcomes of 3 individual methods (Holt, Damped, and Theta), which compose the baseline of the *M4* competition. The *M4* benchmark dataset was selected for this research because: (1) it consists of a reliable and curated benchmark base, adopted by other researchers and practitioners for developing and testing forecasting methods; (2) it has a significant number of series: 100,000 time series, with different frequencies (hourly, daily, monthly, weekly, quarterly, yearly); (3) it has been mostly predominated by statistical

methods of forecasting; (4) and it is composed of univariate and independent series.

The major contributions of the present paper can be summarized as follows:

- the study applies a new method to *M4* competition for benchmark purposes;
- the method is compared with nine classical statistical methods and a combination of the outcomes of three individual methods, which compose the baseline of the competition;
- in addition to applying the method, along with its default parameters, an exhaustive search with hold-out validation is adopted for model selection.

The major conclusions are:

- in the hourly time domain, an average error of 12.9% was obtained using the method with the default parameters, while the competition baseline was 22.1%;
- through the automatic selection of parameters, we boosted the accuracy of the method by 12.31% compared to the method application without parameters selection;
- the method proved to be competitive, both in terms of accuracy and computational cost, over long time series and with high repeatability.

The present paper is organized into 5 sections. Following this Introduction, Section 2 provides a review of the proposed forecasting method. Section 3 provides a background of the datasets and methods applied in this study. Section 4 presents the results and discussions obtained from the application of the methods. Finally, Section 5 concludes the present paper and includes some recommendations for future studies.

## 5.2 Materials and methods

### 5.2.1 *M4* competition dataset

The data used in the current study comes from the *M4* competition dataset ([Makridakis et al., 2018](#)). It is composed of 100,000 time series, taken from different domains such as Economics, Finance, Demographics, and Industry, among others. The time series show different periods: yearly, quarterly, monthly, weekly, daily, or hourly.

Table 8 summarizes the information about the competition's dataset. Domain refers to the time period from which the data have been extracted, ranging from hourly to yearly. The number of Series shows how many time series are available, in total. The dataset is mostly composed of a collection of time series from yearly, quarterly or monthly domains - 95,000 time series. The minimum length is the shorter time series in the given domain: the more aggregated the domain,

like yearly, the more difficult it is to retrieve data. For example, hourly time series are longer, having at least 700 available observation points. Horizon refers to how many steps are being predicted in the future and are being used for metric computation. Seasonality represents the expected recurrence of an event in a given time domain.

Table 8 – Summary of M4 competition dataset, including time-frequency, minimum length of time series, and forecast horizon of each time series.

Domain	Number of series	Min. length	Horizon	Seasonality
Yearly	23,000	13	6	1
Quarterly	24,000	16	8	4
Monthly	48,000	42	18	12
Weekly	359	80	13	52
Daily	4,227	93	14	7
Hourly	414	700	48	24

Source: Research results.

The dataset provides a public and reliable source for comparing statistical, machine learning, or hybrid methods on univariate time series forecasting (Bontempi, 2020). It is internationally recognized by researchers and data scientists as the most important competition in this area (Fildes & Makridakis, 1995).

## 5.2.2 Dynamic time scan forecasting

*DTSF* is a forecasting method based on scan statistics (Glaz & Balakrishnan, 2012) and was originally developed to address the problem of wind forecasting for Brazilian power generation plants. It consists of scanning a time series and identifying past patterns (called "analogs") similar to the last observations available of the time series (called "query") (Costa et al., 2021).

Let  $y_t$  be a time series of length  $N$ ,  $t = 1, \dots, N$ . Firstly, let vector  $\mathbf{y}^{[w]}$  be defined as the last  $w$  observations of the series:

$$\mathbf{y}^{[w]} = [y_{N-w+1}, \dots, y_N]. \quad (5.1)$$

The goal of *DTSF* is to identify analogs in the time series which are greatly correlated with vector  $\mathbf{y}^{[w]}$ . Hence, the set of candidate vectors can be defined by:

$$\mathbf{x}_t^{[w]} = [y_{t-w+1}, \dots, y_t] \quad (5.2)$$

where  $t = 1, \dots, N - 2 \cdot w$ . The upper limit of the time sequence ( $N - 2 \cdot w$ ) guarantees that vector  $\mathbf{x}_t^{[w]}$  does not overlap with vector  $\mathbf{y}^{[w]}$ . Fig. 13 presents the *DTSF* procedure. Given the last  $w$  observed values, which comprises vector  $\mathbf{y}^{[w]}$ , a rolling window with the same size ( $\mathbf{x}_t^{[w]}$ ) is used for scanning previous values of the series.

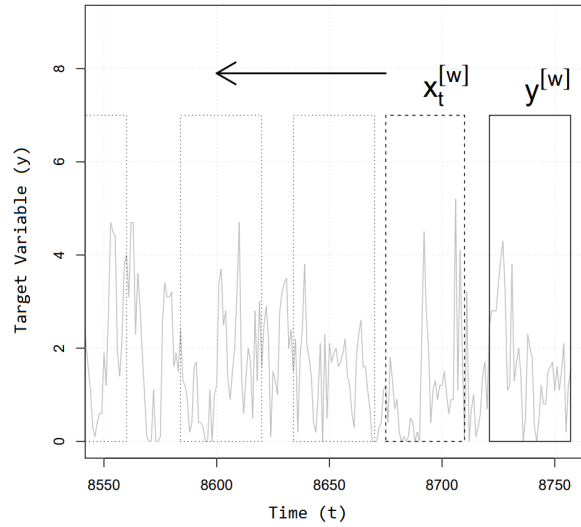


Lastly, *DTSF* provides a  $k - steps$  ahead forecast of the time series,  $y_{N+1}, \dots, y_{N+k}$ . To produce this outcome, the *DTSF* scans the series to find the closest analogs  $\mathbf{x}_t^{[w]}$ . The subsequent values of the time series are used as the forecast values:

$$y_{N+i} = f_{\mathbf{x}_t^{[w]}}(y_{t-w+i}) \quad (5.3)$$

where  $f_{\mathbf{x}_t^{[w]}}$  is a function which correlates the elements of vector  $\mathbf{x}_t^{[w]}$  and the elements of vector  $\mathbf{y}^{[w]}$ .

Figure 13 – Illustration of the DTSF time series scan procedure.



Source: (Costa et al., 2021), adapted by the authors.

According to that, a first constraint can be set on  $k$  :  $1 \leq k \leq w$ . This constraint guarantees that if the most correlated time series window comprises the most recent values, prior to vector  $\mathbf{y}^{[w]}$ , then the forecast values are a function of vector  $\mathbf{y}^{[w]}$ ,

$$y_{N+i} = f_{\mathbf{x}_{N-2w}^{[w]}}(y_{N-w+i}). \quad (5.4)$$

As stated in Equations before, forecast values depend on the window length  $w$  and the function  $f_{\mathbf{x}_t^{[w]}}(\cdot)$ . A intuitive proposal for function  $f_{\mathbf{x}_t^{[w]}}(\cdot)$  is a linear scaling of the elements of vector  $\mathbf{x}_t^{[w]}$ , i.e., a linear model. This occurs due to the fact that previous values are likely similar to the last observations, except for a scale and/or offset shift. So, the method searches for values that may be similar to the last values, after applying a similarity function (Costa et al., 2021).

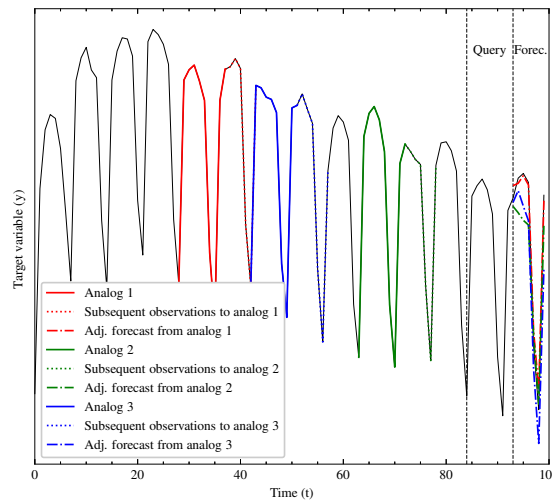
By taking a linear function as the similarity function, the parameters of the model can be estimated to minimize the sum of squares between the elements of vector  $\mathbf{y}^{[w]}$  and the linear equation:  $\beta_0^{[t]} + \beta_1^{[t]} \times \mathbf{x}_t^{[w]}$ . Moreover, the similarity statistic can be assumed as the linear regression coefficient of determination  $R^2$  (Costa et al., 2021; Montgomery et al., 2012):

$$R^2 = 1 - \frac{\sum_j \left( \mathbf{y}_j^{[w]} - \hat{\mathbf{y}}_j^{[w]} \right)^2}{\sum_j \left( \mathbf{y}_j^{[w]} - \bar{\mathbf{y}}_j^{[w]} \right)^2} \quad (5.5)$$

where  $\mathbf{y}_j^{[w]}$  is the  $j$ -th value of vector  $\mathbf{y}^{[w]}$  and  $\hat{\mathbf{y}}_j^{[w]}$  is the  $j$ -th predicted value using the estimated linear function. Finally, the method calculates a similarity profile based on the  $R^2$  score resulting from the comparison of the query with previous windows. The analogs with higher  $R^2$  scores are considered closer analogs. Predictions of future steps are calculated from a predefined number of analogs using aggregation functions, such as median (Costa et al., 2021).

Fig. 14 illustrates the forecasting procedure, using time scanning in a given hourly time series, adopting a window with a length equal to 48 hours, a linear similarity function (degree equal to 1), and the three analogs. Windows 1, 2, and 3 are the ones most similar to the last window of available data. The forecast is given by the median (but other statistics can be used such as the mean) of the subsequential observations of the analogs.

Figure 14 – Example of DTSF application to forecasting a time series. The three colored lines represent the top three analogs correlated to the queried period. The dashed lines are the subsequential observations of the analogs. The forecast is given by the median of the adjusted forecast from the subsequential observations of the top analogs.



Source: Research results.

The *DTSF* model requires three parameters to be selected by the user: the length of the query window, the similarity function specification, and the number of analogs to be considered. The original implementation of *DTSF* is available on the *R* package, *DTScanF*. In the present study, the original implementation is the extent to which the aggregation function applied to analogs can be either the median or the mean, according to the user or the model selection procedure.

As a data-driven method, *DTSF* usually performs better on time series with large numbers of observations and it can also be extended to search the patterns of secondary series related to the prediction. The main disadvantage of the method is the computational cost of scanning the

entire time series and calculating the similarity profile. However, more efficient methods, such as the Maureen's Algorithm of Similarity Search (MASS) which applies convolution, have been applied for speeding up this task (Gontijo, Costa, & de Santis, 2020). To keep it feasible, the linear similarity functions commonly adopted are from the first to the third-degree polynomials.

### 5.2.3 Statistical forecasting methods

A univariate forecasting method is a procedure for estimating a point. The forecast is based on past and present values of a given time series (Chatfield, 2000). This method is generally applied when there is a large number of series to forecast, or when multivariate methods require forecasts for each explanatory variable. Given the advantage of simplicity and high usage, univariate forecasting methods are employed in most of the forecast applications in areas such as business, energy, and finance. The following methods are selected from the latest M4 competition benchmark (Makridakis et al., 2018), and a simple explanation is given for each one, as follows:

1. *Naive*: the simplest, yet still powerful forecasting method; assumes that the next steps to be predicted are equal to the last available observation (Makridakis & Hibon, 1979).
2. *Seasonal Naive (sNaive)*: the same concept as Naive, with the adaptation that the time series is deseasonalized; method adjusted and forecast later, re-adjusted with the seasonal component (Makridakis & Hibon, 1979).
3. *Naive2*: each time series uses the forecast of either Naive or sNaive, based on their score on the validation set.
4. *Simple Exponential Smoothing (SES)*: classic statistical method which applies an exponentially weighted average (R. Hyndman, Koehler, Ord, & Snyder, 2008).
5. *Holt*: exponential smoothing with level and linear trend components (R. Hyndman et al., 2008).
6. *Damped*: exponential smoothing with dampened parameters for flattening trends, after a given period (Gardner & McKenzie, 2011).
7. *Theta*: method based on a coefficient of curvature of the time-series, applied to the second difference of the data (Assimakopoulos & Nikolopoulos, 2000).
8. *Combined (Comb)*: the simple average of the forecasts of the previous three models: Holt, Damped and Theta.
9. *ARIMA*: general forecast method estimated from the autoregressive, moving average and integration components from the time series analysis (G. E. Box & Pierce, 1970).
10. *ETS*: automatic forecasting based on an extended range of exponential smoothing methods (R. J. Hyndman, Koehler, Snyder, & Grose, 2002).

Table 9 – Parameters range adopted for *DTSF*.

Parameters	Range
Polynomial degree	1
Analogs	10
Window size	48
Aggregation function	Median

Source: Research results.

11. *DTSF*: the proposed method, adopting the defined default parameters, which are: (i) polynomial function degree equal to 1, (ii) analogs equal to 10, (iii) window size equal to length of forecast horizon, and (iv) median as aggregation function (Costa et al., 2021).

Table 9 presents the range adopted for the parameters of the proposed method. The polynomial degree is the degree of the function used for approximation, analogs are the number of analogs to be used to estimate the forecast, window size defines the length of the scan window, and aggregation function is the one that transforms the projection of the analogs into the final forecast.

#### 5.2.4 Model selection procedure

The split of the data into training sets and test sets split is predefined and given by the competition organizers. The data come from different files for each of the time series domains. The test set has a fixed horizon for all the time series, and it is used only for computing the final scores. The evaluation metrics adopted are the same ones that are applied in the *M4* Competition, and are those most used in literature (Al-Alawi & Islam, 1996; Azadeh, Ghaderi, & Sohrabkhani, 2008): the Symmetric Mean Absolute Percentage Error (sMAPE), Mean Absolute Scaled Error (MASE) and Overall Weighted Average (OWA). The formula for calculating the metrics is given:

$$sMAPE = \frac{1}{h} \sum_{t=1}^h \frac{2|Y_t - \hat{Y}_t|}{|Y_t| + |\hat{Y}_t|} \quad (5.6)$$

$$MASE = \frac{1}{h} \frac{(n - m) \sum_{t=1}^h |Y_t - \hat{Y}_t|}{\sum_{t=m+1}^n |Y_t - Y_{t-m}|} \quad (5.7)$$

$$OWA = \frac{sMAPE_k / sMAPE_{base} + MASE_k / MASE_{base}}{2} \quad (5.8)$$

where  $Y_t$  is the post sample value of the time series at point  $t$ ,  $\hat{Y}_t$  is the estimated forecast,  $h$  is the forecasting horizon,  $m$  is the frequency of the data,  $k$  is a given regressor, and *base* is the sNaive estimator.

A hold-out cross-validation scheme is adopted to evaluate and select the best parameters for the methods, in which the last  $k$  observations are kept as the validation set,  $k$  being equal to

the forecast horizon. All possible parameter combinations are enumerated within the defined ranges, and the methods are tuned using an exhaustive grid search procedure with  $sMAPE$  as the scorer.

### 5.2.5 Software and hardware

Routines were implemented using the R 3.6.0 programming language with the official benchmarks and evaluation script of M4 Competition, available at the *GitHub* repository (<https://github.com/M4Competition/M4-methods>). The *Forecast 8.7* package is used for the *SES*, *Holt*, *Damped*, *ARIMA*, and *ETS* methods. *DTSF* comes from the official implementation of the method in R and C++, available from the public repository (<https://rdrr.io/github/leandrominetti/DTScanF/>). All data and scripts are available from the authors upon request.

Computer specifications used to execute the algorithms and calculate the forecasts are as follows: CPU 8-core Intel Core *i9* 2.3 GHz, 16 GB of RAM, and *macOS* 12.5 operating system. Once the predictions are calculated, the error arrays are next calculated and saved as *RDS* files, allowing analysis of the results. Fitting time is computed from the time delta of the system, before and after each execution of the methods.

## 5.3 Results and discussion

Table 10 presents the average  $sMAPE$  achieved by each of the statistical methods and by the proposed method, computed for each of the time domains. The Theta method achieved the best scores for the yearly and monthly frequencies (14.603 and 13.003), which composed more than 70% of the total of the series, thus contributing to this particular method outperforming the other methods in the overall average (12.312). In the individual domains, Comb achieved the lowest error for both the daily (10.197) and the quarterly (10.197) domains, while the *ARIMA* method scored the lowest error on the weekly frequency (8.593).

The average error of all methods is the lowest for daily frequency (close to 3.00), and there seems to exist a trend toward increasing as the time domain becomes broader: the weekly average error is around 9, the monthly is around 13, and so on. The exception is for the hourly frequency, in which most of the statistical methods scored errors from 13.912 to 43.003.

*DTSF* exhibited fewer errors in comparison with those benchmark models (12.927). This makes the *DTSF* method interesting for studying applications in which competitive estimators are sought.

Table 11 presents the evaluation of the methods using *OWA*. This metric is understood as showing how one method is more accurate when compared to Naive2. If *OWA* is lower than 1 the method is more adequate than Naive2. Otherwise, Naive2 provides better forecasting

Table 10 – The performance of DTSF compared to M4 benchmark statistical methods – *sMAPE* metric.

<b>sMAPE</b>							
Method	Yearly (23k)	Quarterly (24k)	Monthly (48k)	Weekly (359)	Daily (4,227)	Hourly (414)	Average (100k)
Naive	16.342	11.610	15.255	9.161	3.405	43.003	14.207
sNaive	16.342	12.521	15.994	9.161	3.405	13.912	14.660
Naive2	16.342	11.012	14.429	9.161	3.405	18.383	13.565
SES	16.398	10.600	13.620	9.012	3.405	18.094	13.089
Holt	16.535	10.955	14.833	9.706	3.070	29.474	13.839
Damped	15.162	10.243	13.475	8.867	3.063	19.277	12.655
Theta	<b>14.603</b>	10.312	<b>13.003</b>	9.094	3.053	18.138	<b>12.312</b>
Comb	14.874	<b>10.197</b>	13.436	8.947	<b>2.985</b>	22.114	12.567
ARIMA	15.150	10.408	13.486	<b>8.593</b>	3.185	14.081	12.679
ETS	15.356	10.291	13.525	8.727	3.046	17.307	12.725
DTSF	16.816	11.006	13.823	8.983	3.313	<b>12.927</b>	13.370

Source: Research results.

Table 11 – The performance of DTSF compared to M4 benchmark statistical methods – *OWA* metric.

<b>OWA</b>							
Method	Yearly (23k)	Quarterly (24k)	Monthly (48k)	Weekly (359)	Daily (4,227)	Hourly (414)	Average (100k)
Naive	1.000	1.066	1.095	1.000	1.000	3.593	1.072
sNaive	1.000	1.153	1.147	1.000	1.000	0.628	1.106
Naive2	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SES	1.003	0.970	0.951	0.975	1.000	0.990	0.970
Holt	0.956	0.935	0.989	0.964	0.997	2.760	0.976
Damped	0.888	0.893	0.924	0.916	0.996	1.140	0.912
Theta	0.872	0.917	0.907	0.971	0.999	1.006	<b>0.906</b>
Comb	<b>0.868</b>	0.891	0.920	0.926	<b>0.979</b>	1.559	<b>0.906</b>
ARIMA	0.891	0.898	<b>0.904</b>	0.927	1.041	0.950	<b>0.906</b>
ETS	0.903	<b>0.890</b>	0.914	0.931	0.996	1.824	0.913
DTSF	1.002	0.961	0.950	<b>0.914</b>	1.092	<b>0.552</b>	0.969

Source: Research results.

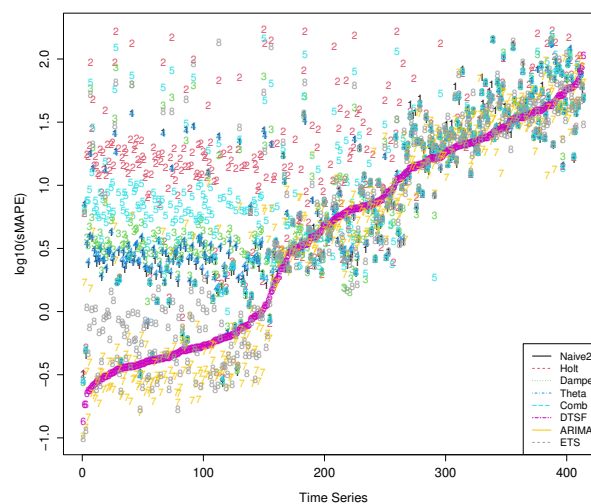
performance. The *DTSF* scores for the hourly series imply a meaningful increase in accuracy over the Naive method (0.552). Moreover, when applying fine-tuning, the gain increases to nearly 50%. For all other domains, the only ones in which the method performed worse than Naive2 were the yearly and the daily, both of which have in common the longer term forecast period and the lowest seasonality traits in common.

The outcome of the experiment can be explained by the intrinsic design of the *DTSF* method, which was originally conceived to deal with very long time series with recurrent patterns, such as its original application to 30 – *min* frequency wind speed forecasting. Comparing results

to Table 8, which presents the seasonality, length, and forecast horizon of each time domain, it is shown that the *DTSF* accuracy is greater when the number of available data points is also greater.

Fig. 15 displays the average *sMAPE* for each one of the 414 hourly time series available in the competition database, listed in ascending order according to the calculated error of the *DTSF* method. The methods Naive, sNaive and SES methods were holdouts of the graphical representation. The y-axis is presented using the base-10 logarithmic scale in order to facilitate visual analysis.

Figure 15 – Forecasting methods average *sMAPE* for each of the 414 hourly time series, ordered by the accuracy of the *DTSF* method. The proposed method obtained fewer errors for most of the time series in this particular domain of application.



Source: Research results.

In the first 170 time series with the lowest *sMAPE* – one-third of the total available – the method proposed in the present article achieved errors close to  $10^{-2}$ , while most of the others obtained errors between  $10^{0.5}$  and  $10^2$ . This shows the enormous predictive power in this specific type of series, and the great gain in accuracy that explains the best performance of this method, on average. Analyzing the sets between the 170<sup>th</sup> and 300<sup>th</sup> time series with the smallest error, there is less distinction between all the methods which, in general, presented errors very close to each other. Other methods have shown a lower errors than *DTSF* along all time series, specially the methods *ARIMA* and *ETS*. In the set between 300<sup>th</sup> and 414<sup>th</sup>, *DTSF* again marginally outperformed the other benchmark methods in most of the series.

Table 12 presents the average *sMAPE* detailed by the forecast horizon, grouped by 6-hour periods. *DTSF* obtained lower errors, for all horizons than the other compared methods. Furthermore, the average error is 12.9%, and the highest errors were obtained during the periods between the hours from 19 to 30.

To provide better visualization of error evolution over time, Fig. 16 presents the mean

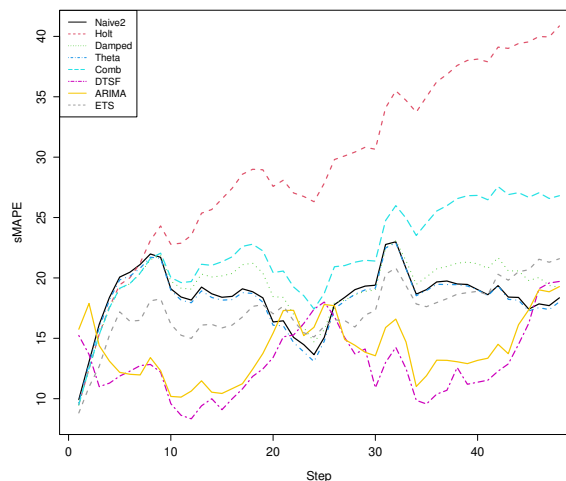
Table 12 – Average  $sMAPE$  obtained in the 414 hourly time series  
by the predicted steps, grouped in 6-hour periods.

Methods	Steps								
	1-6	7-12	13-18	19-24	25-30	33-36	37-42	43-48	1-48
Naive2	16.3	20.1	18.8	15.7	18.2	20.7	19.3	18.0	18.4
Naive2	16.3	20.1	18.8	15.7	18.2	20.7	19.3	18.0	18.1
Holt	15.7	23.0	27.1	27.5	29.9	34.9	37.9	39.8	29.5
Damped	15.5	20.3	20.5	17.5	18.1	21.2	21.2	19.9	19.3
Theta	16.1	19.9	18.5	15.3	17.8	20.5	19.2	17.8	18.1
Comb	15.6	20.6	21.8	19.7	20.8	24.9	26.7	26.8	22.1
ARIMA	14.2	11.4	11.2	15.8	15.4	13.9	13.4	17.0	14.1
ETS	13.6	16.5	16.4	16.6	16.5	19.0	18.9	17.4	17.3
<b>DTSF</b>	<b>12.6</b>	<b>10.7</b>	<b>10.2</b>	<b>15.0</b>	<b>14.8</b>	<b>11.6</b>	<b>11.6</b>	<b>11.6</b>	<b>12.9</b>

Source: Research results.

errors per step of each method (excluding the three from the previous figure), for all hourly time series. An increase in error over time, according to the phenomenon of error propagation, is expected. This is better observed in the Holt method, in which error varied from 10% at the first step to 40% at the last step. Moreover, in such a visual representation, the Theta model is perceived to have been more accurate, on average, than the *DTSF* model for the 1st and 24th hours.

Figure 16 – Average  $sMAPE$  (obtained in the 414 hourly time series  
by all the methods for each step of the prediction, up to 48 hours – forecast horizon).



Source: Research results.

Most statistical methods presented a pattern of very similar curves, with the exception of the *DTSF* method. In *DTSF*, the errors presented a different pattern, alternating peaks, and valleys with the patterns of the other statistical methods. In general, *DTSF* appeared to remain more stable throughout the period, experiencing less of the error propagation effect and not



exceeding the limit of 20%. These are more examples that explain the better performance of the *DTSF* method, compared to the benchmark, in the hourly domain.

Table 13 shows the time necessary to fit the methods for all of the 100,000 time series. The methods *sNaïve* and Comb have been omitted as these two are a combination/selection of individual methods. Total fitting time is given in seconds, while the average time per series is given in microseconds. The Ratio Naive column compares the average time of a particular method compared to the execution time of the *Naïve* method.

Table 13 – Total and average times necessary for fitting the methods.

Methods	Total fit- ting time (s)	Average time per series (ms)	Ratio to naive
Naïve	0.458	1.106	1.00
sNaïve	0.656	1.584	1.43
SES	2.219	5.360	4.85
Holt	5.947	14.365	12.99
Damped	12.789	30.892	27.94
Theta	2.964	7.159	6.47
ARIMA	18437.598	44535.261	40278.22
ETS	1838.638	4441.155	4016.63
DTSF	6.241	15.074	13.63

Source: Research results.

*DTSF* was the method that consumed the most computational time, almost 9 times more than Naive. It is worth mentioning that the default parameters for *DTSF* adopt 10 analogs to estimate the forecast. Also, part of the method is executed in the C compiled language, and part of it is executed in *R*.

## 5.4 Conclusions

The current paper presents the results of applying the dynamic time scan forecasting method with the *M4 – competition* data and compares it with statistical methods used as baselines in the same competition. The results point to a significant gain in accuracy in hourly time domain problems, compared to the reference, which justifies adopting this method for problems of this particular nature.

Since the method was developed for problems with long time series and high repeatability, *DTSF* has been proved competitive. In the present experiment, the *DTSF* method reduced the *sMAPE* by 12.13%.

Furthermore, the dissemination of this method may be interesting for other researchers who wish to extend it to existing methods, either by combining it with other techniques or by adapting its operation to other applications.

Future research should extend the method to multivariate forecasting problems and hierarchical time series and should assess its performance in other applications with this characteristic (the M5 competition, for instance). Also, some extensions of the method itself are foreseen, in order to improve its accuracy on time series for which its performance was less satisfactory than the performance of other statistical methods, for example, adopting k-fold instead of hold-out cross-validation for model selection ([Bergmeir, Hyndman, & Koo, 2018](#)).

# 6 Applications of dynamic time scan forecasting on electricity spot market: a case study based on the Brazilian Difference Settlement Price

## Abstract

Developing predictive models is a complex task since it deals with the uncertainty and the stochastic behavior of variables. Specifically concerning commodities, accurately predicting their future prices allows us to minimize risks and establish more reliable decision support mechanisms. Although the discussion on this question is extensive, there is academic attention being paid to the construction of nonparametric models applied to energy markets, as they have presented promising predictive results, that justifies the present study. Given the above, the following question is formulated: what is the accuracy of Dynamic Time Scan Forecasting (DTSF) regarding energy prices in the Brazilian spot market? This paper applies *DTSF* to the short-term electricity market prices, in Brazil, from 2006 to 2019. *DTSF* consists of scanning a time series and then identifying past patterns (so-called "matches"), similar to the last available observations. We predict Brazilian electricity spot prices, according the most similar matches, using aggregation functions, such as median. Recent research on the electricity spot market is increasing, indicating research significance. Our predictive approach exhibited greater accuracy than seminal statistical models. Our approach was designed for a high frequency series. Its predictive performance remained robust when other models presented both high predictive errors (spring), as well as when those models are highly accurate (winter). For future research, we recommend a more finely-tune study on *DTSF* parameters.

**Keywords:** Electricity Price; Spot Market; Scan Methods; Dynamic Time Scan.

## 6.1 Introduction

The construction of predictive models arouses interest in the literature ([Hamm & Borison, 2006](#); [Kuhn et al., 2008](#); [Bui et al., 2016](#); [Geisser, 2017](#)), since economic agents deal with uncertainty in multiple spheres and aim to achieve the best level of results from the available resources ([Choi, 1993](#)). Therefore, developing models with acceptable accuracy and adequate rigor presents a meaningful challenge to researchers. The Professor George Box synthesized this

scheme: "All models are wrong, but some are useful" (G. Box, 1976). In other words, prediction is a technique that deals with risk, and there will always be a fundamental error associated with it. The best model is the one that adequately represents the phenomenon of interest.

In relation to the object of our study, electricity prices, there are several forecasting applications: (i) classical time series models like the autoregressive moving average, autoregressive integrated moving average, generalized autoregressive conditional heteroscedastic, among others (Pappas et al., 2008; H. Liu & Shi, 2013); (ii) pre-processing techniques, e.g., spectrum analysis, wavelets and Fourier analysis (Simonsen, 2003; Miranian et al., 2013); and, (iii) machine learning approaches such as neural networks, fuzzy systems and support vector machine (X. Chen et al., 2012; Bui et al., 2016). Additionally, an alternative class known as hybrid models aims to combine machine learning representations with deferent methods. Instances of these methods are focused time-delay neural networks (Y. Chen et al., 2019), neural networks with fuzzy inputs (H. Liu et al., 2015), finite-impulse response neural networks (Pir et al., 2017), local feedback dynamic fuzzy neural networks (Nagaraja et al., 2016), type recurrent fuzzy networks (Jain et al., 2014), neuro-fuzzy inference systems (Moreno & Santos Coelho, 2018), among others.

Although the literature on this question is extensive, there is academic interest in the construction of nonparametric models applied to energy markets, as they have presented promising predictive results. In general, these models are designed to deal with long-time series and are chiefly based on analog search (D. Yang & Alessandrini, 2019; D. Yang, Wu, & Kleissl, 2019) and scan-clustering methodologies (Simmhan & Noor, 2013; Costa et al., 2021).

According to this framework, understanding price behavior takes singular importance (Rostamnia & Rashid, 2019) since it allows the minimization of risk and uncertainty (Khosravi, Nahavandi, Creighton, & Naghavizadeh, 2012; Heck, Smith, & Hittinger, 2016) and thereby to provide reliable production plans (Milligan et al., 2016; Schuh, Prote, Sauermann, & Franzkoch, 2019) and to establish a fine electricity market design (Woo & Zarnikau, 2019). In this scenario, it is noteworthy that in Brazil, for instance, the National System Operator defines the spot price of electricity weekly (Resende, Soares, & Ferreira, 2018). Predicting future values can help to establish enterprise decisions and behaviors. For example, if a firm expects that the energy price will scarcely increase, it could provisionally suspend part of its production and sell the surplus electricity in the spot market later at a higher price (Ioakimidis, Oliveira, & Genikomsakis, 2014; Tian, Xiao, Wang, & Ding, 2015). Also, the value of mapping the main conceptual theories used in the literature regarding the electricity spot market is noteworthy, as those theories evolve and become more robust over time (Weron, 2007, 2014).

The present paper contributes to the debate about the electricity spot market in two ways: (i) it illustrates how big data tools can produce relevant information about the electricity market; and, (ii) it illustrates, as a case study, a new forecasting approach to the analyzed market. Accordingly, we aim to review the literature on the electricity spot market, through the employment of big data tools, presenting the main research trends in the electricity spot market;

also, to introduce a new, predictive instrument based on dynamic time scan, using the Brazilian difference settlement price as a case study.

The paper is organized as follows. Section 2 presents the methodology, the data retrieval, and the dynamic time scan forecasting procedure. Section 3 illustrates an application of our forecasting approach, applied to the Brazilian electricity market, as a case study. Finally, section 4 highlights some patterns observed in the literature review and the potential use of dynamic time scan forecasting for future studies focusing on the energy spot market.

## 6.2 Materials and Methods

This section first presents the data and the procedures used to carry out a review of literature. Second, it introduces a promising statistical model based on dynamic time scan forecasting. To illustrate the accuracy of the model, and to present its applicability, the Brazilian "difference settlement price" dataset was used as a case study.

## 6.3 A review of the literature on the Electricity Spot Market

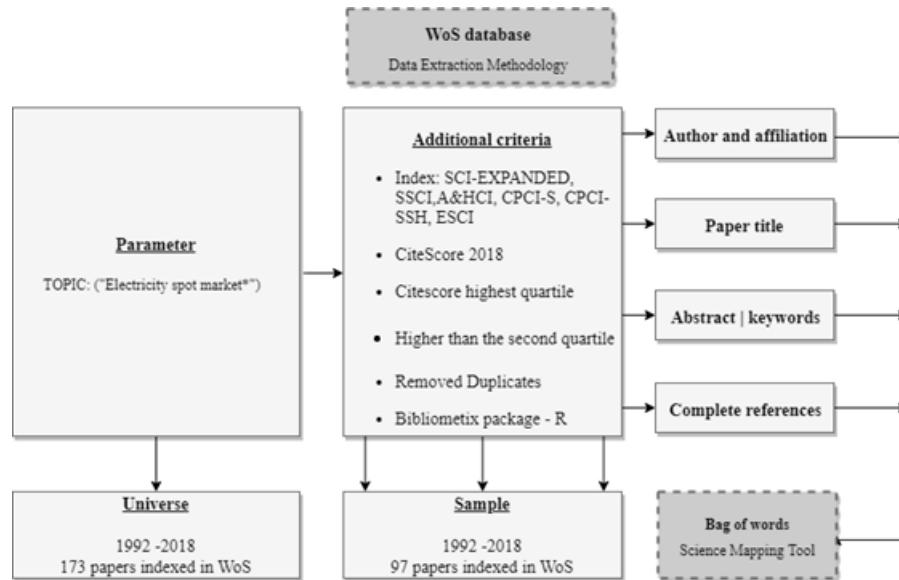
We conducted an exhaustive survey of publications regarding papers on the electricity spot market. It is noteworthy that Web of Science (WoS), along with Scopus, are the most commonly used academic citation databases for delineating fields of study. WoS, formerly known as the ISI, Web of Knowledge database, was used as the most complete and traditional bank of scientific publications in the world (Strozzi, Colicchia, Creazza, & Noè, 2017; Analytics, 2017). We limited our search to papers published in English. Moreover, we considered only the WoS database, since the WoS and Scopus databases may not differ significantly (Weron, 2014). In addition, we chose to analyze only the core publication journals, which are usually indexed in both databases.

Data extraction from the WoS platform (2019 – 05 – 06) took into account publications that responded to the search expression "electricity spot market\*", as shown in Figure 17. The use of the "\*" character in the search expression helps to capture words in both the singular and in the plural, making search results more complete.

The research universe encompasses publications from the years 1992 to 2018 ( $n = 173$ ). 1992 is the year of the first scientific publication on the subject indexed in WoS. 2018 is the last year for which complete information is available. The sample for this project (1992to2018,  $n = 97$ ) is henceforth called "primary publications". To filter these primary publications, we scanned for any of the descriptors shown in Figure 17 in the titles, abstracts, author's keywords, or in the Plus keywords.

Criteria for including a paper in the sample were based on these indicators: (i) CiteScore (measures average citations received per document published in the serial); (ii) SCImago Journal

Figure 17 – Methodological procedure to obtain the analyzed sample.



Source: Research results. Note: *CiteScore*, *SNIP* and *SJR* metrics calculated using data from 30 April 2018.

Rank (*SJR*) (measures weighted citations received by the serial. Citation weighting depends on subject field and prestige, *SJR* of the citing serial); and, (iii) Source Normalized Impact per Paper (*SNIP*) (measures actual citations received relative to citations expected for the serial’s subject field), all of them for the year of 2018. According to that the present study considered only (i) American Economic Review; (ii) Applied Energy; (iii) Econometrica; (iv) Economic Journal; (v) Energy; (vi) Energy Conversion and Management; (vii) Energy Economics, (viii) Energy Policy; (ix) Environmental Science & Technology; (x) European Journal of Operational Research; (xi) IEEE Transactions on Power Systems; (xii) IEEE transactions on Smart Grid; (xiii) International Journal of Electrical Power and Energy Systems; (xiv) International Journal of Forecasting; (xv) Journal of Banking and Finance; Journal of Economic Perspectives; (xvi) Journal of Political Economy; (xvii) Journal of the European Economic Association; (xviii) Mathematical Programming; (xix) Operations Research; (xx) Proceedings of the IEEE; (xi) Production and Operations Management; (xii) Renewable Energy; (xiii) Review of Financial Studies and (xiv) Solar Energy.

After selecting journals, we conducted a descriptive analysis of our bibliographic data frame. To do this, we used the *Bibliometrix R* package developed by (Aria & Cuccurullo, 2017) to analyze the annual publication of electricity spot markets in the most relevant journals, as well as the most productive countries of corresponding authors and seminal works. Next, we developed a dictionary that aggregates similar words. For example, we considered “electricity prices” and “electricity pricing”, or “electricity spot market” and “electricity trading”, as equivalent expressions. Finally, we constructed a WordCloud using the VOSviewer software (N. Van Eck & Waltman, 2010), that filters the hot and cold areas of interest for the theme according to the co-citation keywords, showing in detail how they interact with each other.

### 6.3.1 Data retrieval

The present study is based on the Brazilian difference settlement price (PLD). The *PLD* is determined weekly, considering three load levels for each sub-market. The submarkets are defined by the National Operator of the System, and consider the following geographical divisions: North, Northeast, Southeast/Center-West and South.

The *PLD* is determined ex-ante (considering expected availability and load information), based on weeks counted from Saturday to Friday. The prices must sell out all the energy, not just the contracted energy, among the agents (Chamber of Electric Energy Commercialization, 2019). The analyzed sample consists of weekly *PLD* data ( $R\$/MWh$ ) collected by the Chamber of Electric Energy Commercialization, from January 2006 to May 2019 ( $n = 701weeks$ ). We utilized the twelve available series, divided into the four Brazilian sub-markets: North, Northeast, Southeast/Midwest and South (called, respectively, N, NE, SE and S); and three load levels of energy: Heavy, Average, and Light (called, respectively, P, M and L). According to this, we have:  $L_N; L_{NE}; L_S; L_{SE}; M_N; M_{NE}; M_S; M_{SE}; P_N; P_{NE}; P_S$  and  $P_{SE}$ .

### 6.3.2 Dynamic time scan forecasting methodology and benchmark comparison

Dynamic time scan forecasting (DTSF) is an *R* – package based on scan statistics (J. Chen & Glaz, 2012). It was originally formulated to deal with wind forecasting and power generation by industrial plants (Costa et al., 2021). It consists of scanning a time series and then identifying past patterns (so called "matches") similar to the last available observations. Future values are predicted from the most similar matches using aggregation functions, such as median. Mathematical formulation and the results of original applications can be found in (Costa et al., 2021).

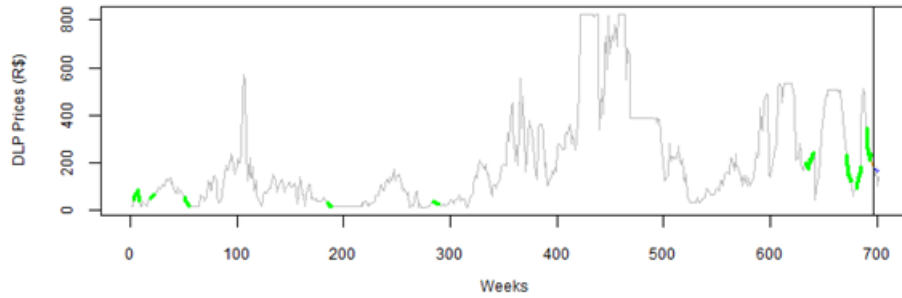
Here, we present an innovative use of *DTSF* based on the electricity spot market. To present the method, we utilize a predictive window equal to four weeks, the number of matches as the ten best-correlated ones and, finally, a polynomial function of order one (Figure 18). In this way, it is possible to find patterns between last available data and the old information, without any statistical test (green highlights). This procedure is based only on high similarity statistics ( $R^2$ ). The most similar patterns, called “median value of matches,” allow us to predict the *PLD* with the highest level of accuracy.

To illustrate the *DTSF*, we compare it to the classical statistical methods utilized in the “Makridakis Competitions” (also known as the M Competitions or M-Competitions). These competitions are a series of open disputes organized by the Professor Spyros Makridakis to evaluate and compare the accuracy of different forecasting methods (Makridakis & Hibon, 2000; R. J. Hyndman, 2020). Table 14 provides our benchmark comparison.

The test set has a fixed horizon for all of the time series, and it is used only to compute



Figure 18 – Time series of the PLD ( $P_{SE}$ ) and ten best matches found using *DTSF*.



Source: Research results.

Table 14 – *DTSF*, benchmarks, and standards for comparison of the *M4* Competition.

Model	Initials	Description
Statistical benchmarks	Naïve	A random walk model; future values will be the same as that of the last known observation
	Naïve2	Forecasts are equal to the last known observation of the same period.
	Mean	The forecasts of all future values are equal to the average (or “mean”) of the historical data.
	Arima	An automatic selection of possible ARIMA models is performed and the best one is chosen using appropriate election criteria.
	ETS	Automatically provides the best exponential smoothing model, indicated through information criteria.

Source: (Makridakis, Spiliotis, & Assimakopoulos, 2020), adapted by the authors.

the final scores. We utilized *M4* competition accuracy metrics, based on the Overall Weighted Average (OWA) of two accuracy measures: The Mean Absolute Scaled Error (MASE) and the Symmetric Mean Absolute Percentage Error (sMAPE), since they are among those most found in the literature (Al-Alawi & Islam, 1996; Azadeh et al., 2008).

## 6.4 Results and Discussion

### 6.4.1 Academic research on electricity spot market

Research on electricity spot markets began with (R. J. Green & Newbery, 1992), who analyzed the competition in the British market and demonstrated the Nash equilibrium in supply schedules, implying a high markup on marginal costs and substantial deadweight losses. However, there was a notable gap of seven years until the next three papers were published (De Vany & Walls, 1999; Wolfram, 1999; R. Green, 1999) The first research (De Vany & Walls, 1999) studied transmission efficiency in the Western US. The second and the third studies (Wolfram, 1999; R. Green, 1999) focused on measuring duopoly power in the British electricity spot market and



the reform of electricity trading in England and Wales, respectively.

In the period between 1992 and 2018, 215 authors published in leading selected journals ( $n = 25$ ) indexed in WoS, totaling ( $n = 97$ ) publications. Here, we have additional information for our sample: (i) Author's Keywords (included in records of papers by the authors) ( $n = 299$ ); (ii) Keywords Plus (index terms automatically generated by WoS, considering the titles of cited papers) ( $n = 253$ ); (iii) all keywords ( $n = 432$ ) (the sum of Author's Keywords and Keywords Plus, excluding duplicates); and, (iv) Average citations per documents ( $n = 45.54$ ). During the survey period, the mean annual growth of publications was 8.33% (from 1 paper in 1992 to 8 papers in 2018). Additionally, an average of  $4.62 \pm 2.78$  papers was published per year, demonstrating academic relevance.

Seminal works on the energy spot market are concentrated, according to Bradford's Law (Bradford, 1934), in a few journals, namely: Energy Economics ( $n = 20$ ), Energy Policy ( $n = 19$ ) and IEEE Transactions on Power Systems ( $n = 14$ ). This growth has aroused the interest of several researchers and favors the emergence of some questions, namely: (i) which theoretical approaches have shown the best results in describing electricity spot market? (ii) what are the potential gaps in the literature that future works should address?

The leading countries in publications are the USA ( $n = 31$ ), followed by Germany ( $n = 23$ ) and Spain ( $n = 21$ ). Together, they account for 41.44% of the world's research on the electricity spot markets. On the other hand, the countries with the highest total number of citations are, respectively, the USA ( $n = 1127$ ), Spain (1042) and the United Kingdom ( $n = 993$ ). Last, but not least, Ireland ( $n = 101$ ), the United Kingdom ( $n = 90.27$ ) and Spain ( $n = 86.83$ ) are the three countries with the highest average number of citations per paper, which is a good indicator of the degree of academic relevance.

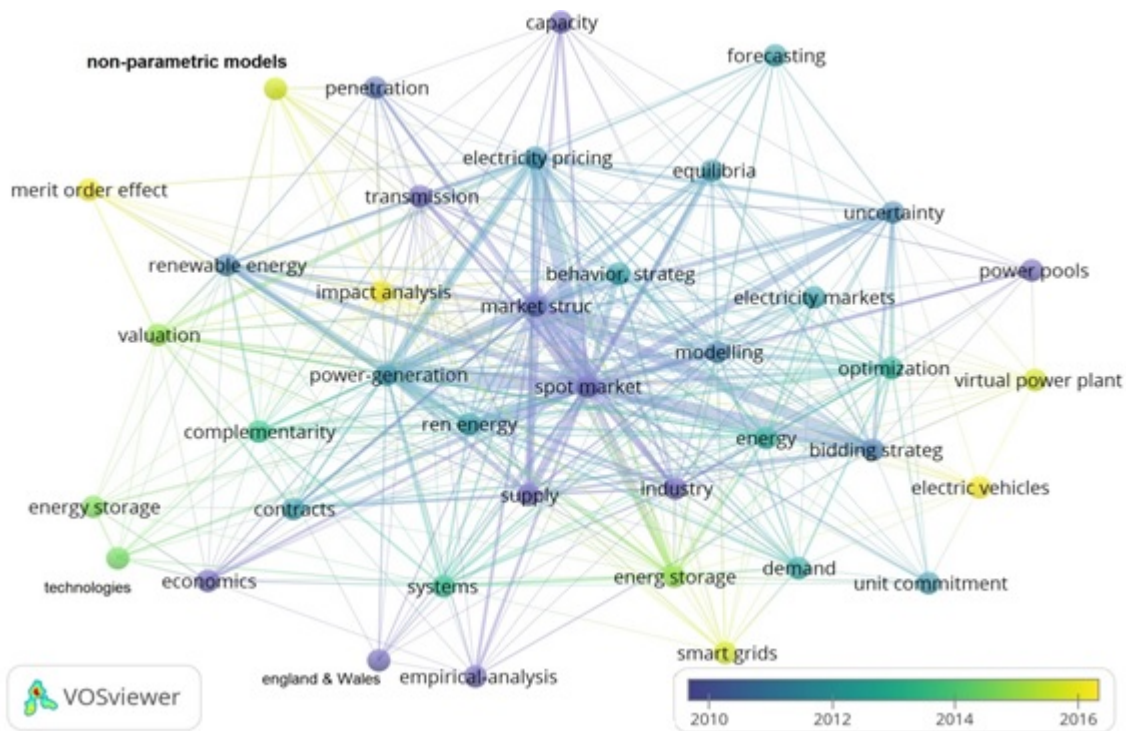
The principal references, cited in the 97 publications, are, respectively: (i) (R. J. Green & Newbery, 1992). Richard Green is Professor of Sustainable Energy Business at the Imperial College London; David Newbery is a professor at the University of Cambridge; (ii) (Klemperer, 2002). Paul Klemperer is an economist and Professor of Economics at Oxford University; and, (iii) (Arroyo & Conejo, 2000). José M. Arroyo is Professor at the Universidad de Castilla-La Mancha; Antonio J. Conejo is Professor at Ohio State University.

It is possible to create maps based on bibliographic information and then explore, for instance, the degree of appearance of a particular keyword and its temporal evolution. This type of analysis allows us to verify the hot and cold areas of interest in the scientific debate on the electricity spot market (Figure 19). Our purpose is to take out the frequency of papers with certain keywords. We set our "breakpoint" ( $n = 2$ ) as the minimum number of occurrences considering all the keywords ( $n = 432$ ). As it is possible to see, below, 37 keywords (nodes) meet that threshold. The diameter of the circle is proportional to the number of occurrences of, and the lines show the number of links between, these keywords. In recent research, some topics have gained prominence: (i) non-parametric models; (ii) impact analysis; (iii) merit order effect;

(iv) virtual power plant; and, (v) electric vehicles.

Figure 19 highlights a defined trajectory of research over time. The first published research (in purple) emphasized economic aspects such as the supply of energy, as well as technical elements (capacity and transmission, e.g.) concerning spot energy markets. Following this, there was a growing importance of studies on electricity generation and pricing (in blue), studies on contract theory, energy storage, and technologies (in green). Finally, the areas of the most current interest are those linked to non-parametric models, impact analysis, merit order effect, virtual power plant, and electric vehicles.

Figure 19 – Temporal WordCloud for the 50 most repeated keywords (all) in the analyzed sample.



Source: Research results.

Let us look at the non-parametric node. This node has been used in papers after 2016, which indicates a possible trend for further research. Studies that have used the keyword "non-parametric models" have also used other keywords, namely: (i) behavior strategies, (ii) complementarity, (iii) electricity pricing, (iv) impact analysis, (v) market structure, (vi) penetration, (vii) power generation, (viii) renewable energy, (ix) spot market, and (x) transmission. Thus, our paper relates to current research trends on electricity and innovates by proposing the first application of the nonparametric model (DTSF) for forecasting energy prices in the spot market.

The main articles that addressed nonparametric models applied to energy focused on: (i) assessing the influence of high penetration of wind power on the market-splitting behavior between West and East Denmark, using logit and non-parametric models (Figueiredo, da Silva, & Cerqueira, 2016); and, (ii) expressing the probability response for market-splitting of day-

ahead spot electricity prices as a function of the explanatory variables representing the main technologies in the generation mix, including wind, hydro, thermal and nuclear power, together with the available transfer capacity and electricity demand (Figueiredo, da Silva, & Cerqueira, 2015).

### 6.4.2 Statistical analysis of difference settlement price time series (PLD)

In order to test the hypothesis of equality of PLD prices, among groups 1 to 12:  $P_{SE}, M_{SE}, L_{SE}, P_S, M_S, L_S, P_{NE}, M_{NE}, L_{NE}, P_N, M_N, L_N: \mu_0 = \mu_1 = \mu_2 = \dots = \mu_{12}$ , we performed an Analysis of Variance (ANOVA) (Figure 20). The initial objective is to identify the existence of at least one difference among the prices in the PLD groups. As the  $P - value$  for the  $F$  statistic is significant, there is evidence that at least one difference among the average prices would be significant.

Figure 20 – Analysis of Variance (ANOVA): response of PLD prices for the analyzed groups.

Response: Prices					
Degrees of Freedom	Sum of Squares	Mean of Squares	F	P-value	Pr(>F)
11	976023	88729	2.299	0.008	**
8400	324076244	38581			

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

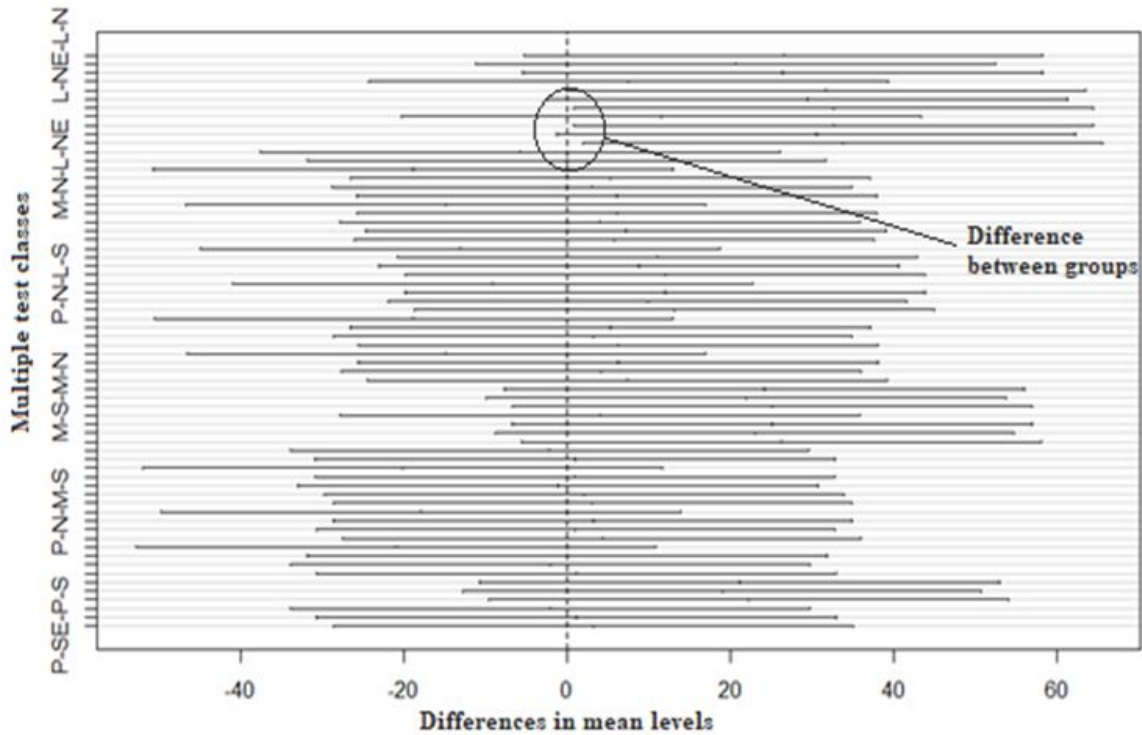
Source: Research results.

To check whether the results obtained in the ANOVA are satisfactory, it is recommended that the following conditions be tested: (i) are the standard deviations of groups constant over time? (ii) is the response distribution normal for each group? To test for homoscedasticity, we performed the Bartlett Test (Bartlett, 1937), and obtained a P-value equal to 0.448. To detect normality, we performed the Shapiro test (Shapiro & Wilk, 1965), which gave a  $P - value$  of 0.501. The non-normality of these results corroborates findings previously reported in the literature (R. V. Gomes, 2015). Furthermore, since the DTSF is based on a non-parametric model, our goal was not to perform statistical inference analysis.

From this evidence, it is interesting to identify where the differences among the means of the analyzed groups lie. In particular, the process of comparing means examines them two by two, using the Multiple Comparisons method proposed by Tukey (Tukey, 1949) (Figure 21). In general, most of the series have equal mean values. This reinforces the thesis that the PLD behavior is stable concerning the charge amount and the associated geographic region.

In general, the main differences are centered in the north of Brazil. This is reasonable, since this region is not fully connected to the national electricity system and, thus, has specific characteristics (Böckler & Pereira, 2019).

Figure 21 – Tukey’s Honestly Significant Difference (HSD) test.



Source: Research results.

### 6.4.3 Case study: applying *DTSF* to the heavy southeast energy sub-market

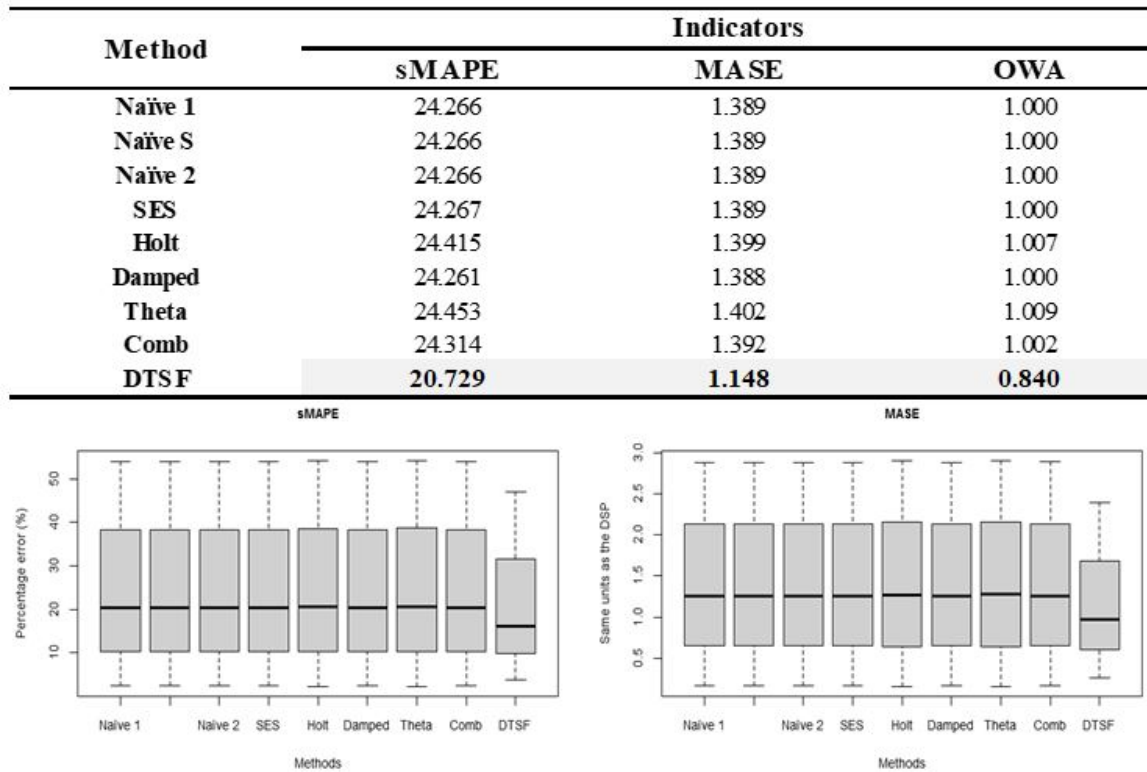
According to the statistical evidence presented in section before, there is robust indication of a similar pattern among the mean values for the twelve analyzed time series. Thus, we assume that the *DTSF* approach will generate similar forecastings for each group. Given that, our case study is based only on the *PLD* in the southeast of Brazil, and considers the heavy load level of energy. This choice is justified because the southeast is the wealthiest region of the country and concentrates nearly 55% of the country’s gross development product (E. S. d. Santos, Juchem, & Maduro, 2017). Lastly, we kept the heavy load level as it relates to industrial energy demand (Munhoz, 2017).

As expressed in the Material and Methods, the new predictive *DTSF* model is faced to the eight “seminal statistical models” (classified here as the benchmark models), frequently used in the technical literature on forecasting competitions, such as the classical one *M4* Competition (Makridakis & Hibon, 2000; R. J. Hyndman, 2020).

There is no evidence of seasonal influence in the time series analyzed, since the variations of the Naïve method resulted in equivalent predictive metrics (Figure 22). We also observed that the benchmark models presented similar results. On the other hand, according to the *OWA*, the *DTSF* method presented better performance than the Naïve (baseline) model. *DTSF* presents an efficient resolution for the Brazilian *PLD* prediction in the southeast region, considering the

heavy load level of energy. Figure 22 graphically illustrates the fit of the models according to the *sMAPE* and *MASE* statistics.

Figure 22 – Comparison between the *DTSF* and the analyzed forecasting models (benchmark).



Source: Research results.

Nevertheless, *DTSF* should be examined in more detail. In Figure 17 we presented the time series utilized and, according to that time series, the discontinuous nature of the observations over the weeks is evident. This occurs for different reasons, ranging from political influence on the energy pricing process (Guimarães & Piefer, 2017) to theoretical limitations in defining the difference settlement price in Brazil (R. V. Gomes, 2015).

To strengthen our predictive study, we divided the time series according to the seasons (spring, summer, autumn, and winter). Following this division, the predictive potential of the models was re-examined (Figure ??amemm03). In general, the best performances were, respectively: (i) spring (*DTSF*); (ii) summer (Naïve and Damped); (iii) autumn (Naïve) and (iv) winter (Damped).

It is noteworthy that our results are generally consistent. *DTSF*, based on the scanning method, exhibited predictive worsening due to the time series cutoff. Nevertheless, even with the cutoff, it was more statistically accurate in the spring and winter forecasts. Although all models have high predictive errors (spring), *DTSF* produces better results. The same goes for the inverse situation: whenever models are highly accurate (winter), *DTSF* is among the best predictors. Finally, due to the high discontinuity pattern of the analyzed series, the *Naïve* model presented good overall performance.



Figure 23 – Comparison between the *DTSF* and the analyzed forecasting models (benchmark).

Seasons	Method	Indicators		
		sMAPE	MASE	OWA
Spring	Naïve 1	45.898	1.149	1.000
	Naïve S	45.898	1.149	1.000
	Naïve 2	45.898	1.149	1.000
	SES	45.949	1.150	1.001
	Holt	45.756	1.144	0.996
	Damped	45.935	1.150	1.001
	Theta	47.993	1.221	1.054
	Comb	45.880	1.148	1.000
	<b>DTSF</b>	<b>41.727</b>	<b>1.013</b>	<b>0.896</b>
Summer	Naïve 1	37.220	2.645	1.000
	Naïve S	37.220	2.645	1.000
	Naïve 2	37.220	2.645	1.000
	SES	37.224	2.645	1.000
	Holt	38.329	2.743	1.034
	Damped	37.224	2.645	1.000
	Theta	37.617	2.680	1.012
	Comb	37.594	2.678	1.011
	<b>DTSF</b>	63.340	5.435	1.878
Autumn	Naïve 1	24.266	1.060	1.000
	Naïve S	24.266	1.060	1.000
	Naïve 2	24.266	1.060	1.000
	SES	24.443	1.070	1.008
	Holt	25.248	1.114	1.046
	Damped	24.437	1.070	1.008
	Theta	25.380	1.122	1.052
	Comb	24.712	1.085	1.021
	<b>DTSF</b>	28.745	1.327	1.218
Winter	Naïve 1	0.901	0.149	1.000
	Naïve S	0.901	0.149	1.000
	Naïve 2	0.901	0.149	1.000
	SES	0.901	0.149	1.000
	Holt	3.081	0.517	3.447
	Damped	0.862	0.142	0.957
	Theta	1.560	0.259	1.737
	Comb	1.614	0.268	1.797
	<b>DTSF</b>	0.901	0.149	1.000

Source: Research results.

Our results, although early, shed light on some future research possibilities such as detailed examination of DTSF performance in other global markets, such as in the Nordic countries. These countries have a sturdy energy pricing system, "Nord Pool" (Botterud, Kristiansen, & Ilic, 2010), that use hourly pricing information for electricity. Thus, the available information exceeds 50,000 observations for each market, (country) which allows it to produce a bright body of investigation. Last, but not least, we suggest the examination of different *DTSF* parameters through a grid search inquiry. Also, we recommend testing the precision of models by cutting time series into hundreds of subsamples and then verifying the predictive power of DTSF under different circumstances.

## 6.5 Conclusions

Analysis of the electricity spot market is complicated, involving the relationship between forecasting models and uncertainty, distinctly regarding the stochastic behavior of variables. The National System Operator regulates the Brazilian spot price of electric energy, which weekly discloses a new electric energy price to minimize the operating cost of the whole system. The present paper is aimed at the policymakers, offering a big data analysis of the scientific research of electricity. It also proposes a new forecasting approach, based on Scan-clustering modeling of the buying and selling of energy in future markets.

Although still nascent, research on the electricity spot market is increasing. As proof of this, the annual growth of publications during the survey period (1992 to 2018) corresponds to 8.33%, with an average of  $4.62 \pm 2.78$  papers published per year. This indicates the growing significance of this research. There may be several reasons for this, including the development of technologies or the growing interest in intelligent and automated networks by agents. Given the above, we noted that recent research shown a particular interest in some specific study domains, such as non-parametric models.

We present the first application of *DTSF* based on the electricity spot market, and apply our results to the eight “seminal statistical models” regularly used in forecasting competitions. In general, the eight benchmark models presented similar results. However, *DTSF* exhibited the best performance, as measured by all accuracy metrics. Subsequently, we examined the accuracy of the *DTSF* against timeframes in the observations, taking into account the seasons. *DTSF* was designed for series with high frequency and its predictive performance remained high when all other models had high predictive errors (spring). Even when other models were highly accurate (winter), *DTSF* remained among the best ones. There were other cases, however, in which we observed the *DTSF* performance to falter.

For future research, we recommend fine-tuning the study of *DTSF* parameters, as well as testing the accuracy of the models by randomly splitting time series into subsamples. Further, we recommend the development of case studies that will take the pricing of the electricity spot market in different regions, such as the Nordic and Iberian countries that dispose of massively available, high frequency data, into consideration.

Finally, the present research contributes to the energy planning processes of different players, given that understanding of the price patterns has singular importance in minimizing risks and supporting reliable production planning. Good forecasts for future energy pricing can support operational arrangements, e.g., when the energy price is high, it may be more valuable for an industry temporarily to delay part of its production, trade the surplus electricity, and carry out preventive maintenance on machines and accessories.

# 7 Application of a data-driven *DTSF* and benchmark models for the prediction of electricity prices in Brazil: a time-series case

## Abstract

The global energy market has developed significantly in recent years, proof of this is the creation and promotion of smart grids and technical advances in energy commercialization and transmission. Specifically in the Brazilian context, with the recent modernization of the electricity sector, energy trading prices, previously published on a weekly frequency, are now available on an hourly domain. In this context, the definition and forecasting of prices become an increasingly important factor for the economic and financial viability of energy projects. In this scenario of changes in the local regulatory framework, there is a lack of publications based on the new hourly prices in Brazil. This paper presents, in a pioneering way, the Dynamic Time Scan Forecasting (DTSF) method for forecasting hourly energy prices in Brazil. This method searches for similarity patterns in time series and, in previous investigations, showed competitive advantages concerning established forecasting methods. This research aims to test the accuracy of the *DTSF* method against classical statistical models and machine learning. We used the short-term prices of electricity in Brazil, made available by the Electric Energy Commercialization Chamber (CCEE).

**Keywords:** Electricity price forecasting. Statistical models. Machine learning. Time series. Dynamic time scan forecasting. Benchmark models. *M4* competition. Hourly prices.

## 7.1 Introduction

Energy planning policies arouse the interest of regulatory agencies, local governments, and the business sector. However, reconciling the interests of all the agents involved is not a simple task (Bhattacharyya, 2019) since, for the management to be fulfilled, it is necessary to achieve simultaneous success in energy supply, attracting investments, the fiscal balance of the government, and tariff modicity (Da Silva, Costa, Ahn, & Lopes, 2019). Additionally, investing in renewable energies in the present portends reducing the use of fossil fuels in the future, thus generating a positive externality for society (Tjørring & Gausset, 2015). Therefore, the promotion



of energy policies favors regional development and, consequently, an improved standard of living for individuals (Xu et al., 2019).

Due to the complexity of this issue, and the number of variables involved, public policies for energy trading occupy a prominent place in the energy industry since such policies should provide security in the investment environment (Lammers & Hoppe, 2018). Thus, a safe marketing regime is one that accurately signals the price of electricity to agents, allowing them adequately to remunerate the efficiency, reliability and flexibility of the energy generating sources (Hussain et al., 2018; Wan et al., 2016).

In this context, the Brazilian government defined the attributions of the Electric Energy Trading Chamber (CCEE) with Decree No. 5,177/2004 (Brazil, 2004). One of the CCEE's main responsibilities is to account for the amount of electricity sold in the National Interconnected System (SIN), as well as to promote settlement for the operational values of the purchase and sale of electricity in the Short-Term Market (MCP) (R. Gomes & Poltronieri, 2018). The same Decree also establishes that the valuation of the amounts settled in the MCP be used for the Settlement Price of Differences (PLD). This price is calculated weekly by the CCEE, considering sub-regional energy markets and load levels to be marketed (Ebert & Sperandio, 2018).

The basis for calculating the *PLD* is the Marginal Operating Cost (CMO), derived from the mathematical methods (Newave and Decomp) used by the National Electric System Operator (ONS) to define the system operation schedule. It should be noted that this arrangement is delimited by a minimum price and a maximum price, established annually by the National Electric Energy Agency (ANEEL) (Maceira, Melo, & Zimmermann, 2016).

Despite its relevance to the free energy market, the Brazilian *PLD* is undergoing reformulations. Accordingly, the Ministry of Mines and Energy (MME) has developed a plan for the modernization of the electrical system with Ordinance No.300/2019. The proposals include improvements to the existing computational models for the operation of the national electricity system and adoption of a new method (based on hourly prices) for pricing electricity in the Brazilian spot market. The hourly *PLD* is evolving and will come into effect completely in 2021. The goal is to bring the price of energy closer to that of the National Electric System (Abreu, de Souza, & Ribeiro, 2020; Marchetti & Rego, 2022; Munhoz, 2021).

The purpose of these methodological arrangements is to stimulate energy pricing in a context of demand response programs (Kalavani et al., 2019; Jordehi, 2019); i.e., to assign value to energy according to the moment of production, with higher prices at times of higher demand or lower generation, for example. This should lead to efficiency gains for the electrical system, in the long term. At the same time, the changes in the *PLD* will bring the Brazilian trading system closer to international systems that already adopt hourly prices. These systems include: (i) The Nordic Electricity Market - Nord Pool (Haugom et al., 2020); (ii) The Italian Electricity Market - Mercati Energetici Manager (GME) (Ilea et al., 2017); and (iii) The Iberic Electricity Market –

Iberian Electricity Market (MIBEL) (Mota et al., 2021), among others.

In this context, one of the most relevant tasks for the proper functioning of energy planning systems is Developing forecasting models, which is a difficult task. Particularly concerning electricity prices, accurately predicting their forthcoming values makes it possible to minimize planning risks. This fact becomes even more relevant in the current scenario of global energy insecurity, derived from factors such as the war between Russia and Ukraine (Steffen & Patt, 2022) and the repeated interventions of the Organization of the Petroleum Exporting Countries (OPEC) in oil prices over the last few decades (Lin, Omoju, & Okonkwo, 2015; Harris, Bitonti, Fleisher, & Binderkrantz, 2022). Given the above, this scenario reinforces the need to encourage research and development projects related to the energy market.

According to that the objectives of the present paper are to compare statistical-computational models and machine learning approaches that allows an accurate prediction of electricity prices in the spot market, considering the new price structure in Brazil.

We organized the present paper into four sections. Following this Introduction, Section 2 brings the material and methods utilized, focusing on datasets, the formulation of dynamic time scan forecasting, benchmark models, and evaluation metrics. Section 3 provides the results and discussions obtained from the proposed methodologies. Finally, Section 4 concludes the paper and includes limitations and recommendations for future research.

## 7.2 Materials and Methods

### 7.2.1 Dataset

The data used in the current paper comes from the Brazilian Chamber of Electric Energy Commercialization, available on <https://www.ccee.org.br/web/guest/precos/painel-precos>. It consists of the short-term electricity prices in Brazil, called "difference settlement price" (PLD). The *PLD* is determined in hourly frequency, considering four submarket and three load levels. The submarkets are defined by the National Operator of the System (ONS) and consider the geographical divisions: North, Northeast, South and Southeast/Center-West (Gontijo et al., 2021; T. Santos, Diniz, Saboia, Cabral, & Cerqueira, 2020).

The *PLD* is *ex-ante* based, considering expected availability and load information. The *PLD* must sell out all the energy, not just the contracted energy, among the agents (R. J. Hyndman & Khandakar, 2008). The analyzed sample consists of hourly *PLD* data (R\$/MWh) collected by the *CCEE*, from 01 January 2019 to 31 December 2021, since they are the first and last year, respectively, with complete information available ( $n = 26305$  hours). According to Gontijo, Costa e De Santis (2021) the *PLD* pattern is stable concerning the load charge and the geographical regions. Thus, this paper used the *PLD* in the Southeast/Center-West of Brazil, since this region is responsible for the leading share of the Brazilian gross domestic product (E. S. d. Santos et al.,

2017). Additionally, we fixed our analysis on the heavy load level of energy because it relates to the industrial demand (Munhoz, 2017).

## 7.2.2 Dynamic Time Scan Forecasting

Dynamic Time Scan Forecasting (*DTSF*) is a forecasting procedure based on scan statistics (Kulldorff, 1999; Abolhassani & Prates, 2021), initially developed to handle wind forecasting for Brazilian power generation plants (Costa et al., 2021). It consists of scanning a time series and identifying past patterns ("analog") like the last available observations at a given time ("query") (Gontijo et al., 2021).

According to (Costa et al., 2021), the *DTSF* method scans a times series utilizing a specified window size. Let  $y_t$  be a time series of length  $N$ ,  $t = 1, \dots, N$ . Firstly, let vector  $\mathbf{y}^{[w]}$  be defined as the last  $w$  observations of the series:

$$\mathbf{y}^{[w]} = [y_{N-w+1}, \dots, y_N]. \quad (7.1)$$

As described, *DTSF* seeks to strongly identify subsets of the time series correlated with the vector  $\mathbf{y}^{[w]}$ . This occurs by running a scanning window with the same size of vector  $\mathbf{y}^{[w]}$  to scan subsets of the previous values in the time series. The set of candidate subsets are:

$$\mathbf{x}_t^{[w]} = [y_{t-w+1}, \dots, y_t] \quad (7.2)$$

where  $t = 1, \dots, N - 2 \cdot w$ . The upper limit of the time sequence ( $N - 2 \cdot w$ ) guarantees that vector  $\mathbf{x}_t^{[w]}$  does not overlap with vector  $\mathbf{y}^{[w]}$ . Given the last  $w$  observed values, which comprises vector  $\mathbf{y}^{[w]}$ , a rolling window with the same size ( $\mathbf{x}_t^{[w]}$ ) is used for scanning previous values of the series.

Lastly, *DTSF* provides a  $k - steps$  ahead forecast of the time series,  $y_{N+1}, \dots, y_{N+k}$ . To produce this outcome, the *DTSF* scans the series to find the closest analogs  $\mathbf{x}_t^{[w]}$ . The subsequent values of the time series are used as the forecast values:

$$y_{N+i} = f_{\mathbf{x}_t^{[w]}}(y_{t-w+i}) \quad (7.3)$$

where  $f_{\mathbf{x}_t^{[w]}}$  is a function which correlates the elements of vector  $\mathbf{x}_t^{[w]}$  and the elements of vector  $\mathbf{y}^{[w]}$ .

According to that, a first constraint can be set on  $k : 1 \leq k \leq w$ . This constraint guarantees that if the most correlated time series window comprises the most recent values, prior to vector  $\mathbf{y}^{[w]}$ , then the forecast values are a function of vector  $\mathbf{y}^{[w]}$ ,

$$y_{N+i} = f_{\mathbf{x}_{N-2w}^{[w]}}(y_{N-w+i}). \quad (7.4)$$

As stated before, forecast values depend on the window length  $w$  and the function  $f_{\mathbf{x}_t^{[w]}}(\cdot)$ . A intuitive proposal for function  $f_{\mathbf{x}_t^{[w]}}(\cdot)$  is a linear scaling of the elements of vector  $\mathbf{x}_t^{[w]}$ , i.e., a linear model. This occurs due to the fact that previous values are likely similar to the last observations, except for a scale and/or offset shift. So, the method searches for values that may be similar to the last values, after applying a similarity function (Costa et al., 2021).

By taking a linear function as the similarity function, the parameters of the model can be estimated to minimize the sum of squares between the elements of vector  $\mathbf{y}^{[w]}$  and the linear equation:  $\beta_0^{[t]} + \beta_1^{[t]} \times \mathbf{x}_t^{[w]}$ . Moreover, the similarity statistic can be assumed as the linear regression coefficient of determination  $R^2$  (Costa et al., 2021; ?, ?):

$$R^2 = 1 - \frac{\sum_j (\mathbf{y}_j^{[w]} - \hat{\mathbf{y}}_j^{[w]})^2}{\sum_j (\mathbf{y}_j^{[w]} - \bar{\mathbf{y}}_j^{[w]})^2} \quad (7.5)$$

where  $\mathbf{y}_j^{[w]}$  is the  $j$ -th value of vector  $\mathbf{y}^{[w]}$  and  $\hat{\mathbf{y}}_j^{[w]}$  is the  $j$ -th predicted value using the estimated linear function. Finally, the method calculates a similarity profile based on the  $R^2$  score resulting from the comparison of the query with previous windows. The analogs with higher  $R^2$  scores are considered closer analogs. Predictions of future steps are calculated from a predefined number of analogs using aggregation functions, such as median (Costa et al., 2021).

### 7.2.3 Benchmark Models

A univariate predicting approach is a procedure for estimating a point forecast. The forecasts are based on past and present values of a given time series (Pal & Prakash, 2017; Bisht & Ram, 2021). Given the benefit of simplicity and high usage, the literature applies univariate forecasting methods in several problems in different areas, such as energy and finance. A 24 – hour predictive window was used in this paper. We selected the following benchmark methods from the classical statistical and machine learning literature, such as the  $M$  – competition (Makridakis et al., 2018), and a description is provided for each one, as follows:

1. *ARIMA*: Auto tuning model, with contains the “auto.arima” function in *R*, a variation of the Hyndman-Khandakar algorithm (S. Wang, 2006; R. J. Hyndman & Khandakar, 2008), which combines unit root tests, minimisation of the Akaike or Bayesian information criterion and maximum likelihood estimation to obtain an *ARIMA* model.
2. *DTSE*: the proposed method, adopting the defined default parameters, which are: (i) polynomial function degree equal to 1, (ii) analogs equal to 10, (iii) window size equal to length of forecast horizon, 24, and (iv) median as aggregation function (Costa et al., 2021).
3. *ETS*: Automatic forecasting procedure based on a range of exponential smoothing methods. Available through the *ets* function in *R*. The *ETS* model deals with trend and

seasonality in datasets and other prior assumptions about the time series (Hand, 2009; R. Hyndman et al., 2008; R. J. Hyndman et al., 2002).

4. *Naïve*: the straightforward, yet a still robust predictive procedure. It assumes that the  $k - step$  ahead forecasts to be predicted are equal to the last given observation. Accessible through the *naive* function in (Makridakis & Hibon, 1979).
5. *TBATS*: It consists in an exponential smoothing procedure, that incorporates *Box – Cox* transformation, an *ARMA* model for residuals, and the trigonometric seasonal component. The trigonometric seasonality term can potentially reduce model parameters when high seasonality frequencies is applicable. Available through the *tbats* function in *R* (Livera, Hyndman, & Snyder, 2011).
6. *Theta*: a procedure based on a coefficient of curvature of the time series, applied to the second difference in the data. Achievable through the *Theta.classic* function in *R* (Assimakopoulos & Nikolopoulos, 2000).
7. *XGBoost*: It applies to time series the Extreme Gradient Boosting procedure (Friedman, 2001). The basic idea of *xgboost* deals with extrapolation into a new range of variables not in the training set. In order treat the the non-stationarity of the series, the first difference in electricity prices was taken. To deal with the seasonality in data we constructed a set of pairs of Fourier transform variables and take them as regressors ones. Disposable through the *xgbar* function in *R* (Khuhawar, Siddiqui, Arain, Siddiqui, & Qureshi, 2021).

## 7.2.4 Forecast Evaluation

A 24 – hour predictive window was used in this paper. The split of the data into training sets and test sets split considered the definition of 10 random test days, considering 24 hours of prices for the whole series, totaling 240 test points. The choice of the number of test points considered evidence from the literature in other forecasting studies (Rayas-Sánchez, Aguilar-Torrentera, & Jasso-Urzúa, 2010; Koziel & Bandler, 2008). The testing points, are, respectively: (i) 2021 – 03 – 17; (ii) 2021 – 05 – 08; (iii) 2021 – 05 – 31; (iv) 2021 – 06 – 13; (v) 2021 – 07 – 26; (vi) 2021 – 08 – 04; (vii) 2021 – 08 – 26; (viii) 2021 – 10 – 08; (ix) 2021 – 12 – 07 e (x) 2021 – 12 – 21.

The forecast evaluation metrics assumed are the same adopted in the *M – Competition* and are those most used in the literature (Islam & Al-Alawi, 1996; Azadeh et al., 2008): Mean Absolute Scaled Error (MASE), Overall Weighted Average (OWA), and the Symmetric Mean Absolute Percentage Error (sMAPE). The equation for calculating these metrics is given:

$$sMAPE = \frac{1}{h} \sum_{t=1}^h \frac{2|Y_t - \hat{Y}_t|}{|Y_t| + |\hat{Y}_t|} \quad (7.6)$$

$$MASE = \frac{1}{h} \frac{(n - m) \sum_{t=1}^h |Y_t - \hat{Y}_t|}{\sum_{t=m+1}^n |Y_t - Y_{t-m}|} \quad (7.7)$$

$$OWA = \frac{sMAPE_k / sMAPE_{base} + MASE_k / MASE_{base}}{2} \quad (7.8)$$

where  $Y_t$  represents the *post - sample* value of the time series at point  $t$ ,  $\hat{Y}_t$  is the estimated forecast,  $h$  is the forecasting horizon,  $m$  is the frequency of the data,  $k$  is a given regressor, and the base is the *sNaïve* estimator.

### 7.2.5 Hardware and Software

All routines utilized in this paper were executed using the *R* 4.1.2 programming language (R Core Team, 2021). The *Forecast* package (version 8.17.0) was used for the estimations of *ARIMA*, *ETS*, *Naïve*, and *TBATS* models (<https://cran.r-project.org/web/packages/forecast/index.html>). *DTSF* and its original implementation in *R* and *C++*, is available from the repository (<https://rdr.io/github/leandrominetti/DTScanF/>). Theta model, the top performing benchmark of the *M4* forecast competition is disposable at the official repository (<https://github.com/Mcompetitions/M4-methods>). Finally, the fitting of the *XGBoost* model was achieved by the *forecastxgb* package (<https://github.com/ellisp/forecastxgb-r-package>). Hardware specifications adopted to perform the forecasting are CPU Intel Core *i3 - 6100U*.

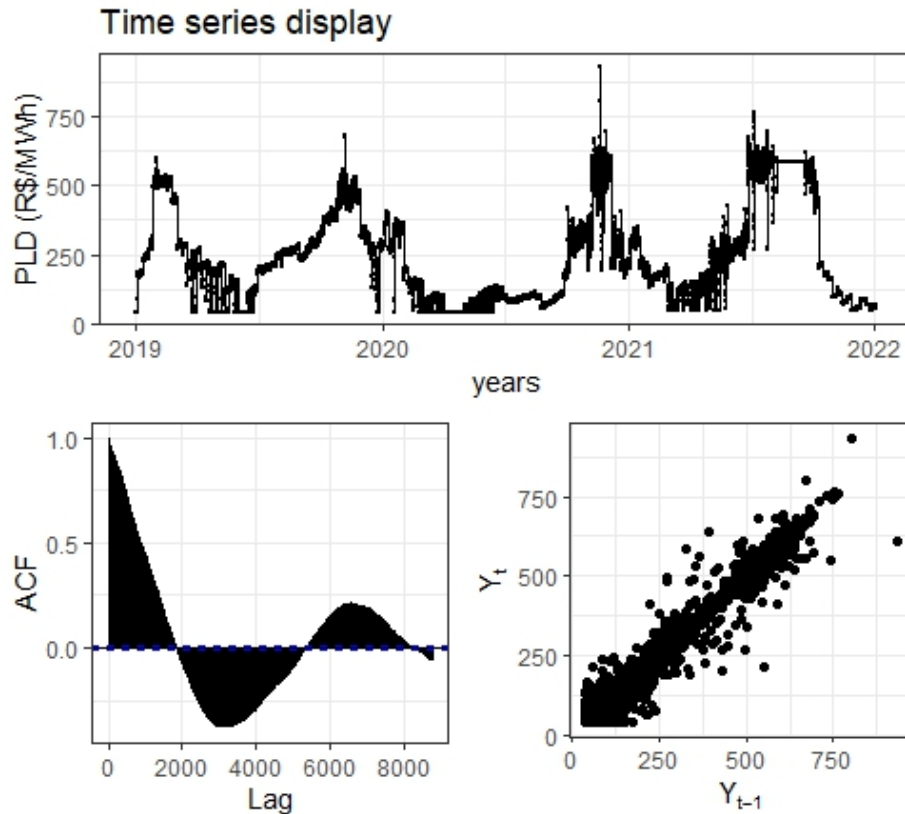
## 7.3 Exploratory data analysis (EDA)

The *PLD* hourly time series is volatile, due to different factors: (i) this is a new arrangement pricing system in Brazil, and energy trading is not yet a complete free trade environment (Marchetti & Rego, 2022; T. Santos et al., 2020); (ii) the recent water crises in Brazil have caused price instability (Hunt, Stilpen, & de Freitas, 2018); (iii) other factors may explain price volatility, such as the uncertainties arising from the coronavirus crisis (Zhong, Tan, He, Xie, & Kang, 2020) and the war between Russia and Ukraine (Johannesson & Clowes, 2022) and (iv) finally, political instability in Brazil and Latin America may have an influence on production and energy consumption, and consequently prices (Chevalier, 2009) (Figure 24).

According to Figure 24 (bottom), the *PLD* time series, through the Autocorrelation function (ACF) and the scatterplot between the current observation and their respective lags, show strong evidence of the need for a modeling that considers autoregressive vectors. One way to model such behavior is to adopt family models, which can also verify the degree of stationarity of the base and the need not to differentiate the data. Given this behavior, we selected the auto Arima model as one of the benchmark models.



Figure 24 – Time series display of PLD (R\$/MWh), considering ACF and lag correlation.



Source: Research results.

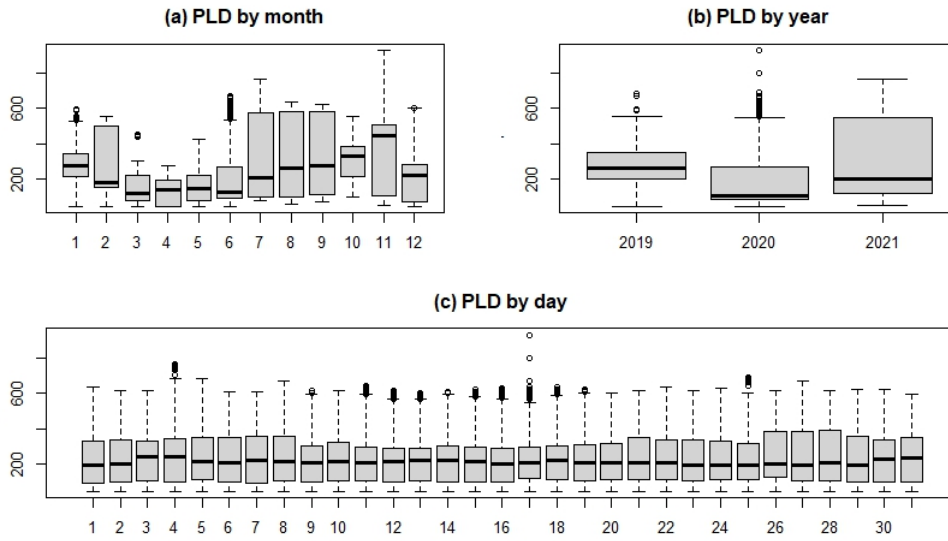
According to Figure 25, the Brazilian PLD exhibited an annual and monthly variation. The effect of the daily variation is relatively smaller than the previous ones. The presence of this seasonal pattern justifies the choice of the automatic adjustment *ETS* model. *ETS* selects the best fit, including the main classes of models with the trend, and seasonality, such as Simple Exponential Smoothing, *Holt* model, *Holt – Winter*, *SeasonalNaïve*, and all their respective extensions.

The next subsection presents the results obtained with the predictive models. The estimates were carried out based on previously established performance metrics, namely: *sMAPE*(%), *MASE* and *OWA*.

### 7.3.1 Forecasting results

As expressed in the material and methods the new predictive model *DTSF* was confronted with six methods: *Arima*; *Ets*; *Naïve*; *Tbats*; *Theta* and *XGBoost*. According to the *sMAPE*(%) indicator the *DTSF* exhibited the best predictive performance. The predictions made by the *Arima*, *DTSF* and *XGBoost* model showed similar values. It is also noteworthy that the *Ets*, *Tbats* and *Theta* showed greater predictive variability compared to the other procedures. The *sMAPE*(%) variations concerning the 24 – steps ahead show that the models present varied performances depending on the horizon. The *DTSF* remains competitive in

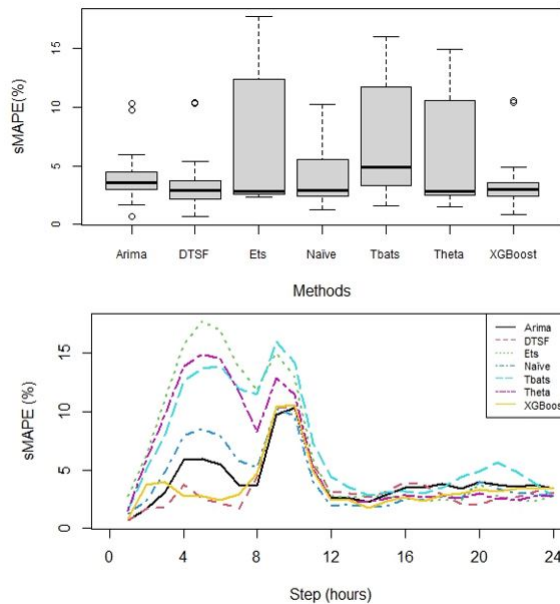
Figure 25 – Boxplot of annual, monthly, and daily PLD (R/*MWh*) variations.



Source: Research results.

all analyzed steps ahead. There is also a slight convergence of predictions from the 12<sup>th</sup> step onward (Figure 26).

Figure 26 – *sMAPE*(%) – Comparison between the *DTSF* and the benchmark models.



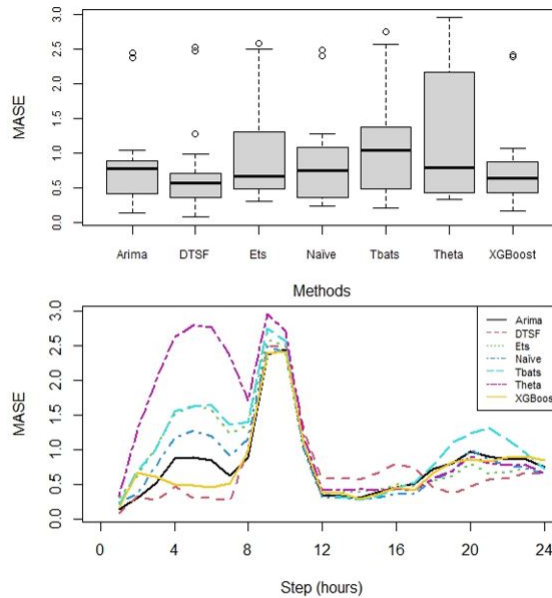
Source: Research results.

Complementarily, Figure 27 below illustrates the performance of the selected models regarding the *MASE* measure. The results are like the previous ones, showing, through another evaluation metric, the predictive power of the *DTSF* compared to the benchmark models. Again, there is no one optimal model for all the predictive steps ahead; this fact suggests that a decision-making system on electricity prices must consider switching models; that is, for certain hours of the day, there are models that stand out over others. Figure 36 and 37, inserted in the



Appendix of this thesis, present the numerical values for each predictive model, concerning the 24 predictive steps ahead, for the *sMAPE*(%) and *MASE* measurements, respectively.

Figure 27 – *MASE* - Comparison between the *DTSF* and the benchmark models.



Source: Research results.

According to (Makridakis et al., 2020) there are several approaches disposable in the literature for evaluating the performance of benchmarking methods (Kim & Kim, 2016; R. J. Hyndman & Koehler, 2006). In the historic of *M – Competitions*, many of the accuracy indicators were used without any justification concerning advantages and disadvantages of each one (Paul & Richard, 1999). This way, we decided to take another accuracy measure, one of the most popular in the forecasting literature, called Overall Weighted Average (OWA). We believe this would reinforce us to obtain higher level of reliability. Thus, the best predictive model based on *OWA* will be evaluated considering the two previously calculated measures, namely: *sMAPE*(%) and *MASE*. According to the *OWA*, the *DTSF* showed the best predictive potential, followed by the *XGBoost* model (Table 15).

Table 15 – Comparison between the *DTSF* and the analyzed forecasting models.

Model	MASE	sMAPE(%)	OWA
Arima	0.803	4.080	1.000
<b>DTSF</b>	<b>0.712</b>	<b>3.448</b>	<b>0.866</b>
Ets	0.950	6.848	1.431
Naïve	0.846	4.231	1.045
Tbats	1.063	6.882	1.505
Theta	1.273	5.979	1.526
XGBoost	0.769	3.588	0.918

Source: Research results.

Despite its positive points, it is relevant to investigate the *DTSF* model in a more

in-depth analysis. Future research should analyze *PLD* time series according to the seasons (spring, summer, autumn, and winter). This procedure could help us to validate the model fitting through other perspectives.

## 7.4 Conclusions

Analysis of the electricity spot market deals with a complicated task since it deals with the relationship between forecasting models and uncertainty. This theme becomes even more relevant in the current context of global energy insecurity, highlighted by the war between Russia and Ukraine and the instability in oil prices, resulting from interventionist actions by the Organization of Petroleum Exporting Countries (OPEC).

According to that, this paper aims to contribute to the policymakers, proposing a new forecasting approach, based on Scan clustering modeling for buying and selling electricity in the Brazilian market. This study is justified because the energy trading system in Brazil is undergoing reformulations. The hourly *PLD* is a new pricing model for the short-term market. It will replace the current methodology called 'level week' to another one with daily price updates.

We noted that recent research had shown a particular concern for some study domains as the non-parametric and hybrid approaches. This paper presents the first application of the DTSF method based on the spot market for electricity. We compared the forecasts generated in this paper with the main benchmarking models regularly used in the *M – Competition*.

This research has limitations, namely: (i) the database is relatively new and has irregularities; (ii) the absence of similar studies carried out in Brazil makes it difficult to compare this result with other local papers. Future research should explore other predictive horizons, considering the short, medium, and long term.

## Part III

### Complementary studies

# 8 Forecasting Hierarchical Time Series in Power Generation

## Abstract

Academic attention is being paid to the study of hierarchical time series. Especially, in the electrical sector, there are several applications in which information can be organized into a hierarchical structure. The present study analyzed hourly power generation in Brazil (2018-2020), grouped according to each of the electrical subsystems and their respective sources of generating energy. The objective was to calculate the accuracy of the main measures of aggregating and disaggregating the forecasts of the *ARIMA* and *ETS* models. Specifically, the following hierarchical approaches were analyzed: (i) *Bottom – Up* (BU), (ii) *Top – down* (TD), and (iii) Optimal Reconciliation. The Optimal Reconciliation models showed the best mean performance, considering the primary predictive windows. It was also found that energy forecasts in the South subsystem presented greater inaccuracy, compared to the others, which signals the need for individualized models for this subsystem.

**Keywords:** Power generation. Electrical subsystems. Time series.

## 8.1 Introduction

The advent of Industry 4.0 revolutionized factories worldwide, since it allowed the connectivity between measuring machines and the automation of companies, distributing the capacity to collect massive volumes of data (Medojevic, Medic, Marjanovic, Lalic, & Majstorovic, 2019). In high-level data analysis, forecasting models allow the extraction of behavior patterns, as well as the prediction of future scenarios for the collected data set (Alcácer & Cruz-Machado, 2019).

There are several forecasting applications relevant to power generation, the object of the present study, including: (i) classical time series models (H. Chen, Wan, Li, & Wang, 2013), (ii) pre-processing techniques (Malvoni, De Giorgi, & Congedo, 2017), (iii) machine learning approaches (Sharifzadeh, Sikinioti-Lock, & Shah, 2019), among others. Additionally, an alternative class known as hierarchical forecasting (Athanasopoulos, Ahmed, & Hyndman, 2009) deals with organized time series that can be aggregated at different levels into groups based on geography, sources of energy, or other, specific features.

Despite this being a recent topic, there is already research that has addressed the application of hierarchical forecasting models in the energy sector. Some examples of applications

are: electrical grids (Almeida, Ribeiro, & Gama, 2016), forecasting models for air pollution (Kosiorowski, Mielczarek, Rydlewski, et al., 2017), solar power generation (Panamtash & Zhou, 2018), energy transport (Abouarghoub, Nomikos, & Petropoulos, 2018), among others.

The papers identified above have calibrated the forecasts using only the *Bottom – Up*, *Top – down*, and *OLS* assumptions (R. J. Hyndman, Ahmed, Athanasopoulos, & Shang, 2011). Thus, the following research question is formulated: how is it possible to make hierarchical predictions using advanced linear regression models with regularization? We can obtain more reliable forecasts if we rewrite the hierarchical problem in terms of finding a set of unbiased, minimum variance measures of projected values across the whole array of data. It is possible to minimize the sum of variances of the reconciled estimate errors under the property of unbiasedness, using the procedure called *MinT* (minimum trace) reconciliation (Wickramasuriya, Athanasopoulos, & Hyndman, 2019).

The present paper presents a case study using a power generation data set from Brazil (2018-2020) organized by electrical subsystems and different generating sources. Specifically, the main approaches used to aggregate and disaggregate predictions made for grouped time series are examined, namely: (i) *Bottom – Up*, (ii) *Top – Down* and (iii) Optimal reconciliation models (OLS, WLS and MinT). The predictive models *ARIMA* and *ETS* were used to test the performance of these reconciliation methods.

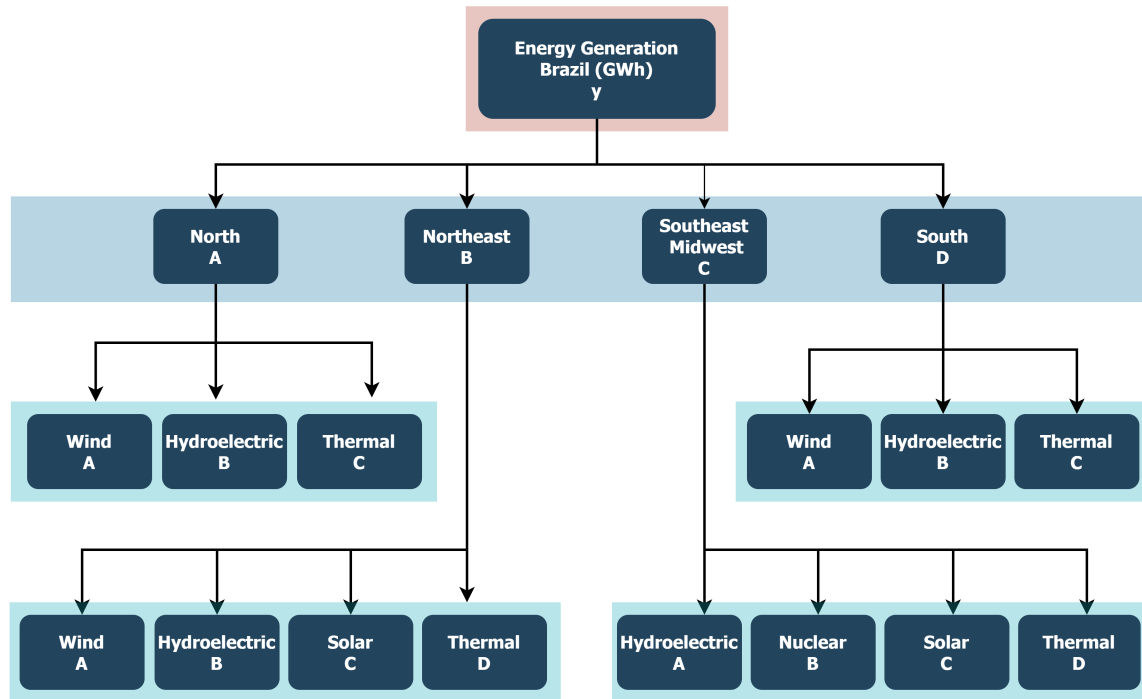
The remainder of the present paper is organized as follows. Section 2 defines the study methodology, describing the data set, hierarchical procedures and forecasting models employed. Section 3 presents the results and discussions of the techniques, in addition to the limitations of this paper. Finally, Section 4 presents the conclusions and guidelines for future work.

## 8.2 Materials and Methods

The secondary data used in this study correspond to the amounts of power generated by each of the Brazilian electrical subsystems (North, Northeast, Southeast/Midwest, and South). We separated these data according to the source of energy (Wind, Hydroelectric, Thermal, Solar, and Nuclear). Data were obtained from the National Electric System Operator (Operator, 2020), due to their reliability. The observations of hourly power generation (GWh) were made during the period from January 2018 to January 2020, making a total of 17521 hours.

Based on R. J. Hyndman et al. (2011), we present a schematic representation of the Brazilian energy generation system, comprising a three-level hierarchical structure (Figure 28). Level 0 represents the total energy generated in Brazil (completely aggregated series). Level 1 denotes each of Brazil's electrical subsystems (first level of disaggregation). The last level, Level 2, represents each of the energy generating sources (Level k). According to this framework, it is possible to identify the most disaggregated time series (in this case  $k = 2$ ).

Figure 28 – Hierarchical aggregation structure for the energy generation in Brazil.



Source: Research results.

Table 16 shows the amounts of power generation in Brazil (GWh), according to generating sources and electrical subsystems. There is a predominance of hydroelectric generation (73%), making the Brazilian electrical matrix one of the cleanest in the world. At the same time, the Southeast/Midwest subsystem accounts for more than half (56%) of all energy generated in the country.

Table 16 – Amounts of power generation in Brazil (GWh).

Subsystem Source	Wind	Hydro	Thermal	Solar	Nuclear	Total (GWh)	%
North (A)	2688	125182	31489	0	0	159359	14.3%
Northeast (B)	85377	37705	36699	4626	0	164407	14.7%
Southeast/Midwest (C)	0	518714	73555	2437	31805	626511	56.1%
South (D)	11326	135914	19472	0	0	166712	14.9%
Total (GWh/Source)	99391	817516	161215	7063	31805	1116989	100%
%	8.9%	73.2%	14.4%	0.6%	2.8%	100%	-

Source: Research results.

Routines were implemented using the *R* programming language (R Core Team, 2021). The *R*–package *HTS* was used to calculate the *bottom–up*, *top–down*, optimal combination reconciliation and trace minimization reconciliation. *HTS* is available at: <https://cran.r-project.org/web/packages/hts/index.html>. Although *HTS* includes functions for creating, plotting and forecasting hierarchical time series, it has some limitations. Those limitations include the fact that it has only three *built–in* forecasting options: *ARIMA*, *ETS*,

and random walks (R. J. Hyndman et al., 2011). This paper will use the *ARIMA* and the *ETS* models since they have automatic adjustment and allow consideration of factors such as the trend and seasonality of the data set. The computer used to execute the algorithms had *CPU* Intel Core *i5* – 7200 2.70 GHz, RAM of 16 GB, and operating system Windows 10x64. In the next subsection, we present the hierarchical reconciliation models used in the present paper, as well as the forecasting models.

### 8.2.1 The *Bottom – Up* (BU) Approach

The *BU* procedure requires first providing forecasts for every series at the *bottom – level*, and then summing these to generate forecasts for all the levels of the hierarchical structure (Orcutt, Watts, & Edwards, 1968). In its simplicity, this approach neglects the relations between time series and works, mainly unsuccessfully, on highly disaggregated data. These data tend to have a low signal-to-noise ratio (Wickramasuriya et al., 2019). According to the hierarchy (Figure 28), we first make *h – step* ahead forecasts for all the bottom-level time series ( $n = 14$ ):

$$\hat{y}_{AA,t}, \hat{y}_{AB,t}, \hat{y}_{AC,t}, \hat{y}_{BA,t}, \hat{y}_{BB,t}, \hat{y}_{BC,t}, \hat{y}_{BD,t}, \hat{y}_{CA,t}, \hat{y}_{CB,t}, \hat{y}_{CC,t}, \hat{y}_{CD,t}, \hat{y}_{DA,t}, \hat{y}_{DB,t}, \hat{y}_{DC,t}. \quad (8.1)$$

Summing these, we obtain *h – step* ahead forecasts for the rest of the series:

$$\begin{aligned} \tilde{y}_t &= \hat{y}_{AA,t} + \hat{y}_{AB,t} + \hat{y}_{AC,t} + \hat{y}_{BA,t} + \hat{y}_{BB,t} + \hat{y}_{BC,t} + \hat{y}_{BD,t} + \hat{y}_{CA,t} + \hat{y}_{CB,t} \\ &\quad + \hat{y}_{CC,t} + \hat{y}_{CD,t} + \hat{y}_{DA,t} + \hat{y}_{DB,t} + \hat{y}_{DC,t}. \\ \tilde{y}_{A,t} &= \hat{y}_{AA,t} + \hat{y}_{AB,t} + \hat{y}_{AC,t} \\ \tilde{y}_{B,t} &= \hat{y}_{BA,t} + \hat{y}_{BB,t} + \hat{y}_{BC,t} + \hat{y}_{BD,t}. \\ \tilde{y}_{C,t} &= \hat{y}_{CA,t} + \hat{y}_{CB,t} + \hat{y}_{CC,t} + \hat{y}_{CD,t}. \\ \tilde{y}_{D,t} &= \hat{y}_{DA,t} + \hat{y}_{DB,t} + \hat{y}_{DC,t}. \end{aligned} \quad (8.2)$$

According to R. J. Hyndman et al. (2011), it is possible to arrange the equations expressed in (9.2) into an algebra notation. Below is a complete notation for this problem:

$$\begin{bmatrix} \tilde{y}_t \\ \tilde{y}_{A,t} \\ \tilde{y}_{B,t} \\ \tilde{y}_{C,t} \\ \tilde{y}_{D,t} \\ \tilde{y}_{AA,t} \\ \tilde{y}_{AB,t} \\ \tilde{y}_{AC,t} \\ \tilde{y}_{BA,t} \\ \tilde{y}_{BB,t} \\ \tilde{y}_{BC,t} \\ \tilde{y}_{BD,t} \\ \tilde{y}_{CA,t} \\ \tilde{y}_{CB,t} \\ \tilde{y}_{CC,t} \\ \tilde{y}_{CD,t} \\ \tilde{y}_{DA,t} \\ \tilde{y}_{DB,t} \\ \tilde{y}_{DC,t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} \hat{y}_{AA,t} \\ \hat{y}_{AB,t} \\ \hat{y}_{AC,t} \\ \hat{y}_{BA,t} \\ \hat{y}_{BB,t} \\ \hat{y}_{BC,t} \\ \hat{y}_{BD,t} \\ \hat{y}_{CA,t} \\ \hat{y}_{CB,t} \\ \hat{y}_{CC,t} \\ \hat{y}_{CD,t} \\ \hat{y}_{DA,t} \\ \hat{y}_{DB,t} \\ \hat{y}_{DC,t} \end{bmatrix} \quad (8.3)$$

Alternatively, the notation presented in (9.3) can be reformulated in a compact way by applying the summing matrix. Thus, the *bottom – up* approach can be represented as:

$$\tilde{y}_t = S\hat{b}_t \quad (8.4)$$

where  $\tilde{y}_t$  is an  $n - dimensional$  vector of  $h - step$  ahead forecasts for the total energy,  $S$  is the summing matrix, and  $\hat{b}_t$  is an  $m - dimensional$  vector of  $h - step$  ahead forecasts for each of the sources of energy at *bottom – level*. An advantage of this procedure is that we are forecasting at the bottom-level of a hierarchy. Consequently, no information is missed due to aggregation (R. J. Hyndman & Khandakar, 2008).

## 8.2.2 The *Top – Down* (TD) Approach

*Top – down* methods operate with strictly hierarchical aggregation structures, not with grouped structures. They involve first making forecasts for the Total level  $y_t$ , and next disaggregating these down the hierarchy (R. J. Hyndman & Khandakar, 2008). Let  $p_1, \dots, p_m$  be a set of disaggregation proportions that deliver the forecasts of the Total series, which are to be distributed in order to obtain forecasts for all series at the bottom-level of the structure. To illustrate, concerning our hierarchy by applying proportions to Figure 28, we get  $p_1, \dots, p_{14}$ :



$$\begin{aligned}
\tilde{y}_{AA,t} &= p_1 \hat{y}_t, \tilde{y}_{AB,t} = p_2 \hat{y}_t, \tilde{y}_{AC,t} = p_3 \hat{y}_t. \\
\tilde{y}_{BA,t} &= p_4 \hat{y}_t, \tilde{y}_{BB,t} = p_5 \hat{y}_t, \tilde{y}_{BC,t} = p_6 \hat{y}_t, \tilde{y}_{BD,t} = p_7 \hat{y}_t. \\
\tilde{y}_{CA,t} &= p_8 \hat{y}_t, \tilde{y}_{CB,t} = p_9 \hat{y}_t, \tilde{y}_{CC,t} = p_{10} \hat{y}_t, \tilde{y}_{CD,t} = p_{11} \hat{y}_t. \\
\tilde{y}_{DA,t} &= p_{12} \hat{y}_t, \tilde{y}_{DB,t} = p_{13} \hat{y}_t, \tilde{y}_{DC,t} = p_{14} \hat{y}_t.
\end{aligned} \tag{8.5}$$

This can be rewritten using matrix notation. If we stack the set of proportions in an  $m$ -dimensional vector  $p = (p_1, \dots, p_m)'$ , we have the bottom-level  $h - step$  ahead predictions. Overall, for a given set of proportions, *top - down* approaches can be written as:

$$\begin{aligned}
\tilde{b}_t &= p_j \hat{y}_t \\
\tilde{y}_t &= S p_j \hat{y}_t.
\end{aligned} \tag{8.6}$$

The main *TD* models stipulate disaggregation proportions according to the historical proportions of the data. Among the main models of this approach, we highlight the following three: (i) *top - down* Gross–Sohl method A (TDGSA), (ii) *top - down* Gross–Sohl method F (TDGSF), and (iii) *Top - down* forecast proportions (TDFP) (Table 2). Additional details and demonstrations of Table 17 can be obtained from (Gross & Sohl, 1990) and (Athanasopoulos et al., 2009).

Table 17 – TD disaggregation proportions according to the historical proportions of the data.

TD Gross-Sohl:A/TDGSA	TD Gross-Sohl:F/TDGSF	TD:TDFP
$p_j = \frac{1}{T} \sum_{t=1}^T \frac{y_{j,t}}{y_t}$	$p_j = \sum_{t=1}^T \frac{y_{j,t}}{T} / \sum_{t=1}^T \frac{y_t}{T}$	$p_j = \prod_{l=0}^{K-1} \frac{\hat{y}_{j,h}^{(l)}}{\hat{S}_{j,t}^{(l+1)}}$
<p>for <math>j = 1, \dots, m</math>. Each proportion <math>p_j</math> reflects the average of the historical proportions of the bottom-level series <math>y_{j,t}</math>, <math>t</math> over the period <math>t = 1, \dots, T</math> relative to the total aggregate <math>y_t</math></p>	<p>for <math>j = 1, \dots, m</math>. Each proportion <math>p_j</math> takes the average historical value of the <i>bottom - level</i> series <math>y_{j,t}</math> related to the average value of the total aggregate <math>y_t</math>.</p>	<p>where <math>j = 1, \dots, m</math>, <math>\hat{y}_{j,h}^{(l)}</math> is the <math>h - step</math> ahead forecast and <math>\hat{S}_{j,t}^{(l)}</math> is the sum of the <math>h - step</math> ahead forecasts below the node that is <math>l</math> levels above node <math>j</math>.</p>

Source: Research results.

### 8.2.3 The Optimal Reconciliation Approaches

The optimal reconciliation approach proposed by R. J. Hyndman et al. (2011) consists of an ordinary least squares problem based on the calculation of independent projections for all hierarchical levels, then applying a regression model to optimize the combination of these forecasts. According to Oliveira and Ramos (2019), we can write the base prediction as:

$$\hat{y}_{t+h|t} = S \beta_{t+h|t} + \varepsilon_h, \tag{8.7}$$

where  $\beta_{t+h|t}$  represents the unknown conditional mean of the most disaggregated series, and  $\varepsilon_h$  is the error with mean of zero and covariance matrix  $\Sigma_h$ . If  $\Sigma_h$  were known, the estimator of  $\beta_{t+h|t}$  would lead to the following weighted least squares, producing reconciled forecasts, as follows:

$$\tilde{y}_{t+h|t} = S\hat{\beta}_{t+h|t} = S(S' \sum_h^{-1} S)^{-1} S^{-1} \sum_h^{-1} \hat{y}_{t+h|t} = SP\hat{y}_{t+h|t}, \quad (8.8)$$

where  $P = (S' \sum_h^{-1} S)^{-1} S^{-1} \sum_h^{-1} S$ . If the base forecasts  $\hat{y}_{t+h|t}$  are unbiased, then the reconciled forecasts  $\tilde{y}_{t+h|t}$  will be unbiased, provided that  $SPS = S$  (R. J. Hyndman et al., 2011). This condition is valid for this reconciliation procedure for the *bottom – up*, although not for the *top – down*, methods. Consequently, the *top – down* approaches will never give unbiased reconciled forecasts, even if the base forecasts are unbiased. Additionally, Wickramasuriya et al. (2019) proved that, in general,  $\Sigma_h$  is not known and not identifiable. The covariance matrix of the  $h$  – step ahead reconciled forecast errors is given by the following expression:

$$Var(y_{t+h} - \tilde{y}_{t+h|t}) = SPW_h P' S', \quad (8.9)$$

for any  $P$  such that  $SPS = S$ , then  $W_h = Var(y_{t+h} - \tilde{y}_{t+h|t}) = E(\hat{\varepsilon}_{t+h|t} \hat{\varepsilon}'_{t+h|t})$  is the covariance matrix of the corresponding  $h$  – step ahead base forecast errors. The purpose is to get the matrix  $P$  that minimizes the error variances of the reconciled forecasts which are on the diagonal of the covariance matrix  $Var(y_{t+h} - \tilde{y}_{t+h|t})$ . Finally, Wickramasuriya et al. (2019) demonstrated that the optimal reconciliation matrix  $P$  that minimizes the trace of  $SPW_h P' S'$ , such that  $SPS = S$ , and the optimal reconciled forecasts, respectively, are given by:

$$P = (S' W_h^{-1} S)^{-1} S' W_h^{-1} \quad (8.10)$$

$$\tilde{y}_{t+h|t} = S(S' W_h^{-1} S)^{-1} S' W_h^{-1} \hat{y}_{t+h|t}$$

which is introduced as the *MinT* (minimum trace) estimator. The next step consists of estimating  $W_h$ , a matrix of order  $n$ . Wickramasuriya, Athanasopoulos and Hyndman (2019) proposed the following procedures (Table 18) to obtain the matrix:

### 8.2.4 ARIMA and ETS Formulation

*ARIMA* is one of the most-widely-used time series approaches for forecasting power generation (D. Yang et al., 2018). Although studies have shown that *ETS* outperforms *ARIMA* (Panigrahi & Behera, 2017), it is recommended to keep *ARIMA* as a reference model during the forecasting process. Moreover, several statistical software packages, like *R*, provide automatic model identification and parameter estimation skills for both *ARIMA* and *ETS* (R. J. Hyndman & Khandakar, 2008). Professor Hyndman (R. J. Hyndman et al., 2011) developed the *HTS* package initially based on these predictive models. The present paper aims to test different

Table 18 – Hierarchical forecasting for electricity generation based on the *ARIMA* procedure.

Procedure	Description
OLS	$W_h = k_h I, \forall h$ where $k_h > 0$ . This is the most simplifying premise, and collapses the <i>MinT</i> estimator to the <i>OLS</i> estimator, proposed by Hyndman et al. (2011). This is optimal when the base forecast errors are uncorrelated and equivariant. $W_h = k_h \text{diag}(\hat{W}_1), \forall h$ where $k_w > 0$ and: $\hat{W} = \frac{1}{T} \sum_{t=1}^T \hat{\epsilon}_t(1)\hat{\epsilon}_t(1)'$
WLSv	is the unbiased sample covariance estimator of the in-sample one-step-ahead base forecast errors. In this case, we can describe <i>MinT</i> as a <i>WLS</i> estimator applying variance scaling (Wickramasuriya et al., 2019).
WLSs	$W_h = k_h \Lambda, \forall h$ where $k_h$ and $\Lambda = \text{diag}(S_1)$ with 1 being a unit column vector of dimension $n$ . We assume that each of the <i>bottom-level</i> base forecast errors has a variance $k_h$ and is uncorrelated between nodes. Consequently, every element of the diagonal $\Lambda$ matrix receives the number of forecast error variances contributing to that aggregation level (Wickramasuriya et al., 2019). This estimator depends only on the grouping structure of the hierarchy.
MinT (Sample)	$W_h = k_w \hat{W}_1, \forall h$ where $k_h > 0$ , the unrestricted sample covariance estimator for $h = 1$ (Wickramasuriya et al., 2019). In the results section, we denote this as MinT (Sample).
MinT (Shrink)	$W_h = k_w \hat{W}_{1,D}^*; \forall h; k_n > 0; W_{1,D}^* = \lambda D \hat{W}_{1,D}^* + (1 - \lambda D) \hat{W}_1$ , is a shrinkage estimator with diagonal target, $\hat{W}_{1,D}$ , which is a diagonal matrix comprising the diagonal entries of $\hat{W}_1$ , and $\Lambda_D$ is the shrinkage intensity parameter. Thus, off-diagonal elements of $\hat{W}_1$ are shrunk toward zero and diagonal elements (variances) remain unchanged (Wickramasuriya et al., 2019).

Source: adapted by authors from (Wickramasuriya et al., 2019).

approaches to optimal forecast reconciliation and, to do so, only the *ARIMA* and *ETS* models will be used. It is recommended that future studies extend these forecasting procedures using different predictive models, such as machine learning ones.

*ARIMA* was proposed by (G. Box, 1976). It is a linear forecasting method for dealing with stationary time series. In the initial step, a time series is built stationary by differencing  $d$  times along with some nonlinear transformations, such as logging (Panigrahi & Behera, 2017). The consequential data are recognized as a linear function of past  $p$  data values and  $q$  errors, i.e., modeled as an autoregressive moving average (ARMA) model:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \Theta_1 \epsilon_{t-1} + \Theta_2 \epsilon_{t-2} + \dots + \Theta_q \epsilon_{t-q} \quad (8.11)$$

where  $y_t$  denotes real value at time  $t$ ,  $\varepsilon_t$  describes the error sequence: it is supposed to be white noise and Gaussian distributed  $(0, \sigma^2)$ .  $\phi_i$  for  $(i = 1, 2, \dots, p)$  are autoregressive (AR) coefficients and  $\Theta_j$  for  $(j = 1, 2, \dots, q)$  are moving average (MA) coefficients.  $p$  and  $q$  are integers referred to as model orders. The time series model is denoted as  $ARIMA(p, d, q)$  (G. Box, 1976; Dong, Yang, Reindl, & Walsh, 2013).

The group of exponential smoothing methods utilizes the principle of weighted averages of past information for making forecasts (Panigrahi & Behera, 2017). Since its formulation in 1950, a variety of exponential smoothing methods have been developed. All exponential smoothing methods were initially classified by Pegels (1969), which has been continued by (Gardner Jr, 1985; R. J. Hyndman et al., 2002; Taylor, 2003). *ETS* stands for error, trend, and seasonality elements. As pointed by Panigrahi and Behera (2017), the usual representation for these patterns involves a state vector  $x_t = (l_t, b_t, s_t, s_{t-1}, \dots, s_{t-m+1})'$ , and the state space equations (R. J. Hyndman et al., 2002) have the resulting structure:

$$\begin{aligned} y_t &= w(x_{t-1}) + r(x_{t-1})\varepsilon_t \\ x_t &= f(x_{t-1}) + g(x_{t-1})\varepsilon_t \end{aligned} \quad (8.12)$$

where  $\varepsilon_t$  denotes a Gaussian white noise  $(0, \sigma^2)$  and  $\mu_t = w(x_{t-1})$ . The model with additive error has  $r_t(x_{t-1}) = 1$ , so  $y_t = \mu_t + \varepsilon_t$ . The model with multiplicative errors has  $r_t(x_{t-1}) = \mu_t$ , so  $y_t = \mu_t(1 + \varepsilon_t)$ . Consequently,  $\varepsilon_t = (y_t - \mu_t)/\mu_t$  is a relative error for the multiplicative model and any value of  $r_t(x_{t-1})$  will lead to the identical point forecast for  $y_t$  (Panigrahi & Behera, 2017; R. J. Hyndman et al., 2002).

## 8.2.5 Evaluating Forecast Accuracy

According to Almeida et al. (2016), there are several accuracy metrics, such as mean absolute percentage error (MAPE), mean absolute error (MAE), mean absolute scaled error (MASE), or root-mean-square error (RMSE), to evaluate the performance of point prediction methods, defined as follows:

$$MAPE = \frac{1}{T} \sum_{t=1}^T \left| \frac{y_t - \hat{y}_t}{y_t} \right|. \quad (8.13)$$

$$MAE = \frac{1}{T} \sum_{t=1}^T |y_t - \hat{y}_t|. \quad (8.14)$$

$$MASE = \frac{MAE}{MAE_{insample,naïve}} \quad (8.15)$$

$$RMSE = \sqrt{\frac{1}{T} \sum_{t=1}^T (y_t - \hat{y}_t)^2} \quad (8.16)$$

where  $y_t$  is the amount of power generation at time  $t$ ,  $\hat{y}_t$  is the fitted value for power generation, and  $MAE_{(insample, naive)}$  is the  $MAE$  generated by a *naïve* forecast.

Specifically, in studies of hierarchical time series, the  $MAPE$  indicator appears the most frequently in the literature (Z. Liu, Yan, Yang, & Hauskrecht, 2015; Weiss, 2018; Hong, Xie, & Black, 2019).  $MAPE$  was also the selected metric for the present paper (Figures 29 and 30). Complementarily,  $MAE$ ,  $MASE$ , and  $RMSE$  were estimated, and the results can be found in the appendix. The values of the  $MAPE$ ,  $MAE$ ,  $MASE$  and  $RMSE$  statistics were obtained using a weighted average, with proportions from Table 16.

### 8.3 Results and Discussion

Figure 29, below, shows the predictive result obtained, using the  $ARIMA$  model, considering a predictive window of nine hours ( $h = 1, \dots, 9$ ). Note that the model was estimated, taking the main hierarchical adjustment approaches into account, for the following levels: (i) total power generation in Brazil (Level 0), (ii) total energy generation by electrical subsystem (Level 1), and (iii) total energy generation by the energy generating source (Level 2). For Level 1, four forecasts (one for each electrical subsystem) were estimated. For Level 2, 14 forecasts (one for each energy source) were estimated.

Therefore, we estimated 1539 predictive models satisfying the following proportions: (i) 81 models for Level 0, (ii) 324 models for Level 1, and (iii) 1134 models for Level 2. The  $MAPE$  calculation for Levels 1 and 2 was based on a weighted average of the predictive errors. The weighting factors used are shown in Table 16.

The performance of each predictive model, divided by the forecast horizon, is illustrated by a color scale. The green colors indicate the most accurate forecasts, while the red colors symbolize less accurate forecasts. The best forecasts, for each of the predictive horizons, are highlighted in bold. The last column of Table 1 presents the average performance for each forecast horizon ( $h$ ) for each hierarchical approach.

As pointed by Wickramasuriya et al. (2019), the  $MinT$  procedure has a useful feature: it systematizes results into a unique analytical solution that incorporates information about the correlation structure of the entire dataset. Additionally, the minimum trace reconciliation, with or without regularization, presented the best results of all linear reconciliation methods, such as  $OLS$  and  $WLS$ , with variations. Moreover, the  $MinT$  (Sample) approach returns the most accurate, coherent forecasts for all levels considering just the first forecast horizons. However, as the predictive window grows, the  $BU$  method becomes more accurate. Furthermore, the performance of the  $BU$  model increases as the time series disaggregate.

As expected, the results obtained using the top-down technique did not present good predictive results, since it is intended to generate forecasts for level 0, with worse accuracy for

the other levels. Both *BU* and *TD* present disadvantages: they do not take the correlation among the series at each level into account.

The other accuracy metrics presented in the appendix (MAE, MAE, and RMSE) reinforce the results found. In general, the performance of the optimal reconciliation models, by trace minimization, provides more uniform estimates and better predictive potential for the first hours of the predictive horizon (Figures 40 and 41).



Figure 29 – Hierarchical forecasting for electricity generation based on the *ARIMA* procedure (MAPE). Note: The performance was indicated into a color scale, where green means better values for calculated accuracy, and red means worse accuracy. The intermediate values are colored yellow.

Predictive model: Autoregressive integrated moving average (ARIMA)											
MAPE	Forecast horizon (h)										
	1	2	3	4	5	6	7	8	9	Mean	
<b>Hierarchical level 0 : Total - Brazil</b>											
Method	BU	2.00	3.53	5.62	8.04	10.17	11.45	11.17	10.76	10.48	8.14
	TDGSA	2.07	3.76	6.10	8.79	11.25	12.87	12.95	12.72	12.63	9.24
	TDGSF	2.07	3.76	6.10	8.79	11.25	12.87	12.95	12.72	12.63	9.24
	TDFP	2.07	3.76	6.10	8.79	11.25	12.87	12.95	12.72	12.63	9.24
	OLS	1.98	3.65	5.94	8.59	10.99	12.54	12.55	12.23	12.06	8.95
	WLSv	1.91	3.51	5.70	8.24	10.51	11.93	11.79	11.32	10.99	8.43
	WLSs	1.88	3.46	5.63	8.15	10.39	11.77	11.59	11.09	10.76	8.30
	MintT (Sample)	1.68	3.29	5.50	8.02	10.24	11.57	11.31	10.85	10.60	8.12
	MinT (Shrink)	1.74	3.36	5.59	8.12	10.35	11.69	11.44	10.94	10.69	8.21
	<b>Hierarchical level 1 : Electrical subsystems</b>										
Method	BU	1.97	3.64	6.12	8.75	10.78	11.93	11.90	11.70	11.88	8.74
	TDGSA	31.97	31.74	30.37	28.93	28.12	27.49	26.71	26.04	25.36	28.53
	TDGSF	32.38	32.14	30.71	29.21	28.21	27.46	26.71	26.06	25.41	28.70
	TDFP	1.86	3.88	6.68	9.89	9.89	12.52	14.19	14.45	14.34	9.75
	OLS	1.90	3.55	6.30	9.20	11.70	13.36	13.64	13.56	13.66	9.65
	WLSv	1.77	3.35	5.84	8.62	10.84	12.38	12.56	12.40	12.41	8.91
	WLSs	1.81	3.41	5.92	8.74	11.00	12.57	12.79	12.68	12.75	9.07
	MintT (Sample)	1.64	3.20	5.66	8.50	10.76	12.23	12.40	12.21	12.20	8.76
	MinT (Shrink)	1.66	3.28	5.75	8.57	10.84	12.33	12.50	12.31	12.28	8.83
	<b>Hierarchical level 2 : Energy sources</b>										
Method	BU	2.66	5.05	6.53	7.71	8.88	9.46	9.40	9.22	9.11	7.56
	TDGSA	46.33	44.34	41.72	40.35	39.87	39.29	38.44	37.52	36.58	40.49
	TDGSF	47.66	45.70	42.87	41.24	40.42	39.64	38.80	37.90	36.96	41.24
	TDFP	2.83	5.51	7.53	9.45	9.45	11.46	12.79	13.20	13.33	9.50
	OLS	2.51	5.07	6.78	8.29	9.78	10.62	10.73	10.63	10.56	8.33
	WLSv	2.60	5.11	6.74	8.09	9.42	10.18	10.21	10.07	9.97	8.04
	WLSs	2.56	4.98	6.64	8.00	9.31	10.00	9.95	9.70	9.63	7.86
	MintT (Sample)	2.48	4.96	6.58	7.91	9.27	10.10	10.22	10.10	10.00	7.96
	MinT (Shrink)	2.52	5.04	6.68	8.02	9.38	10.20	10.30	10.19	10.10	8.05

Source: Research results.

In addition to the ARIMA predictive model, Figure 30 presents the same forecasting procedures. However, they are based on the *ETS* automatic adjustment model. The objective is to show the influence of different forecasting methods for each hierarchical reconciliation model. In general, the error percentage produced by the *ETS* model was slightly higher than that produced by the *ARIMA* model. Figure 30 also shows the influence of trace minimization procedures (MinT) on the improvement of predictive performance. In particular, the MinT models have good predictive performance, even with the increase of the forecast horizon hours.

Figure 30 – Hierarchical forecasting for electricity generation based on the *ETS* procedure. Note: The performance was indicated into a color scale, where green means better values for calculated accuracy, and red means worse accuracy. The intermediate values are colored yellow.

Predictive model: Error, trend, seasonality (ETS)											
MAPE	Forecast horizon (h)										
		1	2	3	4	5	6	7	8	9	Mean
<b>Hierarchical level 0 : Total - Brazil</b>											
Method	BU	2.60	4.89	7.73	10.90	13.79	15.78	16.20	16.28	16.48	11.63
	TDGSA	2.11	4.21	6.90	9.95	12.73	14.64	14.98	14.99	15.12	10.62
	TDGSF	2.11	4.21	6.90	9.95	12.73	14.64	14.98	14.99	15.12	10.62
	TDFP	2.11	4.21	6.90	9.95	12.73	14.64	14.98	14.99	15.12	10.62
	OLS	2.13	4.24	6.93	9.98	12.77	14.68	15.03	15.03	15.17	10.66
	WLSv	2.26	4.42	7.16	10.25	13.06	15.00	15.37	15.39	15.55	10.94
	WLSs	2.29	4.46	7.20	10.29	13.11	15.05	15.42	15.45	15.61	10.99
	MintT (Sample)	2.06	4.14	6.80	9.84	12.61	14.51	14.84	14.84	14.96	10.51
	MinT (Shrink)	2.07	4.15	6.82	9.86	12.63	14.53	14.87	14.87	14.99	10.53
	<b>Hierarchical level 1 : Electrical subsystems</b>										
Method	BU	2.59	4.94	8.09	11.43	14.21	16.13	16.71	16.93	17.36	12.04
	TDGSA	31.98	31.78	30.44	29.05	28.29	27.76	27.08	26.51	25.91	28.76
	TDGSF	32.38	32.17	30.78	29.32	28.37	27.72	27.08	26.53	25.96	28.92
	TDFP	2.17	4.37	7.37	10.58	10.58	13.24	15.05	15.54	15.68	10.51
	OLS	2.15	4.34	7.35	10.55	13.22	15.04	15.53	15.68	16.04	11.10
	WLSv	2.30	4.55	7.60	10.85	13.56	15.42	15.94	16.11	16.49	11.42
	WLSs	2.27	4.50	7.54	10.79	13.49	15.34	15.85	16.03	16.40	11.36
	MintT (Sample)	1.89	3.98	6.92	10.09	12.74	14.55	15.03	15.20	15.56	10.66
	MinT (Shrink)	1.94	4.04	7.00	10.17	12.83	14.64	15.13	15.29	15.64	10.74
	<b>Hierarchical level 2 : Energy sources</b>										
Method	BU	3.13	6.56	9.01	11.16	13.24	14.40	14.74	15.21	15.63	11.45
	TDGSA	46.34	44.42	41.84	40.48	40.00	39.45	38.69	37.87	37.00	40.68
	TDGSF	47.66	45.77	42.98	41.35	40.54	39.77	38.99	38.17	37.32	41.40
	TDFP	2.90	6.31	8.74	10.97	10.97	13.05	14.29	14.58	14.94	10.75
	OLS	3.14	6.60	9.07	11.30	13.38	14.62	14.92	15.31	15.74	11.56
	WLSv	2.76	6.08	8.42	10.56	12.53	13.72	13.97	14.18	14.56	10.75
	WLSs	3.13	6.57	9.02	11.21	13.29	14.49	14.79	15.21	15.63	11.48
	MintT (Sample)	2.67	5.88	8.10	10.10	11.99	13.05	13.20	13.47	13.81	10.25
	MinT (Shrink)	2.63	5.89	8.17	10.21	12.14	13.25	13.43	13.72	14.09	10.39

Source: Research results.

The average performance of the trace minimization (MinT) models shows stability, considering all hierarchical levels. As shown in Figure 29, the *ETS*-based predictive model shares some similarities with the *ARIMA* model. The BU technique is better for the most disaggregated levels, whereas the *TD* technique stands out only at the more aggregated levels. Note that the trace minimization procedures show significant gains over the classic linear models, namely *OLS*, and *WLS*.

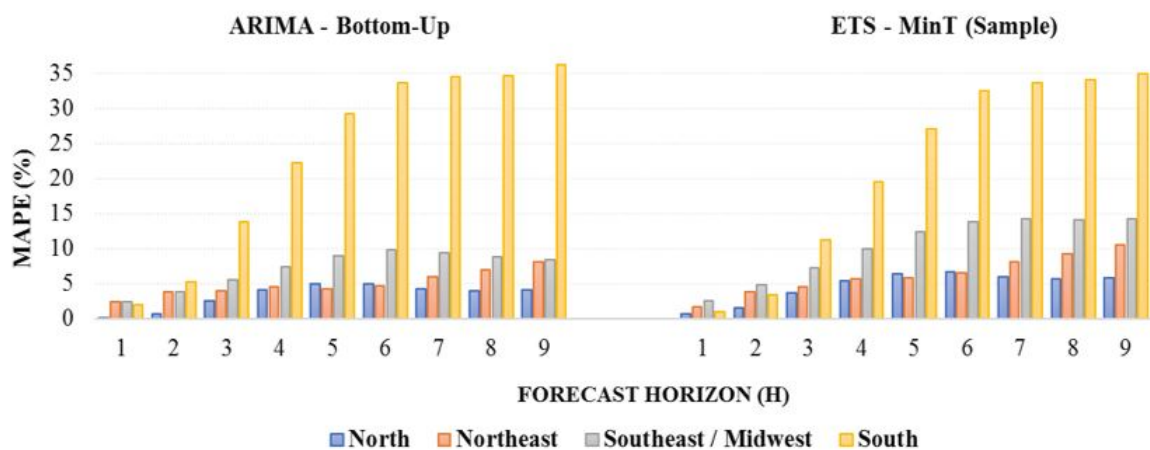
Figures 29 and 30 present some limitations. In general, it is not possible to test the predictive influence of each of the subsystems within the established forecast horizon. To show



this problem, Figure 31 presents a predictive comparison (MAPE) for each of the Brazilian electrical subsystems, considering the nine-hour predictive horizon. On the left is the technique with the best aggregation/disaggregation performance (BU) for the *ARIMA* model. On the right is the technique with the best average performance (MinT) for the *ETS* automatic selection model.

Figure 31 thus shows a negative influence of the "south" electrical subsystem in the global measures of accuracy, especially from a predictive horizon of three hours onward. This system should be analyzed more thoroughly to identify energy sources located in the "south" subsystem that contributed most to the predictive instability of this system. Simultaneously, the use of individualized predictive models for this "south" system can be a good strategy, since unique climatic conditions exist in southern Brazil.

Figure 31 – Hierarchical forecasting for power generation: electrical subsystem versus forecast horizon.



Source: Research results.

Figures 38 and 39 (Appendix) present the accuracy measure of the *ARIMA* and *ETS* models in detail, considering energy sources versus electrical subsystems. These results reinforce those in Figure 31, indicating instability in the southern subsystem, especially wind energy data.

Finally, some limitations of the present paper are recognized here. First, predictive models are based on past information evaluable, so the presented results cannot be extrapolated for different contexts and other time periods. Additionally, it is necessary to incorporate other predictive models to make the results more robust. In future research, it is recommended that models which integrate high-frequency data, e.g., the Wavelet approach, be adopted.

## 8.4 Conclusions

Analysis of the energy market is complicated. It involves the relationship between forecasting models and uncertainty, distinctly regarding the stochastic behavior of variables. The present paper is aimed at policymakers, offering a forecasting tool that deals with grouped time

series. It also proposes a new forecasting approach, based on hierarchical modeling of the energy generation in Brazil.

The present paper introduces the use of trace minimization procedures (MinT) to aggregate and disaggregate forecasts based on the *ARIMA* and *ETS* models. *MinT* models performed better than the classic linear approaches, such as *OLS* and *WLS*. The *MinT* models also have high reliability for short predictive horizons. It is noteworthy that both hierarchical procedures and forecasting methods influence the predictive values of power generation in Brazil.

Therefore, the use of other predictive models, such as those based on analogs, machine learning, and other hybrid techniques, for example, is recommended. For future research, fine-tuning forecasts of the “south” electrical subsystem, as well as testing the accuracy of the hierarchical methods by using new forecasting approaches, is also recommended.

Finally, the present study contributes to the energy planning processes of different agents, given that understanding energy generation patterns is singularly important for minimizing risks and supporting reliable production planning. Good forecasts for future energy generation can support operational arrangements since energy supply and demand impact on spot market sales prices.

## 9 Conclusions

In this thesis, we present seven works developed and published in the central theme of the research: electricity price forecasting. The works are interconnected and presented sequentially to their development. We found that changes in the energy trading system in Brazil, especially the adoption of hourly prices, fostered the need to build new predictive models. This way, this thesis innovated by bringing the theme of analogs and the dynamic time scan forecasting methodology to electricity price forecasting systems.

Here we highlight relevant conclusions from each chapter.

In a pioneering way, chapter three presents a systematic literature review on electricity price forecasting (EPF). This chapter applied research standards derived from the medical field, such as the PRISMA methodology, and used a system developed in *R* to unify Scopus and Web of Science metadata, automatically removing their duplicates. This chapter showed that research on *EPF* has grown substantially and that hybrid models have stood out for this purpose. The methodology developed in this chapter served as the basis for Chapter 4.

Still on the literature review, in chapter four, we verified the exponential increase of publications on applications of predictive maintenance techniques in the hydroelectric sector. The application of machine learning models has been widely advocated by researchers in the reported case studies. In particular, deep learning techniques, given their efficiency in developing models of high accuracy and generalization capacity, indicate a huge opportunity for advancement in this kind of predictive system. With the development of cloud platforms and their respective computational power, these algorithms can deal with gigantic databases, such as those found in the monitoring health of machines context. These recent advances, in harmony with new machine understanding techniques, indicate a future in which the interaction between specialists and intelligent systems will be closer and more dynamic.

In chapter five, we discuss the theme of searching for similarities in time series. According to our methodology, we reduced the search time of the similarity profile of the analogs by up to 17.5%. This result was relevant due to guaranteeing the competitiveness of the method with other approaches, such as the univariate ones.

In chapter six, we compare the accuracy of the *DTSF* method with nine classic statistical methods and with the baseline of the *M4* competition. The *DTSF* method proved to be competitive, especially in long time series and with high repeatability as hourly frequencies.

In chapters seven and eight, the *DTSF* methodology was used for the first time in EPF applications, considering price data with weekly and hourly frequency, respectively. The results were promising and indicated that the *DTSF* could be a relevant methodology for energy planning.

The DTSTF forecasts were confronted with those obtained from classic statistics and machine learning models and indicated that there are ideal models for seasons of the year and times of the day.

Finally, in Chapter nine, an experimental study was conducted. We used the hierarchical time series (HTS) methodology to deal with power generation in Brazil. Despite being a promising forecasting tool, the prediction calibration of those methodologies did not perform well in the *M5* competition. This way, the investigations started in chapter 9 were not deepened.

## References

- Abolhassani, A., & Prates, M. O. (2021). An up-to-date review of scan statistics. *Statistics Surveys*, *15*, 111–153. Cited in page 83.
- Abouarghoub, W., Nomikos, N. K., & Petropoulos, F. (2018). On reconciling macro and micro energy transport forecasts for strategic decision making in the tanker industry. *Transportation Research Part E: Logistics and Transportation Review*, *113*, 225–238. Cited in page 93.
- Abreu, T. M., de Souza, A. Z., & Ribeiro, P. F. (2020). Economic analysis of an energy storage system in the context of hourly electricity spot price in brazil. In *2020 IEEE Power & Energy Society General Meeting (PESGM)* (pp. 1–5). Cited in page 81.
- Aigner, W., Miksch, S., Müller, W., Schumann, H., & Tominski, C. (2007). Visual methods for analyzing time-oriented data. *IEEE transactions on visualization and computer graphics*, *14*(1), 47–60. Cited in page 42.
- Al-Alawi, S. M., & Islam, S. M. (1996). Principles of electricity demand forecasting. i. methodologies. *Power Engineering Journal*, *10*(3), 139–143. Cited 2 times in pages 60 and 72.
- Alamaniotis, M., Bargiotas, D., Bourbakis, N. G., & Tsoukalas, L. H. (2015). Genetic optimal regression of relevance vector machines for electricity pricing signal forecasting in smart grids. *IEEE transactions on smart grid*, *6*(6), 2997–3005. Cited in page 35.
- Alcácer, V., & Cruz-Machado, V. (2019). Scanning the industry 4.0: A literature review on technologies for manufacturing systems. *Engineering Science and Technology, an International Journal*, *22*(3), 899–919. Cited in page 92.
- Almeida, V., Ribeiro, R., & Gama, J. (2016). Hierarchical time series forecast in electrical grids. In *Information science and applications (icisa) 2016* (pp. 995–1005). Springer. Cited 2 times in pages 93 and 100.
- Analytics, C. (2017). *Web of science core collection. citation database. web of science*. Retrieved from <https://www.webofknowledge.com/WOS>. Cited in page 69.
- Aneel. (2013). *Normative resolution no. 583/2013. establishes the procedures and conditions for obtaining and maintaining the situation operational and definition of installed power and net of generating enterprise electricity*. Brasília: National Electric Energy Agency. Cited in page 19.
- Antonopoulos, I., Robu, V., Couraud, B., Kirli, D., Norbu, S., Kiprakis, A., . . . Wattam, S. (2020). Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review. *Renewable and Sustainable Energy Reviews*, *130*, 109899. Cited in page 28.
- Aria, M., & Cuccurullo, C. (2017). bibliometrix: An r-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. Cited 2 times in pages 31 and 70.
- Arroyo, J. M., & Conejo, A. J. (2000). Optimal response of a thermal unit to an electricity spot

- market. *IEEE Transactions on power systems*, 15(3), 1098–1104. Cited in page 73.
- Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International journal of forecasting*, 16(4), 521–530. Cited 2 times in pages 59 and 85.
- Assuncao, H., Escobedo, J., & Oliveira, A. (2003). Modelling frequency distributions of 5 minute-averaged solar radiation indexes using beta probability functions. *Theoretical and Applied Climatology*, 75(3-4), 213–224. Cited in page 42.
- Athanasopoulos, G., Ahmed, R. A., & Hyndman, R. J. (2009). Hierarchical forecasts for australian domestic tourism. *International Journal of Forecasting*, 25(1), 146–166. Cited 2 times in pages 92 and 97.
- Azadeh, A., Ghaderi, S., & Sohrabkhani, S. (2008). Annual electricity consumption forecasting by neural network in high energy consuming industrial sectors. *Energy Conversion and management*, 49(8), 2272–2278. Cited 3 times in pages 60, 72, and 85.
- Bandyopadhyay, A., Roy, S., & Ghosh, D. (2013). Forecasting day-ahead price of electricity—a dynamic regression approach. *International Journal of Business Excellence*, 6(5), 584–604. Cited in page 33.
- Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901), 268–282. Cited in page 75.
- Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120, 70–83. Cited in page 66.
- Bernardini, E., & Cubadda, G. (2015). Macroeconomic forecasting and structural analysis through regularized reduced-rank regression. *International Journal of Forecasting*, 31(3), 682–691. Cited in page 53.
- Bhattacharyya, S. C. (2019). Markets for electricity supply. In *Energy economics* (pp. 699–733). Springer. Cited 2 times in pages 19 and 80.
- Billah, B., Hyndman, R. J., & Koehler, A. B. (2005). Empirical information criteria for time series forecasting model selection. *Journal of Statistical Computation and Simulation*, 75(10), 831–840. Cited in page 53.
- Bisht, D. C., & Ram, M. (2021). Recent advances in time series forecasting. Cited in page 84.
- Böckler, L., & Pereira, M. G. (2019). Consumer (co-) ownership in renewables in brazil. In *Energy transition* (pp. 535–557). Springer. Cited in page 75.
- Bontempi, G. (2020). Comments on m4 competition. *International Journal of Forecasting*, 36(1), 201–202. Cited in page 56.
- Botterud, A., Kristiansen, T., & Ilic, M. D. (2010). The relationship between spot and futures prices in the nord pool electricity market. *Energy Economics*, 32(5), 967–978. Cited in page 78.

- Box, G. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791–799. Cited 4 times in pages 41, 68, 99, and 100.
- Box, G. E., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association*, 65(332), 1509–1526. Cited in page 59.
- Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering*, 137, 85–86. Cited in page 73.
- Brazil. (2004). *Decree no. 5177/2004. regulates articles 4th and 5th of law no. 10,848, of march 15, 2004, and provides for the organization, attributions and functioning of the electric energy commercialization chamber - ccee*. Brasília: Presidency of the Republic. Cited 2 times in pages 19 and 81.
- Brazil. (2019). *Ordinance no 300/2019. approves, under the terms of this ordinance, the improvements proposed by the methodology working group of the standing committee for analysis of methodologies and computational programs in the electricity sector*. Brasília: Ministry of Mines and Energy/Office of the Minister. Cited in page 19.
- Bui, D., Tuan, T., Klempe, H., Pradhan, B., & Revhaug, I. (2016). Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides*, 13(2), 361–378. Cited 4 times in pages 28, 42, 67, and 68.
- Burrus, C., & Parks, T. (1985). *and convolution algorithms*. New York: John Wiley and Sons. Cited in page 48.
- Cai, W., Chen, J., Hong, J., & Jiang, F. (2017). Forecasting chinese stock market volatility with economic variables. *Emerging Markets Finance and Trade*, 53(3), 521–533. Cited in page 53.
- Campos, M. S., Costa, M. A., Gontijo, T. S., & Lopes-Ahn, A. L. (2022). Robust stochastic frontier analysis applied to the brazilian electricity distribution benchmarking method. *Decision Analytics Journal*, 3, 100051. Cited in page 151.
- Capeletti, M. (2019). Análise da implementação do preço de liquidação das diferenças (pld) horário no brasil e a relação com variáveis de entrada do modelo de cálculo no horizonte de curto prazo. Cited in page 20.
- Catarino, F. R. S., Santos, M., Gontijo, T. S., & Rodrigues, A. (2017). Gestão de estoque em uma microempresa do ramo alimentício: Comparação entre a curva abc e o método xyz. *Revista Caribeña de Ciencias Sociales*. ISSN, 2254–7630. Cited in page 152.
- Cei. (2023). *Overview*. Retrieved from <https://ceienergetica.gupy.io/> Cited in page 21.
- Chang, Z., Zhang, Y., & Chen, W. (2019). Electricity price prediction based on hybrid model of adam optimized lstm neural network and wavelet transform. *Energy*, 187, 115804. Cited in page 33.

- Chatfield, C. (2000). *Time-series forecasting*. Chapman and Hall/CRC. Cited in page 59.
- Chen, H., Wan, Q., Li, F., & Wang, Y. (2013). Garch in mean type models for wind power forecasting. In *2013 IEEE Power & Energy Society General Meeting* (pp. 1–5). Cited in page 92.
- Chen, J., & Glaz, J. (2012). Approximations for the distribution and the. *Scan Statistics and Applications*, 27. Cited in page 71.
- Chen, X., Dong, Z. Y., Meng, K., Xu, Y., Wong, K. P., & Ngan, H. (2012). Electricity price forecasting with extreme learning machine and bootstrapping. *IEEE Transactions on Power Systems*, 27(4), 2055–2062. Cited in page 68.
- Chen, Y., He, Z., Shang, Z., Li, C., Li, L., & Xu, M. (2019). A novel combined model based on echo state network for multi-step ahead wind speed forecasting: A case study of nrel. *Energy conversion and management*, 179, 13–29. Cited 3 times in pages 28, 42, and 68.
- Chevalier, J.-M. (2009). The new energy crisis. In *The new energy crisis* (pp. 6–59). Springer. Cited in page 86.
- Choi, Y. (1993). *Paradigms and conventions: Uncertainty, decision making, and entrepreneurship*. University of Michigan Press. Cited 2 times in pages 41 and 67.
- Constantopoulos, P., Schweppe, F. C., & Larson, R. C. (1991). Estia: A real-time consumer control scheme for space conditioning usage under spot electricity pricing. *Computers & operations research*, 18(8), 751–765. Cited in page 33.
- Costa, M. A., Ruiz-Cárdenas, R., Mineti, L. B., & Prates, M. O. (2021). Dynamic time scan forecasting for multi-step wind speed prediction. *Renewable Energy*, 177, 584–595. Cited 10 times in pages 42, 54, 56, 57, 58, 60, 68, 71, 83, and 84.
- Da Silva, A. V., Costa, M. A., Ahn, H., & Lopes, A. L. M. (2019). Performance benchmarking models for electricity transmission regulation: Caveats concerning the Brazilian case. *Utilities Policy*, 60, 100960. Cited in page 80.
- da Silva, D. J. A., de Matos, L. F., & Gontijo, T. S. (2015). A utilização de ferramentas da qualidade em uma empresa de manutenção de equipamentos eletromédicos. *Revista Petra*, 1(2). Cited in page 152.
- de Azevedo, A. A., & Gontijo, T. S. (2017). Habilidades, competências e o perfil do profissional de engenharia de produção no sudeste brasileiro. *Formação@ Docente*, 9(2), 96–109. Cited in page 152.
- de Azevedo, A. A., Gontijo, T. S., Victor, E. F., de Souza, L. L., & de Oliveira, T. J. (2018). Um estudo sobre as causas que geram a indisponibilidade no processo de fabricação de peças automotivas. *ForScience*, 6(3). Cited in page 152.
- de Barros, R. A., Teixeira, F. S., & Gontijo, T. S. (2018). Estudo de caso em uma trefilaria: proposta de redução da perda de maior representatividade. *Sistemas & Gestão*, 13(1), 88–96. Cited in page 152.
- de Cássio Rodrigues, A., de Azevedo, A. A., Gontijo, T. S., de Oliveira, A. A., Ferreira, H. K. G., Aquino, L. M. S., ... de Souza, M. M. L. (2018). Aplicativos educacionais na engenharia



- de produção: O caso do enade nota 10. *Brazilian Journal of Production Engineering-BJPE*, 21–30. Cited in page 152.
- de Cássio Rodrigues, A., De Muylder, C. F., & Gontijo, T. S. (2018). Eficiência das unidades do cefet-mg: uma avaliação por data envelopment analysis. *ForScience*, 6(3). Cited in page 152.
- de Cássio Rodrigues, A., Gonçalves, C. A., & Gontijo, T. S. (2019). A two-stage dea model to evaluate the efficiency of countries at the rio 2016 olympic games”. *Economics Bulletin*, 39(2), 1538–1545. Cited in page 152.
- de Cássio Rodrigues, A., & Gontijo, T. S. (2019). Incorporando julgamentos de especialistas em educação na avaliação da eficiência de cursos de graduação: uma abordagem por data envelopment analysis. *Revista Gestão & Tecnologia*, 19(1), 113–139. Cited in page 152.
- de Cássio Rodrigues, A., Gontijo, T. S., & de Almeida, G. C. D. (2020). O valor do projeto de uma mina de ouro: Uma análise comparativa pelos modelos de fluxo de caixa descontado e de opções reais. *South American Development Society Journal*, 5(15), 122. Cited in page 151.
- de Cássio Rodrigues, A., Gontijo, T. S., & De Muylder, C. F. (2019). Measuring the technical and scale efficiency of rio de janeiro samba schools: a dea approach. *Exacta*, 17(4), 201–210. Cited in page 152.
- de Cássio Rodrigues, A., Gontijo, T. S., Gonçalves, C. A., & Pereira, T. H. M. (2022). Efeitos da incorporação de julgamentos na avaliação da eficiência de clubes de futebol: uma abordagem por data envelopment analysis. *Exacta*, 20(1), 234–251. Cited in page 151.
- de Marcos, R. A., Bello, A., & Reneses, J. (2019). Short-term electricity price forecasting with a composite fundamental-econometric hybrid methodology. *Energies*, 12(6), 1067. Cited in page 34.
- de Santis, R. B., Gontijo, T. S., & Costa, M. A. (2022a). Condition-based maintenance in hydroelectric plants: A systematic literature review. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 236(5), 631–646. Cited in page 151.
- de Santis, R. B., Gontijo, T. S., & Costa, M. A. (2022b). A data-driven framework for small hydroelectric plant prognosis using tsfresh and machine learning survival models. *Sensors*, 23(1), 12. Cited in page 151.
- De Santis, R. B., Gontijo, T. S., & Costa, M. A. (2023). Dynamic time scan forecasting: A benchmark with m4 competition data. *IEEE Latin America Transactions*, 21(2), 320–327. Cited in page 151.
- De Vany, A. S., & Walls, W. D. (1999). Cointegration analysis of spot electricity prices: insights on transmission efficiency in the western us. *Energy Economics*, 21(5), 435–448. Cited in page 72.
- Dong, Z., Yang, D., Reindl, T., & Walsh, W. M. (2013). Short-term solar irradiance forecasting using exponential smoothing state space model. *Energy*, 55, 1104–1113. Cited in page

100.

- Duarte Filho, G. D., de Souza, R. d. S., & Gontijo, T. S. (2016). A eficiência no processo de impressão em uma indústria de embalagens plásticas da região metropolitana de belo horizonte. *Revista Petra*, 2(1). Cited in page 152.
- Dudek, G. (2016). Pattern-based local linear regression models for short-term load forecasting. *Electric Power Systems Research*, 130, 139–147. Cited in page 53.
- Dutra, N. F., De Souza, F. R., Gontijo, T. S., de Cássio Rodrigues, A., & de Matos Andrade, I. C. (2018). O impacto da política nacional de resíduos sólidos nas publicações científicas sobre logística reversa. *Brazilian Journal of Production Engineering-BJPE*, 66–82. Cited in page 152.
- Ebert, P., & Sperandio, M. (2018). Influence of integration of wind power in planning the operation of a hydrothermal system using dynamic systems. *IEEE Latin America Transactions*, 16(5), 1432–1438. Cited 2 times in pages 19 and 81.
- Eckel, F., & Delle Monache, L. (2016). A hybrid nwp–analog ensemble. *Monthly Weather Review*, 144(3), 897–911. Cited in page 42.
- Figueiredo, N. C., da Silva, P. P., & Cerqueira, P. A. (2015). Evaluating the market splitting determinants: evidence from the iberian spot electricity prices. *Energy Policy*, 85, 218–234. Cited in page 75.
- Figueiredo, N. C., da Silva, P. P., & Cerqueira, P. A. (2016). It is windy in denmark: Does market integration suffer? *Energy*, 115, 1385–1399. Cited in page 74.
- Fildes, R., & Makridakis, S. (1995). The impact of empirical accuracy studies on time series analysis and forecasting. *International Statistical Review/Revue Internationale de Statistique*, 289–308. Cited in page 56.
- Flores, M. P. P., Solís, J. F., Valdez, G. C., Barbosa, J. J. G., Ortega, J. P., & Villanueva, J. D. T. (2019). Hurst exponent with arima and simple exponential smoothing for measuring persistency of m3-competition series. *IEEE Latin America Transactions*, 17(05), 815–822. Cited in page 54.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232. Cited in page 85.
- Gardner, E. S., & McKenzie, E. (2011). Why the damped trend works. *Journal of the Operational Research Society*, 62(6), 1177–1180. Cited in page 59.
- Gardner Jr, E. S. (1985). Exponential smoothing: The state of the art. *Journal of forecasting*, 4(1), 1–28. Cited in page 100.
- Geisser, S. (2017). *Predictive inference*. Routledge. Cited 2 times in pages 41 and 67.
- Gellert, A., Florea, A., Fiore, U., Palmieri, F., & Zanetti, P. (2019). A study on forecasting electricity production and consumption in smart cities and factories. *International Journal of Information Management*, 49, 546–556. Cited in page 33.
- Gensler, A., Sick, B., & Pankraz, V. (2016). An analog ensemble-based similarity search technique for solar power forecasting. In *2016 IEEE International Conference on Systems*,

- man, and cybernetics (smc* (p. 002850–002857). IEEE. Cited in page [42](#).
- Ghayekhloo, M., Azimi, R., Ghofrani, M., Menhaj, M., & Shekari, E. (2019). A combination approach based on a novel data clustering method and bayesian recurrent neural network for day-ahead price forecasting of electricity markets. *Electric Power Systems Research*, *168*, 184–199. Cited in page [35](#).
- Girish, G., Tiwari, A. K., et al. (2016). A comparison of different univariate forecasting models for spot electricity price in india. *Economics Bulletin*, *36*(2), 1039–1057. Cited in page [53](#).
- Glaz, J., & Balakrishnan, N. (2012). *Scan statistics and applications*. Springer Science & Business Media. Cited in page [56](#).
- Gomes, R., & Poltronieri, R. (2018). The electricity sector and the structure of the short-term market in brazil. In *Energy law and regulation in brazil* (pp. 113–135). Springer. Cited in page [81](#).
- Gomes, R. V. (2015). Modelo de opções reais para avaliar a estratégia de produção em uma indústria eletrointensiva face ao preço da energia elétrica.  
Cited 3 times in pages [27](#), [75](#), and [77](#).
- Gontijo, T. S., & Alves, F. A. M. (2019). A bibliometric study on industry 4.0. *International Journal of Professional Business Review: Int. J. Prof. Bus. Rev.*, *4*(2), 35–42. Cited in page [152](#).
- Gontijo, T. S., Costa, M. A., & de Santis, R. B. (2020). Similarity search in electricity prices: An ultra-fast method for finding analogs. *Journal of Renewable and Sustainable Energy*, *12*(5), 056103. Cited 3 times in pages [54](#), [59](#), and [151](#).
- Gontijo, T. S., Costa, M. A., & de Santis, R. B. (2021). Electricity price forecasting on electricity spot market: a case study based on the brazilian difference settlement price. In *E3s web of conferences* (Vol. 239, p. 00002). Cited 4 times in pages [54](#), [82](#), [83](#), and [151](#).
- Gontijo, T. S., de Cássio Rodrigues, A., De Muyllder, C. F., Falce, J. L. 1., & Pereira, T. H. M. (2020). Analysis of olive oil market volatility using the arch and garch techniques. *International Journal of Energy Economics and Policy*, *10*(3), 423–428. Cited in page [152](#).
- Gontijo, T. S., De Muyllder, C. F., & de Cássio Rodrigues, A. (2018). Incorporating managed preferences in the evaluation of public organizations efficiency: a dea approach. *Independent Journal of Management & Production*, *9*(4), 1108–1126. Cited in page [152](#).
- Gontijo, T. S., dos Santos, P. M., Franco, T. A. B., & de Azevedo, A. A. (2017). Um estudo de caso sobre o impacto das restrições médicas nos custos ergonômicos escolares de um município. *Revista Produção Online*, *17*(3), 909–930. Cited in page [152](#).
- Gontijo, T. S., Fernandes, E. A., & Saraiva, M. B. (2011). Análise da volatilidade do retorno da commodity dendê: 1980-2008. *Revista de Economia e Sociologia Rural*, *49*, 857–874. Cited in page [152](#).
- Gontijo, T. S., & Reis, I. A. (2021). Os determinantes da eficiência na atenção primária à saúde dos municípios paulistas: um modelo georreferenciado. *Physis: Revista de Saúde Coletiva*,

31. Cited in page 151.
- Gontijo, T. S., Rodrigues, F. D. M., de Cássio Rodrigues, A., da Silva, S. A., & de Azevedo, A. A. (2017). Consumo industrial de energia elétrica: um estudo comparativo entre métodos preditivos. *Brazilian Journal of Production Engineering-BJPE*, 31–45. Cited in page 152.
- Green, R. (1999). Draining the pool: the reform of electricity trading in england and wales. *Energy Policy*, 27(9), 515–525. Cited in page 72.
- Green, R. J., & Newbery, D. M. (1992). Competition in the british electricity spot market. *Journal of political economy*, 100(5), 929–953. Cited 2 times in pages 72 and 73.
- Gross, C. W., & Sohl, J. E. (1990). Disaggregation methods to expedite product line forecasting. *Journal of forecasting*, 9(3), 233–254. Cited in page 97.
- Guimarães, C., & Piefer, N. (2017). Brazil:(future) green energy power and strategic partner for the eu? *Comparative European Politics*, 15(1), 23–44. Cited in page 77.
- Hamm, G., & Borison, A. (2006). Forecasting long-run electricity prices. *The electricity journal*, 19(7), 47–57. Cited in page 67.
- Hand, D. J. (2009). *Forecasting with exponential smoothing: The state space approach by rob j. hyndman, anne b. koehler, j. keith ord, ralph d. snyder*. Wiley Online Library. Cited in page 85.
- Hannah Jessie Rani, R., & Aruldoss Albert Victoire, T. (2019). A hybrid elman recurrent neural network, group search optimization, and refined vmd-based framework for multi-step ahead electricity price forecasting. *Soft Computing*, 23(18), 8413–8434. Cited in page 34.
- Harris, P., Bitonti, A., Fleisher, C. S., & Binderkrantz, A. S. (2022). *The palgrave encyclopedia of interest groups, lobbying and public affairs*. Springer Nature. Cited in page 82.
- Hassani, H., & Silva, E. S. (2018). Forecasting uk consumer price inflation using inflation forecasts. *Research in Economics*, 72(3), 367–378. Cited in page 53.
- Haugom, E., Molnár, P., & Tysdahl, M. (2020). Determinants of the forward premium in the nord pool electricity market. *Energies*, 13(5), 1111. Cited 2 times in pages 20 and 81.
- Heck, N., Smith, C., & Hittinger, E. (2016). A monte carlo approach to integrating uncertainty into the levelized cost of electricity. *The Electricity Journal*, 29(3), 21–30. Cited in page 68.
- Hill, T., Marquez, L., O’Connor, M., & Remus, W. (1994). Artificial neural network models for forecasting and decision making. *International journal of forecasting*, 10(1), 5–15. Cited in page 53.
- Hoeltgebaum, L. E. B., Dias, N. L., & Costa, M. A. (2021). An analog period method for gap-filling of latent heat flux measurements. *Hydrological Processes*, 35(4), e14105. Cited in page 54.
- Hong, T., Xie, J., & Black, J. (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*, 35(4), 1389–1399. Cited in page 101.
- Hubicka, K., Marcjasz, G., & Weron, R. (2018). A note on averaging day-ahead electricity price

- forecasts across calibration windows. *IEEE Transactions on sustainable energy*, 10(1), 321–323. Cited in page 35.
- Hunt, J. D., Stilpen, D., & de Freitas, M. A. V. (2018). A review of the causes, impacts and solutions for electricity supply crises in brazil. *Renewable and Sustainable Energy Reviews*, 88, 208–222. Cited in page 86.
- Hussain, H. M., Javaid, N., Iqbal, S., Hasan, Q. U., Aurangzeb, K., & Alhussein, M. (2018). An efficient demand side management system with a new optimized home energy management controller in smart grid. *Energies*, 11(1), 190. Cited in page 81.
- Hyndman, R., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media. Cited 2 times in pages 59 and 85.
- Hyndman, R. J. (2020). A brief history of forecasting competitions. *International Journal of Forecasting*, 36(1), 7–14. Cited 2 times in pages 71 and 76.
- Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational statistics & data analysis*, 55(9), 2579–2589. Cited 5 times in pages 93, 95, 97, 98, and 99.
- Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for r. *Journal of statistical software*, 27, 1–22. Cited 4 times in pages 82, 84, 96, and 98.
- Hyndman, R. J., & Koehler, A. B. (2006). Effect of question formats on item endorsement rates in web surveys. *International Journal of Forecasting*, 22(4), 679–688. Cited in page 89.
- Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of forecasting*, 18(3), 439–454. Cited 3 times in pages 59, 85, and 100.
- Höppner, F., & Klawonn, F. (2009). Compensation of translational displacement in time series clustering using cross correlation. In *International symposium on intelligent data analysis* (p. 71–82). Berlin, Heidelberg: Springer. Cited in page 44.
- Ilea, V., Bovo, C., et al. (2017). Impact of the price coupling of regions project on the day-ahead electricity market in italy. In *2017 ieee manchester powertech* (pp. 1–6). Cited 2 times in pages 20 and 81.
- Immink, K., & Weber, J. (2015). Hybrid minimum pearson and euclidean distance detection. *IEEE Transactions on Communications*, 63(9), 3290–3298. Cited in page 43.
- Ioakimidis, C. S., Oliveira, L. J., & Genikomsakis, K. N. (2014). Wind power forecasting in a residential location as part of the energy box management decision tool. *IEEE Transactions on Industrial Informatics*, 10(4), 2103–2111. Cited in page 68.
- Islam, S., & Al-Alawi, S. (1996). Principles of electricity demand forecasting. i. methodologies. *IEEE Power Engineering Journal*, 10(3), 139–143. Cited in page 85.
- Iwabuchi, K., Kato, K., Watari, D., Taniguchi, I., Catthoor, F., Shirazi, E., & Onoye, T. (2022). Flexible electricity price forecasting by switching mother wavelets based on wavelet



- transform and long short-term memory. *Energy and AI*, 10, 100192. Cited in page 28.
- Jain, L., Seera, M., Lim, C., & Balasubramaniam, P. (2014). A review of online learning in supervised neural networks. *Neural computing and applications*, 25(3-4), 491–509. Cited 3 times in pages 28, 42, and 68.
- Janke, L., McDonagh, S., Weinrich, S., Murphy, J., Nilsson, D., Hansson, P., & Nordberg, (2020). Optimizing power-to-h2 participation in the nord pool electricity market: Effects of different bidding strategies on plant operation. *Renewable Energy*. Cited in page 48.
- Jaradat, M., Jarrah, M., Bousseham, A., Jararweh, Y., & Al-Ayyoub, M. (2015). The internet of energy: smart sensor networks and big data management for smart grid. *Procedia Computer Science*, 56, 592–597. Cited in page 42.
- Jin, C. H., Pok, G., Paik, I., & Ryu, K. H. (2015). Short-term electricity load and price forecasting based on clustering and next symbol prediction. *IEEE Transactions on Electrical and Electronic Engineering*, 10(2), 175–180. Cited in page 35.
- Johannesson, J., & Clowes, D. (2022). Energy resources and markets—perspectives on the russia–ukraine war. *European Review*, 30(1), 4–23. Cited in page 86.
- Jordehi, A. R. (2019). Optimisation of demand response in electric power systems, a review. *Renewable and sustainable energy reviews*, 103, 308–319. Cited 2 times in pages 20 and 81.
- Kalavani, F., Mohammadi-Ivatloo, B., & Zare, K. (2019). Optimal stochastic scheduling of cryogenic energy storage with wind power in the presence of a demand response program. *Renewable Energy*, 130, 268–280. Cited 2 times in pages 20 and 81.
- Kelley, T. (1925). Measures of correlation determined from groups of varying homogeneity. *Journal of the American Statistical Association*, 20(152), 512–521. Cited in page 44.
- Khan Jaffur, Z. R., Sookia, N.-U.-H., Nunkoo Gonpot, P., & Seetanah, B. (2017). Out-of-sample forecasting of the canadian unemployment rates using univariate models. *Applied Economics Letters*, 24(15), 1097–1101. Cited in page 53.
- Khosravi, A., Nahavandi, S., Creighton, D., & Naghvizadeh, R. (2012). Uncertainty quantification for wind farm power generation. In *The 2012 international joint conference on neural networks (ijcnn)* (pp. 1–6). Cited in page 68.
- Kuhawar, U. Z., Siddiqui, I. F., Arain, Q. A., Siddiqui, M. M., & Qureshi, N. M. F. (2021). On-ground distributed covid-19 variant intelligent data analytics for a regional territory. *Wireless Communications and Mobile Computing*, 2021. Cited in page 85.
- Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, 32(3), 669–679. Cited in page 89.
- Klemperer, P. (2002). What really matters in auction design. *Journal of economic perspectives*, 16(1), 169–189. Cited in page 73.
- Kosiorowski, D., Mielczarek, D., Rydlewski, J., et al. (2017). Forecasting of a hierarchical functional time series on example of macromodel for day and night air pollution in silesia region: a critical overview. *arXiv preprint arXiv:1712.03797*. Cited in page 93.

- Koziel, S., & Bandler, J. W. (2008). Modeling of microwave devices with space mapping and radial basis functions. *International journal of numerical modelling: electronic networks, devices and fields*, 21(3), 187–203. Cited in page 85.
- Kuhn, M., et al. (2008). Building predictive models in r using the caret package. *Journal of statistical software*, 28(5), 1–26. Cited in page 67.
- Kulldorff, M. (1999). Spatial scan statistics: models, calculations, and applications. In *Scan statistics and applications* (pp. 303–322). Springer. Cited in page 83.
- Lammers, I., & Hoppe, T. (2018). Analysing the institutional setting of local renewable energy planning and implementation in the eu: A systematic literature review. *Sustainability*, 10(9), 3212. Cited in page 81.
- Li, R., Woo, C.-K., & Cox, K. (2021). How price-responsive is residential retail electricity demand in the us? *Energy*, 232, 120921. Cited in page 28.
- Lin, B., Omoju, O. E., & Okonkwo, J. U. (2015). Will disruptions in opec oil supply have permanent impact on the global oil market? *Renewable and Sustainable Energy Reviews*, 52, 1312–1321. Cited in page 82.
- Liu, H., & Shi, J. (2013). Applying arma–garch approaches to forecasting short-term electricity prices. *Energy Economics*, 37, 152–166. Cited 3 times in pages 28, 42, and 68.
- Liu, H., Tian, H., Liang, X., & Li, Y. (2015). Wind speed forecasting approach using secondary decomposition algorithm and elman neural networks. *Applied Energy*, 157, 183–194. Cited 3 times in pages 28, 42, and 68.
- Liu, Z., Yan, Y., Yang, J., & Hauskrecht, M. (2015). Missing value estimation for hierarchical time series: A study of hierarchical web traffic. In *2015 ieee international conference on data mining* (pp. 895–900). Cited in page 101.
- Livera, A., Hyndman, R., & Snyder, R. (2011). Forecasting time series with complex smoothing exponential using seasonal patterns. *Journal of the American Statistical Association*, 106, 1513–1527. Cited in page 85.
- Longman, R., Giambelluca, T., Nullet, M., Frazier, A., Kodama, K., Crausbay, S., & Arnold, J. (2018). Compilation of climate data from heterogeneous networks across the hawaiian islands. *Scientific data*, 5, 180012. Cited in page 42.
- Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington academy of sciences*, 16(12), 317–323. Cited in page 32.
- The m6 financial forecasting competition*. (n.d.). Retrieved from <https://m6competition.com/> Cited in page 54.
- Ma, Z., Zhong, H., Xie, L., Xia, Q., & Kang, C. (2018). Month ahead average daily electricity price profile forecasting based on a hybrid nonlinear regression and svm model: an ercot case study. *Journal of Modern Power Systems and Clean Energy*, 6(2), 281–291. Cited in page 33.
- Maceira, M. E. P., Melo, A. C., & Zimmermann, M. P. (2016). Application of stochastic programming and probabilistic analyses as key parameters for real decision making

- regarding implementing or not energy rationing-a case study for the brazilian hydrothermal interconnected system. In *2016 power systems computation conference (pscc)* (pp. 1–7). Cited in page 81.
- Madadi, S., Nazari-Heris, M., Mohammadi-Ivatloo, B., & Tohidi, S. (2018). Application of big data analysis to operation of smart power systems. In *Big data in engineering applications* (p. 347–362). Singapore: Springer. Cited in page 42.
- Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. F. (1993). The m2-competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting*, 9(1), 5–22. Cited in page 54.
- Makridakis, S., & Hibon, M. (1979). Accuracy of forecasting: An empirical investigation. *Journal of the Royal Statistical Society: Series A (General)*, 142(2), 97–125. Cited 3 times in pages 54, 59, and 85.
- Makridakis, S., & Hibon, M. (2000). The m3-competition: results, conclusions and implications. *International journal of forecasting*, 16(4), 451–476. Cited 3 times in pages 54, 71, and 76.
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4), 802–808. Cited 4 times in pages 54, 55, 59, and 84.
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1), 54–74. Cited 2 times in pages 72 and 89.
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2021). The m5 competition: Background, organization, and implementation. *International Journal of Forecasting*. Cited in page 54.
- Malvoni, M., De Giorgi, M. G., & Congedo, P. M. (2017). Forecasting of pv power generation using weather input data-preprocessing techniques. *Energy Procedia*, 126, 651–658. Cited in page 92.
- Marchetti, I., & Rego, E. E. (2022). The impact of hourly pricing for renewable generation projects in brazil. *Renewable Energy*, 189, 601–617. Cited 2 times in pages 81 and 86.
- McDermott, P., & Wikle, C. (2016). A model-based approach for analog spatio-temporal dynamic forecasting. *Environmetrics*, 27(2), 70–82. Cited in page 43.
- Medojevic, M., Medic, N., Marjanovic, U., Lalic, B., & Majstorovic, V. (2019). Exploring the impact of industry 4.0 concepts on energy and environmental management systems: Evidence from serbian manufacturing companies. In *Ifip international conference on advances in production management systems* (pp. 355–362). Cited in page 92.
- Milligan, M., Frew, B. A., Bloom, A., Ela, E., Botterud, A., Townsend, A., & Levin, T. (2016). Wholesale electricity market design with increasing levels of renewable generation: Revenue sufficiency and long-term reliability. *The Electricity Journal*, 29(2), 26–38. Cited in page 68.
- Mirani, A., Abdollahzade, M., & Hassani, H. (2013). Day-ahead electricity price analysis and



- forecasting by singular spectrum analysis. *IET Generation, Transmission Distribution*, 7(4), 337–346. Cited 3 times in pages 28, 42, and 68.
- Mišnić, N., Pejović, B., Jovović, J., Rogić, S., & Đurišić, V. (2022). The economic viability of pv power plant based on a neural network model of electricity prices forecast: A case of a developing market. *Energies*, 15(17), 6219. Cited in page 28.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, P. (2009). Reprint—preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Physical therapy*, 89(9), 873–880. Cited in page 29.
- Mohsenian-Rad, A.-H., & Leon-Garcia, A. (2010). Optimal residential load control with price prediction in real-time electricity pricing environments. *IEEE transactions on Smart Grid*, 1(2), 120–133. Cited in page 33.
- Montgomery, D., Peck, E., & Vining, G. (2012). *Introduction to linear regression analysis* (Vol. 821). John Wiley Sons. Cited 2 times in pages 46 and 57.
- Moreno, S., & Santos Coelho, L. (2018). Wind speed forecasting approach based on singular spectrum analysis and adaptive neuro fuzzy inference system. *Renewable energy*, 126, 736–754. Cited 3 times in pages 28, 42, and 68.
- Mota, B., Gomes, L., Faria, P., Ramos, C., Vale, Z., & Correia, R. (2021). Production line optimization to minimize energy cost and participate in demand response events. *Energies*, 14(2), 462. Cited in page 82.
- Mueen, A., Zhu, Y., Yeh, M., Kamgar, K., Viswanathan, K., Gupta, C., & Keogh, E. (2017). *The fastest similarity search algorithm for time series subsequences under euclidean distance*. Retrieved from <https://www.cs.unm.edu/~mueen/FastestSimilaritySearch.html> (accessed 14 November, 2019.) Cited 4 times in pages 42, 43, 48, and 49.
- Munhoz, F. C. (2017). The necessity of more temporal granularity in the brazilian short-term electricity market. *Sustainable Energy, Grids and Networks*, 11, 26–33. Cited 2 times in pages 76 and 83.
- Munhoz, F. C. (2021). Two-settlement system for the brazilian electricity market. *Energy Policy*, 152, 112234. Cited in page 81.
- Nafade, V., Nash, M., Huddart, S., Pande, T., Gebreselassie, N., Lienhardt, C., & Pai, M. (2018). A bibliometric analysis of tuberculosis research, 2007–2016. *PloS one*, 13(6), e0199706. Cited in page 28.
- Nagaraja, Y., Devaraju, T., Kumar, M., & Madichetty, S. (2016). A survey on wind energy, load and price forecasting: (forecasting methods. In *2016 international conference on electrical, electronics, and optimization techniques (iceeot)* (p. 783–788). IEEE. Cited 3 times in pages 28, 42, and 68.
- Oliveira, J. M., & Ramos, P. (2019). Assessing the performance of hierarchical forecasting methods on the retail sector. *Entropy*, 21(4), 436. Cited in page 97.
- Operator, N. S. (2020). Operation history (report of power generation) [Computer software

- manual]. Brasilia, Brazil. Retrieved from <http://www.ons.org.br/paginas/resultados-da-operacao/historico-da-operacao> Cited in page 93.
- Orang, M., & Shiri, N. (2012). A probabilistic approach to correlation queries in uncertain time series data. In *Proceedings of the 21st acm international conference on information and knowledge management* (p. 2229–2233). ACM. Cited in page 44.
- Orcutt, G. H., Watts, H. W., & Edwards, J. B. (1968). Data aggregation and information loss. *The American Economic Review*, 58(4), 773–787. Cited in page 95.
- Pablo-Romero, M. d. P., Pozo-Barajas, R., & Yñiguez, R. (2017). Global changes in residential energy consumption. *Energy Policy*, 101, 342–352. Cited in page 19.
- Pai, P.-F., & Lin, C.-S. (2005). A hybrid arima and support vector machines model in stock price forecasting. *Omega*, 33(6), 497–505. Cited in page 53.
- Pal, A., & Prakash, P. (2017). *Practical time series analysis: master time series data processing, visualization, and modeling using python*. Packt Publishing Ltd. Cited in page 84.
- Panamtash, H., & Zhou, Q. (2018). Coherent probabilistic solar power forecasting. In *2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)* (pp. 1–6). Cited in page 93.
- Panigrahi, S., & Behera, H. S. (2017). A hybrid ets–ann model for time series forecasting. *Engineering applications of artificial intelligence*, 66, 49–59. Cited 3 times in pages 98, 99, and 100.
- Pappas, S. S., Ekonomou, L., Karamousantas, D. C., Chatzarakis, G., Katsikas, S., & Liatsis, P. (2008). Electricity demand loads modeling using autoregressive moving average (arma) models. *Energy*, 33(9), 1353–1360. Cited in page 68.
- Pastor, R., Da Silva, N. P., Esteves, J., & Pestana, R. (2018). Market-based bidding strategy for variable renewable generation in the mibel. In *2018 15th International Conference on the European Energy Market (EEM)* (pp. 1–5). Cited in page 20.
- Paul, G., & Richard, L. (1999). On the asymmetry of the symmetric mape. *International Journal of Forecasting*. Cited in page 89.
- Paulino, R. V. F., Mendonça, A. C., de Azevedo, A. A., Gontijo, T. S., & Casagrande, V. G. (2017). Avaliação da eficiência na construção civil: um estudo de caso em uma obra da região metropolitana de belo horizonte. *Revista Petra*, 3(1). Cited in page 152.
- Pearson, K. (1895). Correlation coefficient. In *Royal society proceedings* (Vol. 58, p. 214). Cited in page 44.
- Pegels, C. C. (1969). Exponential forecasting: Some new variations. *Management Science*, 311–315. Cited in page 100.
- Peng, C., Liu, G., & Xiang, L. (2013). Short-term electricity price forecasting using relief-correlation analysis based on feature selection and differential evolution support vector machine. *Diangong Jishu Xuebao(Transactions of China Electrotechnical Society)*, 28(1), 277–284. Cited in page 35.
- Pir, M., Shah, F., & Asger, M. (2017). Comparative study of different wavelet based neural

- network models for iip growth forecasting using different yield spreads. *International Journal of Electrical Electronics Computer Science Engineering*, 4(6), 5–13. Cited 3 times in pages 28, 42, and 68.
- Pool, N. (2020). *Historical market data (report of spot prices)*. Retrieved from <https://www.nordpoolgroup.com/historical-market-data/>. (Retrieved from) Cited in page 48.
- Pousinho, H. M. I., Mendes, V., & Catalão, J. P. d. S. (2012). Short-term electricity prices forecasting in a competitive market by a hybrid pso–anfis approach. *International Journal of Electrical Power & Energy Systems*, 39(1), 29–35. Cited 2 times in pages 33 and 34.
- Primc, K., & Slabe-Erker, R. (2020). Social policy or energy policy? time to reconsider energy poverty policies. *Energy for Sustainable Development*, 55, 32–36. Cited in page 27.
- R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/> Cited in page 94.
- Radack, G., & Badler, N. (1989). Local matching of surfaces using a boundary-centered radial decomposition. *Computer vision, graphics, and image processing*, 45(3), 380–396. Cited in page 43.
- Rana, M., Koprinska, I., & Agelidis, V. G. (2016). Univariate and multivariate methods for very short-term solar photovoltaic power forecasting. *Energy Conversion and Management*, 121, 380–390. Cited in page 53.
- Rao, S. L. (2004). *Governing power: A new institution of governance, the experience with independent regulation of electricity*. The Energy and Resources Institute (TERI). Cited in page 19.
- Raviv, E., Bouwman, K. E., & Van Dijk, D. (2015). Forecasting day-ahead electricity prices: Utilizing hourly prices. *Energy Economics*, 50, 227–239. Cited in page 53.
- Rayas-Sánchez, J. E., Aguilar-Torrentera, J., & Jasso-Urzúa, J. A. (2010). Surrogate modeling of microwave circuits using polynomial functional interpolants. In *2010 IEEE MTT-S International Microwave Symposium* (pp. 197–200). Cited in page 85.
- Resende, L., Soares, M., & Ferreira, P. (2018). Electric power load in Brazil: view on the long-term forecasting models. *Production*, 28. Cited in page 68.
- Rodrigues, A. d. C., Gontijo, T. S., & Gonçalves, C. A. (2021). Eficiência do gasto público em atenção primária em saúde nos municípios do Rio de Janeiro, Brasil: escores robustos e seus determinantes. *Ciência & Saúde Coletiva*, 26, 3567–3579. Cited in page 151.
- Rostamnia, N., & Rashid, T. A. (2019). Investigating the effect of competitiveness power in estimating the average weighted price in electricity market. *The Electricity Journal*, 32(8), 106628. Cited in page 68.
- Santos, E. G., Calipo, E. R., & Gontijo, T. S. (2017). Otimização da produtividade através da redução do tempo de setup em terminais de cartão de crédito. *Revista Gestão Industrial*, 13(1). Cited in page 152.

- Santos, E. S. d., Juchem, L., & Maduro, L. A. R. (2017). Performance sport, tax waiver and sports incentive law. *Journal of Physical Education*, 28. Cited 3 times in pages 76, 82, and 83.
- Santos, T., Diniz, A., Saboia, C., Cabral, R., & Cerqueira, L. (2020). Hourly pricing and day-ahead dispatch setting in brazil: The dessem model. *Electric Power Systems Research*, 189, 106709. Cited 2 times in pages 82 and 86.
- Schuh, G., Prote, J.-P., Sauermann, F., & Franzkoch, B. (2019). Databased prediction of order-specific transition times. *CIRP Annals*, 68(1), 467–470. Cited in page 68.
- Shanmugam, R. (2006). Book review. *Journal of Statistical Computation and Simulation*, 76(10), 935-940. Retrieved from <https://doi.org/10.1080/00949650412331321034> doi: 10.1080/00949650412331321034 Cited in page 53.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4), 591–611. Cited in page 75.
- Sharifzadeh, M., Sikinioti-Lock, A., & Shah, N. (2019). Machine-learning methods for integrated renewable power generation: A comparative study of artificial neural networks, support vector regression, and gaussian process regression. *Renewable and Sustainable Energy Reviews*, 108, 513–538. Cited in page 92.
- Shrivastava, N. A., Panigrahi, B. K., & Lim, M.-H. (2016). Electricity price classification using extreme learning machines. *Neural Computing and Applications*, 27(1), 9–18. Cited in page 35.
- Silveira Gontijo, T., & Azevedo Costa, M. (2020). Forecasting hierarchical time series in power generation. *Energies*, 13(14), 3722. Cited in page 151.
- Simmhan, Y., & Noor, M. U. (2013). Scalable prediction of energy consumption using incremental time series clustering. In *2013 IEEE International Conference on Big Data* (pp. 29–36). Cited in page 68.
- Simonsen, I. (2003). Measuring anti-correlations in the nordic electricity spot market by wavelets. *Physica A: Statistical Mechanics and its applications*, 322, 597–606. Cited in page 68.
- Soeiro, S., & Dias, M. F. (2020). Renewable energy community and the european energy market: Main motivations. *Heliyon*, 6(7), e04511. Cited in page 28.
- Steffen, B., & Patt, A. (2022). A historical turning point? early evidence on how the russia-ukraine war changes public support for clean energy policies. *Energy Research & Social Science*, 91, 102758. Cited 3 times in pages 27, 28, and 82.
- Steinert, R., & Ziel, F. (2019). Short-to mid-term day-ahead electricity price forecasting using futures. *The Energy Journal*, 40(1). Cited in page 34.
- Strehl, A., Ghosh, J., & Mooney, R. (2000). Impact of similarity measures on web-page clustering. In *Workshop on artificial intelligence for web search (aaai 2000)* (Vol. 58, p. 64). Cited in page 44.
- Strozzi, F., Colicchia, C., Creazza, A., & Noè, C. (2017). Literature review on the ‘smart

- factory' concept using bibliometric tools. *International Journal of Production Research*, 55(22), 6572–6591. Cited in page 69.
- Tahmasebifar, R., Sheikh-El-Eslami, M. K., & Kheirollahi, R. (2017). Point and interval forecasting of real-time and day-ahead electricity prices by a novel hybrid approach. *IET Generation, Transmission & Distribution*, 11(9), 2173–2183. Cited in page 35.
- Taylor, J. W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*, 54(8), 799–805. Cited in page 100.
- Tchrakian, T. T., Basu, B., & O'Mahony, M. (2011). Real-time traffic flow forecasting using spectral analysis. *IEEE Transactions on Intelligent Transportation Systems*, 13(2), 519–526. Cited in page 53.
- Tian, P., Xiao, X., Wang, K., & Ding, R. (2015). A hierarchical energy management system based on hierarchical optimization for microgrid community economic operation. *IEEE Transactions on Smart Grid*, 7(5), 2230–2241. Cited in page 68.
- Tjørring, L., & Gausset, Q. (2015). Energy renovation models in private households in denmark. *Community governance and citizen-driven initiatives in climate change mitigation*, 89–106. Cited 2 times in pages 19 and 80.
- Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 99–114. Cited in page 75.
- Tularam, G. A., & Saeed, T. (2016). Oil-price forecasting based on various univariate time-series models. *American Journal of Operations Research*, 6(03), 226. Cited in page 53.
- Uniejewski, B., Marcjasz, G., & Weron, R. (2019). On the importance of the long-term seasonal component in day-ahead electricity price forecasting: Part ii—probabilistic forecasting. *Energy Economics*, 79, 171–182. Cited in page 33.
- Van Eck, N., & Waltman, L. (2010). Software survey: Vosviewer, a computer program for bibliometric mapping. *scientometrics*, 84(2), 523–538. Cited 2 times in pages 31 and 70.
- Van Eck, N. J., & Waltman, L. (2017). Citation-based clustering of publications using citnetexplorer and vosviewer. *Scientometrics*, 111(2), 1053–1070. Cited in page 28.
- van Raan, A. F. (2017). Patent citations analysis and its value in research evaluation: A review and a new approach to map technology-relevant research. *Journal of Data and Information Science*, 2(1), 13–50. Cited in page 28.
- Voronin, S., & Partanen, J. (2014). Forecasting electricity price and demand using a hybrid approach based on wavelet transform, arima and neural networks. *International Journal of Energy Research*, 38(5), 626–637. Cited in page 42.
- Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, 105, 569–582. Cited in page 53.
- Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science*



- and Technology*, 63(12), 2378–2392. Cited in page 31.
- Wan, C., Lin, J., Wang, J., Song, Y., & Dong, Z. Y. (2016). Direct quantile regression for nonparametric probabilistic forecasting of wind power generation. *IEEE Transactions on Power Systems*, 32(4), 2767–2778. Cited 2 times in pages 19 and 81.
- Wang, S. (2006). Hyndman, 2006 wang x., smith k., hyndman r. *Characteristic-based clustering for time series data*, *Data Mining and Knowledge Discovery*, 13, 335–364. Cited in page 84.
- Wang, W., Chen, Q., Yan, D., & Geng, D. (2019). A novel comprehensive evaluation method of the draft tube pressure pulsation of Francis turbine based on EEMD and information entropy. *Mechanical Systems and Signal Processing*, 116, 772–786. Retrieved from <https://doi.org/10.1016/j.ymsp.2018.07.033> doi: 10.1016/j.ymsp.2018.07.033 Cited in page 35.
- Weiss, C. (2018). *Essays in hierarchical time series forecasting and forecast combination* (Unpublished doctoral dissertation). University of Cambridge. Cited in page 101.
- Weron, R. (2007). *Modeling and forecasting electricity loads and prices: A statistical approach* (Vol. 403). John Wiley & Sons. Cited in page 68.
- Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International journal of forecasting*, 30(4), 1030–1081. Cited 4 times in pages 28, 29, 68, and 69.
- Wickramasuriya, S. L., Athanasopoulos, G., & Hyndman, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526), 804–819. Cited 5 times in pages 93, 95, 98, 99, and 101.
- Windler, T., Busse, J., & Rieck, J. (2019). One month-ahead electricity price forecasting in the context of production planning. *Journal of Cleaner Production*, 238, 117910. Cited in page 35.
- Wolfram, C. D. (1999). Measuring duopoly power in the british electricity spot market. *American Economic Review*, 89(4), 805–826. Cited in page 72.
- Woo, C. K., & Zarnikau, J. (2019). A nice electricity market design. *The Electricity Journal*, 32(9), 106638. Cited in page 68.
- Xu, F., Ouyang, D.-l., Rene, E. R., Ng, H. Y., Guo, L.-l., Zhu, Y.-j., ... others (2019). Electricity production enhancement in a constructed wetland-microbial fuel cell system for treating saline wastewater. *Bioresource technology*, 288, 121462. Cited 2 times in pages 19 and 81.
- Yang, D. (2019). Ultra-fast analog ensemble using kd-tree. *Journal of Renewable and Sustainable Energy*, 11(5), 053703. Cited in page 43.
- Yang, D., & Alessandrini, S. (2019). An ultra-fast way of searching weather analogs for renewable energy forecasting. *Solar Energy*, 185, 255–261. Cited 7 times in pages 42, 43, 45, 47, 49, 54, and 68.

- Yang, D., Kleissl, J., Gueymard, C. A., Pedro, H. T., & Coimbra, C. F. (2018). History and trends in solar irradiance and pv power forecasting: A preliminary assessment and review using text mining. *Solar Energy*, *168*, 60–101. Cited 3 times in pages 42, 43, and 98.
- Yang, D., Wu, E., & Kleissl, J. (2019). Operational solar forecasting for the real-time market. *International Journal of Forecasting*, *35*(4), 1499–1519. Cited in page 68.
- Yang, J., Astitha, M., Delle Monache, L., & Alessandrini, S. (2018). An analog technique to improve storm wind speed prediction using a dual nwp model approach. *Monthly Weather Review*, *146*(12), 4057–4077. Cited 2 times in pages 42 and 43.
- Yang, W., Sun, S., Hao, Y., & Wang, S. (2022). A novel machine learning-based electricity price forecasting model based on optimal model selection strategy. *Energy*, *238*, 121989. Cited in page 28.
- Yuan, X. (2013). Overview of problems in large-scale wind integrations. *Journal of Modern Power Systems and Clean Energy*, *1*(1), 22–25. Cited in page 33.
- Zahid, M., Ahmed, F., Javaid, N., Abbasi, R. A., Zainab Kazmi, H. S., Javaid, A., ... Ilahi, M. (2019). Electricity price and load forecasting using enhanced convolutional neural network and enhanced support vector regression in smart grids. *Electronics*, *8*(2), 122. Cited in page 33.
- Zhang, J., Tan, Z., & Wei, Y. (2020). An adaptive hybrid model for short term electricity price forecasting. *Applied Energy*, *258*, 114087. Cited in page 28.
- Zhang, J.-L., Zhang, Y.-J., Li, D.-Z., Tan, Z.-F., & Ji, J.-F. (2019). Forecasting day-ahead electricity prices using a new integrated model. *International journal of electrical power & energy systems*, *105*, 541–548. Cited in page 35.
- Zhang, L., Wu, Q., Ma, Z., & Wang, X. (2019). Transient vibration analysis of unit-plant structure for hydropower station in sudden load increasing process. *Mechanical Systems and Signal Processing*, *120*(79), 486–504. Retrieved from <https://doi.org/10.1016/j.ymsp.2018.10.037> doi: 10.1016/j.ymsp.2018.10.037 Cited in page 35.
- Zhang, Y., Li, C., & Li, L. (2018). Wavelet transform and kernel-based extreme learning machine for electricity price forecasting. *Energy Systems*, *9*(1), 113–134. Cited in page 33.
- Zhang, Y., Zhong, M., Geng, N., & Jiang, Y. (2017). Forecasting electric vehicles sales with univariate and multivariate time series models: The case of china. *PloS one*, *12*(5), e0176729. Cited in page 53.
- Zhong, H., Tan, Z., He, Y., Xie, L., & Kang, C. (2020). Implications of covid-19 for the electricity industry: A comprehensive review. *CSEE Journal of Power and Energy Systems*, *6*(3), 489–495. Cited in page 86.
- Zhu, Y., Imamura, M., Nikovski, D., & Keogh, E. (2019). Introducing time series chains: a new primitive for time series data mining. *Knowledge and Information Systems*, *60*(2), 1135–1161. Cited in page 43.
- Zuhaira, Z., Li, J., & Mohammed, H. D. (2022). The future of the shale industry in light of the fluctuations in global oil prices. *Energy & Environment*, 0958305X221129223. Cited in

page 28.



# Appendix

# APPENDIX A – Complementary results

## A.1 Chapter 3

Figure 32 – Annual scientific production.

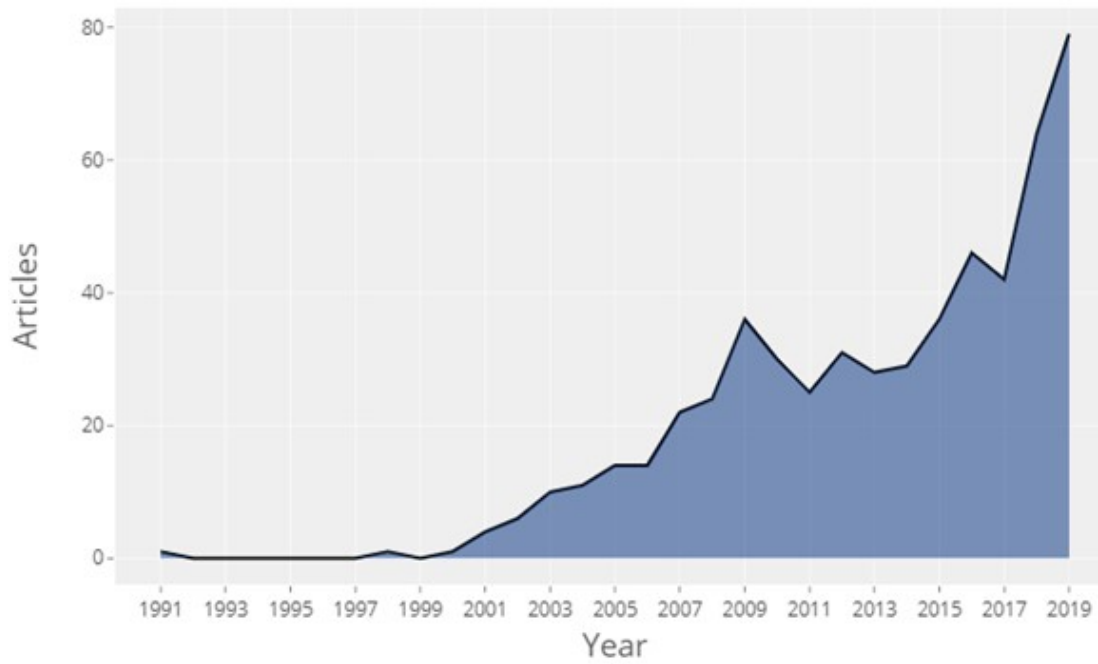


Figure 33 – Average article citations per year.

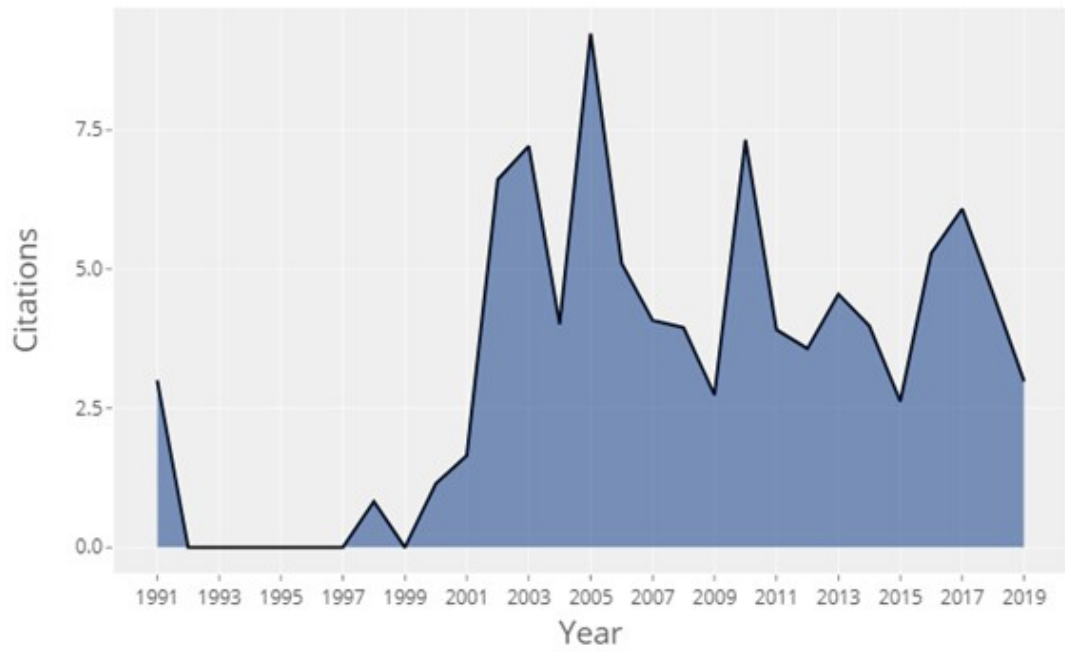


Figure 34 – Sources dynamics.

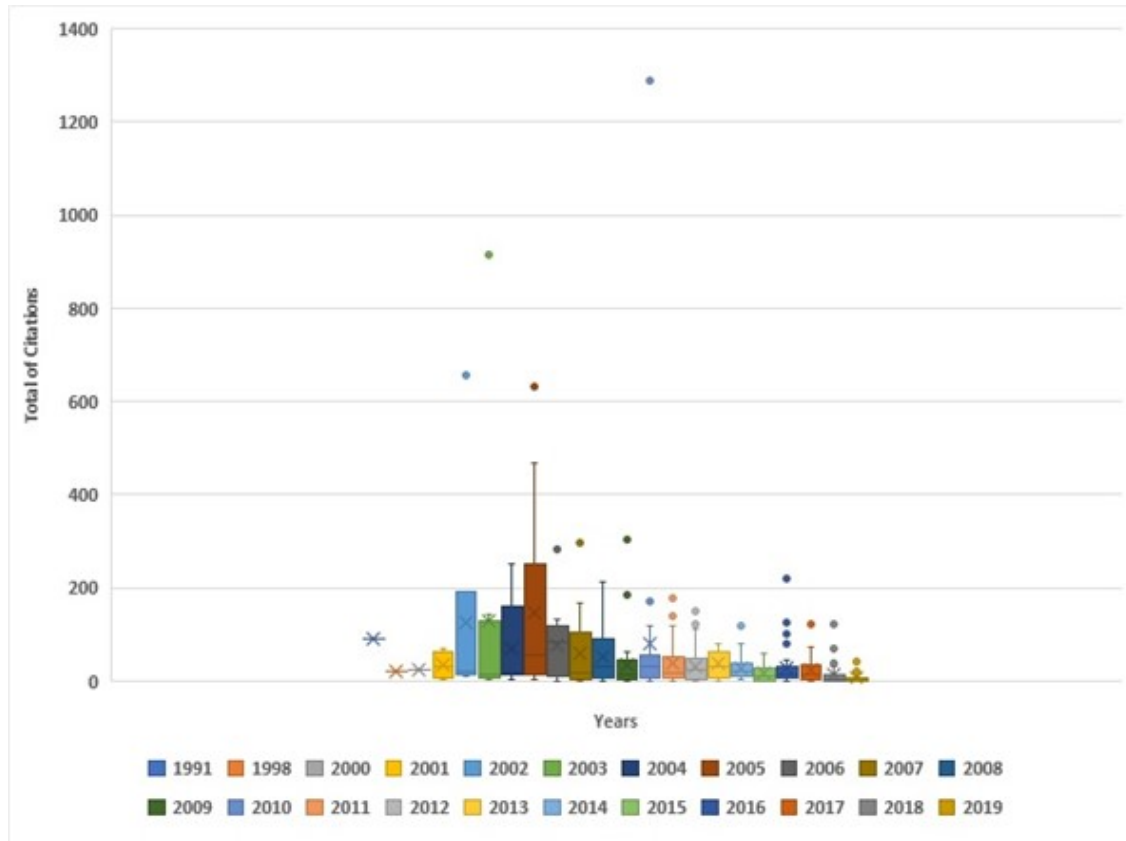
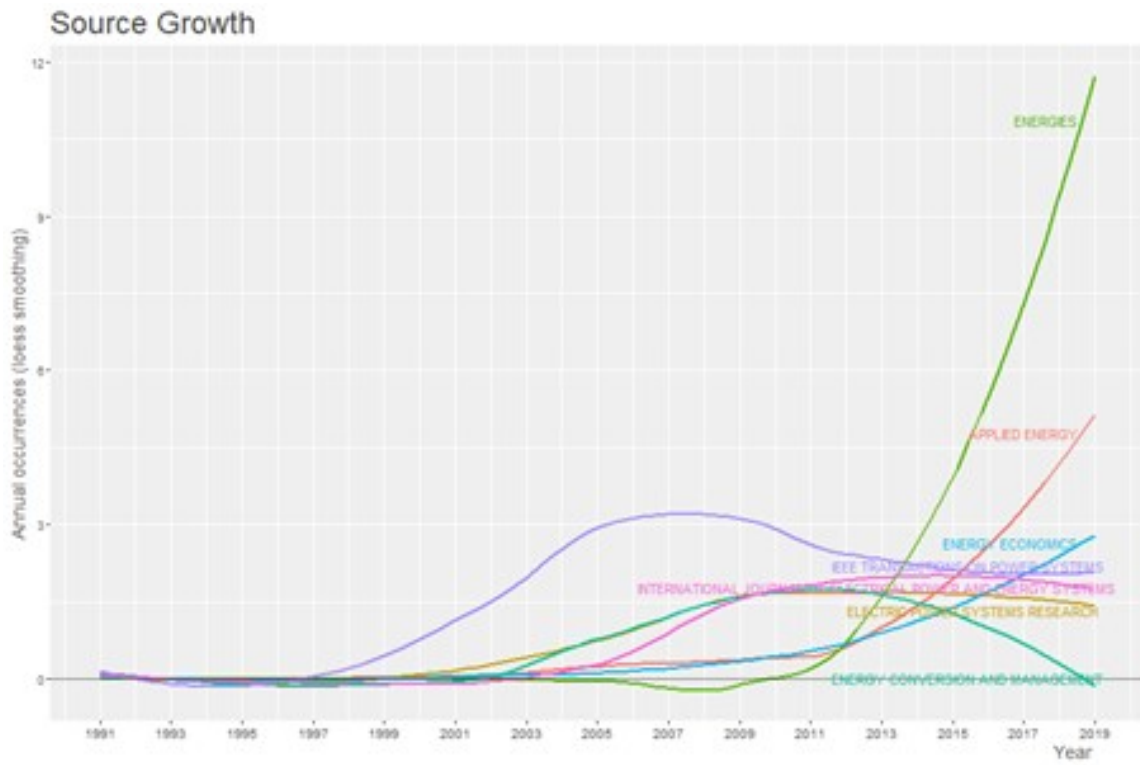


Figure 35 – Sources dynamics.



## A.2 Chapter 8

Figure 36 –  $sMAPE(\%)$  by step - comparison between the *DTSF* and the benchmark models.

Step - h	$sMAPE(\%)$						
	Arima	DTSF	Ets	Naïve	Tbats	Theta	XGBoost
1	0.705	<b>0.649</b>	2.773	1.261	1.557	1.513	0.868
2	1.678	<b>1.651</b>	6.217	2.345	5.038	5.861	3.717
3	3.095	<b>1.771</b>	10.860	4.757	7.911	9.673	3.942
4	5.833	3.722	15.794	7.930	12.610	13.888	<b>2.758</b>
5	5.982	<b>2.535</b>	17.691	8.513	13.700	14.917	2.734
6	5.434	<b>2.175</b>	16.911	7.924	13.829	14.512	2.429
7	3.706	<b>1.635</b>	13.864	5.798	11.987	11.686	2.875
8	<b>3.621</b>	4.422	11.778	5.229	11.438	8.282	4.726
9	<b>9.700</b>	10.288	14.990	10.194	15.966	12.846	10.395
10	10.303	10.428	12.987	<b>9.693</b>	14.161	11.471	10.573
11	4.983	5.380	5.975	<b>4.085</b>	7.243	5.083	4.881
12	2.554	3.156	2.748	<b>1.950</b>	4.377	2.488	2.383
13	2.573	2.954	2.669	<b>2.045</b>	3.512	2.399	2.368
14	2.266	2.692	2.784	1.774	2.846	2.234	<b>1.713</b>
15	2.909	3.177	2.691	<b>1.941</b>	3.088	2.531	2.285
16	3.476	3.967	2.533	<b>2.481</b>	3.136	2.811	2.680
17	3.496	3.740	2.326	2.392	2.989	2.635	<b>2.282</b>
18	3.792	2.918	<b>2.441</b>	2.731	3.471	2.753	2.791
19	3.420	<b>2.108</b>	2.409	2.599	4.384	2.549	3.007
20	4.005	<b>2.019</b>	3.823	3.948	4.878	3.018	3.311
21	3.747	2.578	2.712	3.397	5.650	<b>2.453</b>	3.183
22	3.554	2.633	<b>2.324</b>	2.959	4.894	2.438	3.359
23	3.695	3.159	<b>2.322</b>	2.872	3.771	2.734	3.390
24	3.380	2.996	2.724	2.725	2.733	<b>2.712</b>	3.457

Source: Research results.

Figure 37 – *MASE* by step - comparison between the *DTSF* and the benchmark models.

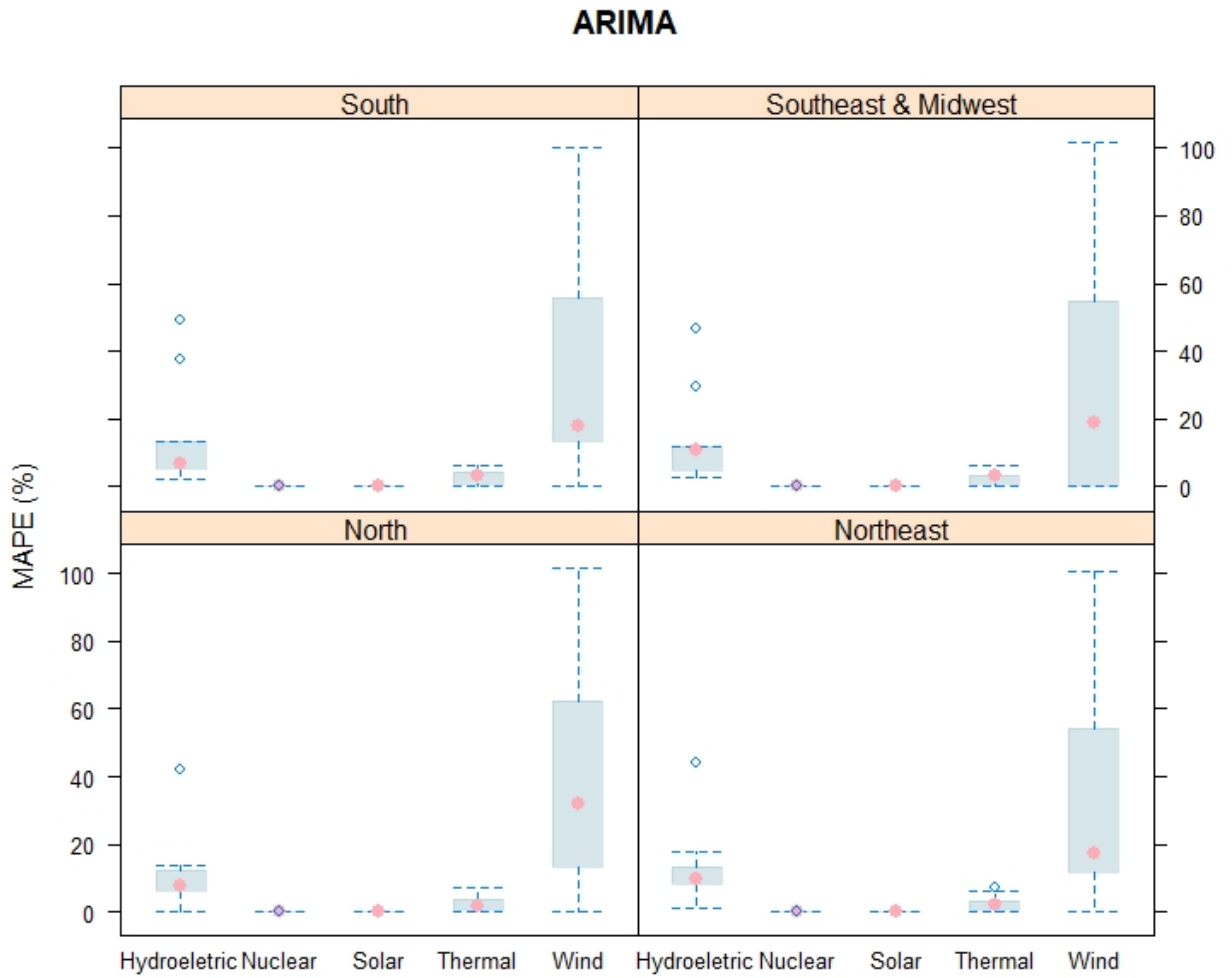
MASE							
Step - h	Arima	DTSF	Ets	Naïve	Tbats	Theta	XGBoost
1	0.145	<b>0.083</b>	0.303	0.247	0.207	0.336	0.175
2	<b>0.312</b>	0.339	0.629	0.368	0.693	1.288	0.674
3	0.521	<b>0.272</b>	1.032	0.772	1.012	1.992	0.615
4	0.882	<b>0.482</b>	1.524	1.182	1.563	2.633	0.499
5	0.896	<b>0.313</b>	1.647	1.283	1.633	2.800	0.496
6	0.850	<b>0.303</b>	1.574	1.193	1.651	2.765	0.469
7	0.625	<b>0.263</b>	1.247	0.923	1.360	2.365	0.513
8	<b>0.889</b>	0.992	1.362	1.156	1.407	1.720	0.973
9	<b>2.382</b>	2.473	2.586	2.498	2.752	2.958	2.396
10	2.451	2.535	2.510	<b>2.411</b>	2.574	2.716	2.428
11	1.053	1.279	1.107	<b>1.004</b>	1.075	1.192	1.079
12	0.332	0.587	0.371	<b>0.304</b>	0.388	0.433	0.374
13	0.332	0.581	0.391	<b>0.314</b>	0.320	0.418	0.373
14	0.316	0.573	0.391	0.286	<b>0.276</b>	0.435	0.298
15	0.388	0.667	0.420	<b>0.306</b>	0.338	0.426	0.369
16	0.461	0.812	0.500	<b>0.381</b>	0.449	0.439	0.456
17	0.524	0.740	0.478	<b>0.361</b>	0.525	0.427	0.426
18	0.730	<b>0.463</b>	0.563	0.573	0.785	0.582	0.664
19	0.804	<b>0.379</b>	0.628	0.689	1.105	0.692	0.818
20	0.981	<b>0.483</b>	0.817	1.009	1.240	0.914	0.861
21	0.910	<b>0.575</b>	0.653	0.868	1.326	0.806	0.834
22	0.863	<b>0.581</b>	0.688	0.778	1.164	0.779	0.890
23	0.864	<b>0.683</b>	0.722	0.741	0.952	0.784	0.907
24	0.760	<b>0.642</b>	0.658	0.658	0.721	0.657	0.857

Source: Research results.



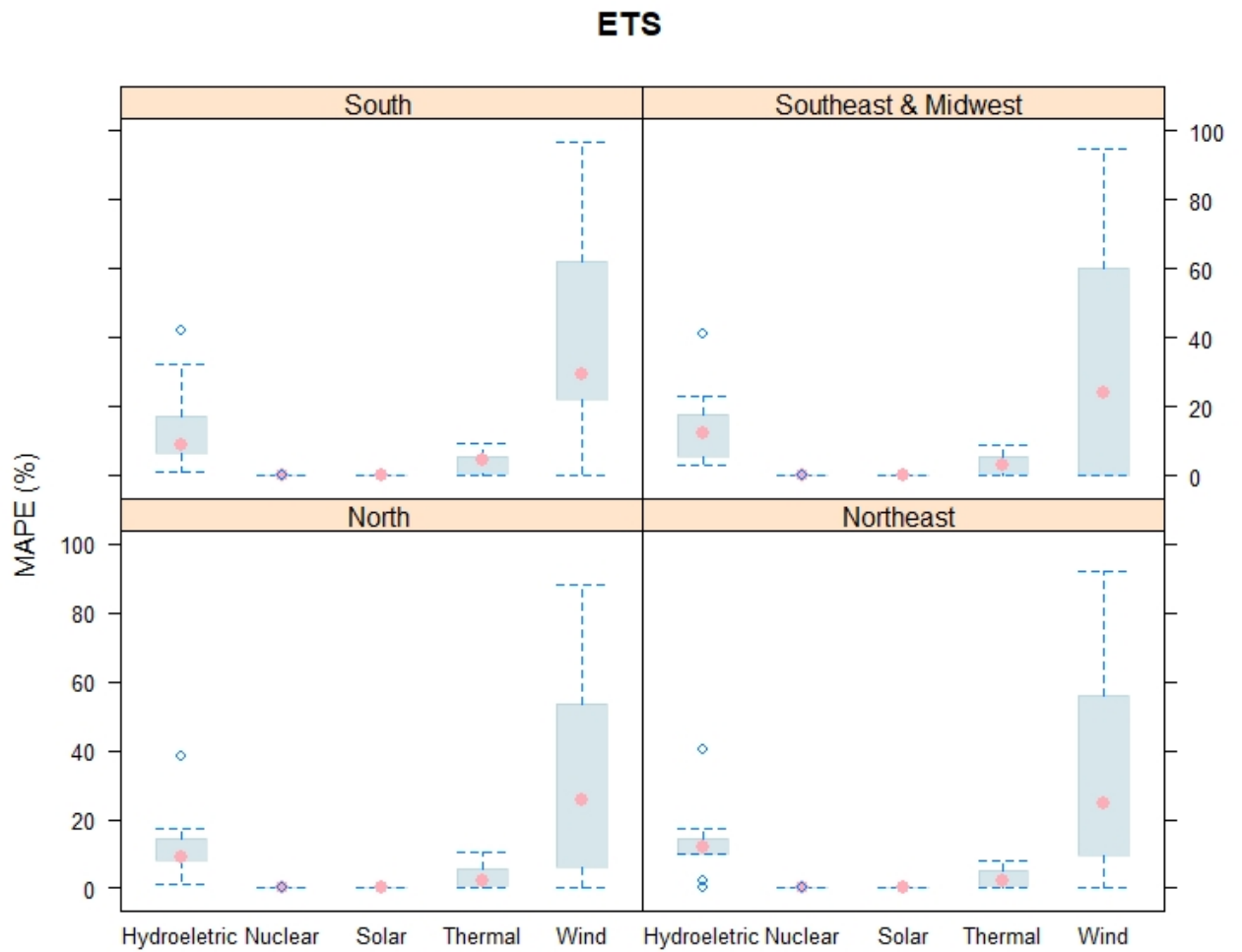
### A.3 Chapter 9

Figure 38 – Hierarchical forecasting for power generation: electrical subsystem versus generating source (ARIMA).



Source: Research results.

Figure 39 – Hierarchical forecasting for power generation: electrical subsystem versus generating source (ETS).



Source: Research results.

Figure 40 – Hierarchical forecasting for electricity generation based on the ARIMA procedure (RMSE, MAE, MASE) Note: The performance was indicated into a color scale, where green means better values for calculated accuracy, and red means worse accuracy. The intermediate values are colored yellow.

Predictive model: Autoregressive integrated moving average (ARIMA)												
Method	RMSE				MAE				MASE			
	Forecast horizon (h) - mean of the interval											
	[1-3]	[4-6]	[7-9]	Mean	[1-3]	[4-6]	[7-9]	Mean	[1-3]	[4-6]	[7-9]	Mean
<b>Hierarchical level 0 : Total - Brazil</b>												
BU	2.36	7.77	8.50	6.21	2.20	6.46	7.14	5.27	1.11	3.27	3.62	2.67
TDGSA	2.47	8.07	9.33	6.62	2.35	7.11	8.35	5.93	1.19	3.60	4.23	3.01
TDGSF	2.47	8.07	9.33	6.62	2.35	7.11	8.35	5.93	1.19	3.60	4.23	3.01
TDFP	2.47	8.07	9.33	6.62	2.35	7.11	8.35	5.93	1.19	3.60	4.23	3.01
OLS	2.41	7.97	9.12	6.50	2.28	6.95	8.05	5.76	1.15	3.52	4.08	2.92
WLSv	2.34	7.82	8.75	6.31	2.19	6.66	7.49	5.45	1.11	3.37	3.80	2.76
WLSs	2.32	7.79	8.67	6.26	2.16	6.59	7.36	5.37	1.09	3.34	3.73	2.72
MintI (Sample)	2.22	7.71	8.56	6.16	2.06	6.49	7.22	5.26	1.05	3.29	3.65	2.66
MinI (Shrink)	2.25	7.75	8.63	6.21	2.10	6.56	7.28	5.31	1.07	3.32	3.69	2.69
<b>Hierarchical level 1 : Electrical subsystems</b>												
BU	1.04	2.99	3.18	2.40	0.97	2.52	2.66	2.05	1.04	2.74	2.97	2.25
TDGSA	5.48	6.88	7.06	6.47	5.46	6.72	6.88	6.35	6.30	7.44	7.45	7.06
TDGSF	5.51	6.91	7.08	6.50	5.50	6.75	6.90	6.38	6.28	7.40	7.41	7.03
TDFP	1.16	2.95	3.99	2.70	1.13	2.75	3.71	2.53	1.20	2.93	3.95	2.70
OLS	1.08	3.20	3.65	2.64	1.04	2.89	3.27	2.40	1.12	3.14	3.64	2.64
WLSv	1.05	3.10	3.44	2.53	1.00	2.73	3.01	2.24	1.06	2.91	3.26	2.41
WLSs	1.05	3.10	3.44	2.53	1.00	2.73	3.00	2.24	1.07	2.95	3.33	2.45
MintI (Sample)	1.01	3.07	3.41	2.50	0.96	2.70	2.95	2.20	1.00	2.87	3.18	2.35
MinI (Shrink)	1.02	3.08	3.43	2.51	0.97	2.72	2.97	2.22	1.02	2.90	3.22	2.38
<b>Hierarchical level 2 : Energy sources</b>												
BU	0.89	2.32	2.38	1.87	0.83	1.98	2.04	1.62	1.22	2.70	2.95	2.29
TDGSA	3.39	4.70	4.93	4.34	3.37	4.57	4.79	4.24	14.02	14.92	15.04	14.66
TDGSF	3.36	4.68	4.90	4.32	3.35	4.55	4.76	4.22	14.79	15.66	15.77	15.41
TDFP	0.98	2.30	2.97	2.08	0.95	2.15	2.77	1.95	1.62	3.57	4.91	3.37
OLS	0.90	2.37	2.49	1.92	0.85	2.07	2.15	1.69	1.85	5.25	7.12	4.74
WLSv	0.91	2.42	2.63	1.98	0.86	2.18	2.33	1.79	1.25	2.88	3.22	2.45
WLSs	0.90	2.35	2.44	1.89	0.84	2.04	2.09	1.66	1.60	4.22	5.37	3.73
MintI (Sample)	0.88	2.40	2.62	1.97	0.84	2.17	2.31	1.77	1.23	2.87	3.24	2.45
MinI (Shrink)	0.89	2.41	2.64	1.98	0.85	2.18	2.33	1.79	1.24	2.89	3.26	2.46

Source: Research results.

Figure 41 – Hierarchical forecasting for electricity generation based on the ETS procedure (RMSE, MAE, MASE). Note: The performance was indicated into a color scale, where green means better values for calculated accuracy, and red means worse accuracy. The intermediate values are colored yellow.

Predictive model: Error, trend, seasonality (ETS)												
Method	RMSE				MAE				MASE			
	Forecast horizon (h) - mean of the interval											
	[1-3]	[4-6]	[7-9]	Mean	[1-3]	[4-6]	[7-9]	Mean	[1-3]	[4-6]	[7-9]	Mean
<b>Hierarchical level 0 : Total - Brazil</b>												
BU	3.02	9.14	11.26	7.81	2.98	8.62	10.57	7.39	1.51	4.37	5.35	3.74
TDGSA	2.65	8.62	10.51	7.26	2.59	7.98	9.76	6.78	1.31	4.04	4.94	3.43
TDGSF	2.65	8.62	10.51	7.26	2.59	7.98	9.76	6.78	1.31	4.04	4.94	3.43
TDFP	2.65	8.62	10.51	7.26	2.59	7.98	9.76	6.78	1.31	4.04	4.94	3.43
OLS	2.67	8.64	10.54	7.28	2.61	8.01	9.78	6.80	1.32	4.06	4.96	3.44
WLSv	2.77	8.78	10.75	7.43	2.72	8.19	10.01	6.97	1.38	4.15	5.07	3.53
WLSs	2.78	8.81	10.78	7.46	2.73	8.22	10.05	7.00	1.38	4.16	5.09	3.54
MintI (Sample)	2.61	8.56	10.43	7.20	2.55	7.91	9.66	6.71	1.29	4.01	4.89	3.40
MinI (Shrink)	2.62	8.58	10.45	7.21	2.56	7.92	9.68	6.72	1.30	4.01	4.90	3.40
<b>Hierarchical level 1 : Electrical subsystems</b>												
BU	1.30	3.66	4.48	3.15	1.29	3.47	4.18	2.98	1.44	3.84	4.66	3.31
TDGSA	5.54	7.09	7.45	6.70	5.54	6.98	7.30	6.61	6.37	7.65	7.81	7.28
TDGSF	5.58	7.12	7.48	6.73	5.57	7.00	7.32	6.63	6.34	7.62	7.77	7.24
TDFP	1.13	2.95	4.07	2.72	1.11	2.77	3.79	2.56	1.22	3.03	4.14	2.80
OLS	1.15	3.42	4.15	2.91	1.13	3.21	3.84	2.73	1.23	3.50	4.22	2.99
WLSv	1.20	3.50	4.25	2.98	1.18	3.29	3.95	2.81	1.31	3.61	4.37	3.10
WLSs	1.20	3.50	4.26	2.98	1.18	3.29	3.95	2.81	1.30	3.60	4.36	3.08
MintI (Sample)	1.20	3.50	4.25	2.98	1.17	3.28	3.93	2.79	1.25	3.52	4.26	3.01
MinI (Shrink)	1.18	3.47	4.22	2.96	1.16	3.26	3.91	2.77	1.25	3.51	4.25	3.00
<b>Hierarchical level 2 : Energy sources</b>												
BU	1.15	2.93	3.53	2.54	1.14	2.83	3.34	2.44	1.64	3.87	4.68	3.40
TDGSA	3.43	4.85	5.21	4.50	3.42	4.75	5.08	4.42	14.21	15.56	16.08	15.28
TDGSF	3.41	4.83	5.18	4.47	3.40	4.73	5.05	4.39	14.97	16.29	16.78	16.01
TDFP	1.05	2.47	3.28	2.27	1.03	2.36	3.11	2.17	1.87	3.88	5.06	3.60
OLS	1.13	2.89	3.48	2.50	1.11	2.79	3.29	2.40	2.64	5.55	6.80	5.00
WLSv	1.07	2.81	3.36	2.41	1.05	2.69	3.17	2.30	1.55	3.74	4.50	3.26
WLSs	1.13	2.91	3.50	2.51	1.12	2.80	3.31	2.41	2.31	4.99	6.09	4.46
MintI (Sample)	1.01	2.71	3.23	2.32	0.99	2.59	3.03	2.20	1.44	3.38	3.99	2.93
MinI (Shrink)	1.03	2.75	3.28	2.35	1.01	2.63	3.08	2.24	1.43	3.49	4.19	3.03

Source: Research results.

# APPENDIX B – Approvals

## Chapter 3

Encontro de Gestão e Negócios (EGEN 2021) - Universidade Federal de Uberlândia

<https://www.even3.com.br/egen/>

**Title:** Melhor artigo na área de Produção e Logística - EGEN 2021.

**Date:** 27-29 September 2021





## Chapter 4

### Proceedings of the Institution of Mechanical Engineers Part O-Journal of Risk and Reliability

<https://journals.sagepub.com/doi/abs/10.1177/1748006X211035623>

**Title:** Condition-based maintenance in hydroelectric plants: A systematic literature review.

**Date:** 2021



Review Article

Institution of  
**MECHANICAL  
ENGINEERS**



## Condition-based maintenance in hydroelectric plants: A systematic literature review

Proc IMechE Part O:  
*J Risk and Reliability*  
1–16  
© IMechE 2021  
Article reuse guidelines:  
[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)  
DOI: 10.1177/1748006X211035623  
[journals.sagepub.com/home/pio](https://journals.sagepub.com/home/pio)

Rodrigo Barbosa de Santis<sup>ORCID</sup>, Tiago Silveira Gontijo and Marcelo Azevedo Costa

### Abstract

Industrial maintenance has become an essential strategic factor for profit and productivity in industrial systems. In the modern industrial context, condition-based maintenance guides the interventions and repairs according to the machine's health status, calculated from monitoring variables and using statistical and computational techniques. Although several literature reviews address condition-based maintenance, no study discusses the application of these techniques in the hydroelectric sector, a fundamental source of renewable energy. We conducted a systematic literature review of articles published in the area of condition-based maintenance in the last 10 years. This was followed by quantitative and thematic analyses of the most relevant categories that compose the phases of condition-based maintenance. We identified a research trend in the application of machine learning techniques, both in the diagnosis and the prognosis of the generating unit's assets, being vibration the most frequently discussed monitoring variable. Finally, there is a vast field to be explored regarding the application of statistical models to estimate the useful life, and hybrid models based on physical models and specialists' knowledge, of turbine-generators.

### Keywords

Condition based maintenance, hydroelectric, fault diagnostics, fault isolation, fault monitoring, fault prognostics, system health management

Date received: 18 September 2020; accepted: 8 July 2021

## Chapter 5 (part a)

### The 40th International Symposium on Forecasting

<https://isf.forecasters.org/>

**Title:** Similarity search in electricity prices: an ultra-fast method for finding analogs

**Date:** 26-28 October 2020

#### Organizing Committee

**George Athanasopoulos**  
IIF President  
Monash University, Australia  
george.athanasopoulos@monash.edu

**Tao Hong**  
PROGRAM CHAIR  
University of North Carolina, USA  
hongtao01@gmail.com

**Pam Stroud**  
IIF Business Director  
Medford, MA USA  
isf@forecasters.org



International Institute of Forecasters

40th INTERNATIONAL SYMPOSIUM ON FORECASTING  
Virtual ~ October 26-28, 2020

26 October 2020

Tiago Silveira Gontijo  
Universidade Federal de Minas Gerais  
Av. Antônio Carlos 6627  
Belo Horizonte Minas Gerais 31270-901  
Brazil

Dear Tiago Silveira Gontijo,

We would like to thank you for your participation at the 40th International Symposium on Forecasting, which took place virtually from October 26-28, 2020. Your presentation entitled *Similarity search in electricity prices: an ultra-fast method for finding analogs* (isf-abs-6346e) was a welcome addition to the conference program. We thank you for presenting.

ISF 2020 was hosted by the International Institute of Forecasters, the pre-eminent organization for scholars and practitioners in the field of forecasting. The IIF is dedicated to stimulating the generation, distribution and use of knowledge on forecasting in a wide range of fields.

Your participation at the International Symposium on Forecasting was most welcomed. We look forward to seeing you at future events ~ Mark your calendar for ISF 2021 ~ June 27-30<sup>th</sup>!

Sincerely,

George Athanasopoulos  
IIF President



## Chapter 5 (part b)

Journal of Renewable and Sustainable Energy - ISSN: 1941-7012

<https://doi.org/10.1063/5.0021557>

**Title:** Similarity search in electricity prices: An ultra-fast method for finding analogs

**Date:** 26 October 2020

Journal of Renewable  
and Sustainable Energy


ARTICLE

scitation.org/journal/rse

### Similarity search in electricity prices: An ultra-fast method for finding analogs

Cite as: J. Renewable Sustainable Energy **12**, 056103 (2020); doi:10.1063/5.0021557

Submitted: 12 July 2020 · Accepted: 1 October 2020 ·  
Published Online: 26 October 2020





Tiago Silveira Gontijo,<sup>✉</sup> Marcelo Azevedo Costa, and Rodrigo Barbosa de Santis 

**AFFILIATIONS**  
Industrial Engineering Department, UFMG Federal University of Minas Gerais, Belo Horizonte, MG, Brazil

<sup>✉</sup>Author to whom correspondence should be addressed: [tgontijo@hotmail.com](mailto:tgontijo@hotmail.com). Tel: +55 31 3409-4697

**ABSTRACT**  
Accurately predicting electricity prices allows us to minimize risks and establish more reliable decision support mechanisms. In particular, the theory of analogs has gained increasing prominence in this area. The analog approach is constructed from the similarity measurement, using fast search methods in time series. The present paper introduces a rapid method for finding analogs. Specifically, we intend to: (i) simplify the leading algorithms for similarity searching and (ii) present a case study with data from electricity prices in the Nordic market. To do so, Pearson's distance correlation coefficient was rewritten in simplified notation. This new metric was implemented in the main similarity search algorithms, namely: Brute Force, JustInTime, and Max. Next, the results were compared to the Euclidean distance approach. Pearson's correlation, as an instrument for detecting similarity patterns in time series, has shown promising results. The present study provides innovation in that Pearson's distance correlation notation can reduce the computational time of similarity profiles by an average of 17.5%. It is worth noting that computational time was reduced in both short and long time series. For future research, we suggest testing the impact of other distance measurements, e.g., Cosine correlation distance and Manhattan distances.

Published under license by AIP Publishing. <https://doi.org/10.1063/5.0021557>

**1 INTRODUCTION**

The construction of predictive models is gaining prominence in the literature (Geisser, 2017), since economic agents deal with uncertainty and aim to achieve the best results using available resources (Choi, 1993). Therefore, developing models with acceptable accuracy presents a meaningful challenge to researchers. George Box stated, "All models are wrong, but some are useful" (Box, 1976). In other words, prediction is a technique that deals with risk, and there will always be a fundamental error associated with it. The best model is the one that most adequately represents the phenomenon of interest.

In relation to the object of our study, electricity prices, there are several forecasting applications: (i) classical time series models like the autoregressive moving average, autoregressive integrated moving average and generalized autoregressive conditional heteroscedastic, among others (Liu and Shi, 2013); (ii) pre-processing techniques like spectrum analysis, wavelets, and Fourier analysis (Miranian, Abdollahzade, and Hassani, 2013); and (iii) machine learning approaches such as neural networks, fuzzy systems, and support vector machine (Bui et al., 2016). Additionally, an alternative class known as hybrid models aims to combine machine learning representations with different methods. Instances of these methods are focused time-delay neural networks (Chen et al., 2019), neural networks with fuzzy inputs (Liu et al., 2015), finite impulse response neural networks (Pir, Shah, and Aqger, 2017), local feedback dynamic fuzzy neural networks (Nagaraja et al., 2016), type recurrent fuzzy networks (Jain et al., 2014), and neuro-fuzzy inference systems (Moreno and dos Santos Coelho, 2018), among others.

The energy market is known for being an industry with high-frequency data (Madadi et al., 2018), for several reasons. First, sensor usage is widespread in energy (Jaradat et al., 2015). Second, high-frequency data can better represent specific weather conditions, enabling the improvement of energy modeling (Aqger et al., 2007). Examples are diverse, such as: (i) solar radiation, which can be collected in minutes (Assmann, Escobedo, and Oliveira, 2003) and (ii) air humidity, atmospheric pressure, temperature, and wind speed, which can also be measured in minutes (Longman et al., 2018).

In particular, the pricing of electricity also has significant volumes of information, in most cases, arranged on an hourly scale (Veronin and Partanen, 2014). Although the literature on this question is extensive, there is academic interest in the construction of nonparametric models applied to electricity prices, as they have presented promising predictive results. In general, these models are designed to deal with long-time series and are chiefly based on analog ensemble (AnEn) searches (Yang and Alessandrini, 2019; Yang, Wu, and Kleissl, 2019)

2015), finite impulse response neural networks (Pir, Shah, and Aqger, 2017), local feedback dynamic fuzzy neural networks (Nagaraja et al., 2016), type recurrent fuzzy networks (Jain et al., 2014), and neuro-fuzzy inference systems (Moreno and dos Santos Coelho, 2018), among others.

The energy market is known for being an industry with high-frequency data (Madadi et al., 2018), for several reasons. First, sensor usage is widespread in energy (Jaradat et al., 2015). Second, high-frequency data can better represent specific weather conditions, enabling the improvement of energy modeling (Aqger et al., 2007). Examples are diverse, such as: (i) solar radiation, which can be collected in minutes (Assmann, Escobedo, and Oliveira, 2003) and (ii) air humidity, atmospheric pressure, temperature, and wind speed, which can also be measured in minutes (Longman et al., 2018).

In particular, the pricing of electricity also has significant volumes of information, in most cases, arranged on an hourly scale (Veronin and Partanen, 2014). Although the literature on this question is extensive, there is academic interest in the construction of nonparametric models applied to electricity prices, as they have presented promising predictive results. In general, these models are designed to deal with long-time series and are chiefly based on analog ensemble (AnEn) searches (Yang and Alessandrini, 2019; Yang, Wu, and Kleissl, 2019)

## Chapter 6

### IEEE Latin America Transactions

<https://latamt.ieeer9.org/index.php/transactions/article/view/6948>




**Title:** Dynamic Time Scan Forecasting: A Benchmark with M4 Competition Data.

**Date:** 2023

320

IEEE LATIN AMERICA TRANSACTIONS, VOL. 21, NO. 2, FEBRUARY 2023

# Dynamic Time Scan Forecasting: A Benchmark With M4 Competition Data

Rodrigo Barbosa de Santis , Tiago Silveira Gontijo , *Reviewer, IEEE* and Marcelo Azevedo Costa , *Reviewer, IEEE*

**Abstract**—Univariate forecasting methods are fundamental for many different application areas. M-competitions provide important benchmarks for scientists, researchers, statisticians, and engineers in the field, for evaluating and guiding the development of new forecasting techniques. In this paper, the Dynamic Time Scan Forecasting (DTSF), a new univariate forecasting method based on scan statistics, is presented. DTSF scans an entire time series, identifies past patterns which are similar to the last available observations and forecasts based on the median of the subsequent observations of the most similar windows in past. In order to evaluate the performance of this method, a comparison with other statistical forecasting methods, applied in the M4 competition, is provided. In the hourly time domain, an average sMAPE of 12.9% was achieved using the method with the default parameters, while the baseline competition – the simple average of the forecasts of Holt, Damped, and Theta methods – was 22.1%. The method proved to be competitive in longer time series, with high repeatability.

**Index Terms**—Univariate methods, M4 competition, benchmarking, dynamic time scan forecasting.

## I. INTRODUCTION

The development of predictive models is widely debated in the literature [1]–[4], since it assists the control of

[21]. The 5th edition took place in 2020, and focused on a retail sales application with 42,850 unit sales hierarchical series, with the objective to produce the most accurate point forecast as well as the most accurate estimation of the uncertainty of these forecasts [22]. The 6th competition will take place this year and it will focus on predicting the overall market returns of individual stocks [23].

Whereas most well-known forecasting methods are based on identifying intrinsic components of the time series, such as level, trend, or seasonality, a particular group of methods based on similarity searches have been arousing interest in the areas of meteorology and renewable energy [24], [25]. These methods consist of identifying past weather patterns (“analog”) that closely resemble the current state. These methods are capable of handling lengthy historical time series in order to produce accurate and interpretive forecasts.

Among these methods, Dynamic Time Scan Forecasting (DTSF) consists of a new and simple analog-based forecasting technique [26]. It generates forecasts based on similar patterns, those with the highest R2 scores, calculated from the last available window.

The accuracy of analog-based methods is scarcely reported

## Chapter 7 (part a)

**ICREN 2020 | 3rd International Conference on Renewable Energy**

<https://premc.org/conferences/icren-renewable-energy/>

**Title:** Electricity price forecasting on electricity spot market: a case study based on the Brazilian Difference Settlement Price

**Date:** 25-27 November 2020



## Chapter 7 (part b)

### E3S Web of Conferences

<https://doi.org/10.1051/e3sconf/202123900002>

**Title:** Electricity price forecasting on electricity spot market: a case study based on the Brazilian Difference Settlement Price

**Date:** 2021

E3S Web of Conferences **239**, 00002 (2021)  
ICREN 2020

<https://doi.org/10.1051/e3sconf/202123900002>

### Electricity price forecasting on electricity spot market: a case study based on the Brazilian Difference Settlement Price

Tiago Silveira Gontijo<sup>1</sup>, Marcelo Azevedo Costa<sup>1,2</sup>, Rodrigo Barbosa de Santis<sup>1,\*</sup>

<sup>1</sup> Graduate Program in Industrial Engineering - Universidade Federal de Minas Gerais - Av. Antônio Carlos 6627, 31270-901 Belo Horizonte, MG, Brazil.

<sup>2</sup> Department of Industrial Engineering - Universidade Federal de Minas Gerais - Av. Antônio Carlos 6627, 31270-901 Belo Horizonte, MG, Brazil.

**Abstract.** Developing predictive models is a complex task since it deals with the uncertainty and the stochastic behavior of variables. Specifically concerning commodities, accurately predicting their future prices allows us to minimize risks and establish more reliable decision support mechanisms. Although the discussion on this question is extensive, there is academic attention being paid to the construction of nonparametric models applied to energy markets, as they have presented promising predictive results, what justifies the present study. This paper applies classical statistical models and Dynamic Time Scan Forecasting (DTSF) to the short-term electricity market prices, in Brazil, from 2006 to 2019. DTSF consists of scanning a time series and then identifying past patterns (so-called "matches"), similar to the last available observations. We predict Brazilian electricity spot prices, according the most similar matches, using aggregation functions, such as median. Recent research on the electricity spot market is increasing, indicating research significance. Our predictive approach exhibited greater accuracy than seminal statistical models. Our approach was designed for a high frequency series. Its predictive performance remained robust when other models presented both high predictive errors (spring), as well as when those models are highly accurate (winter). For future research, we recommend a more finely-tune study on DTSF parameters.



## Chapter 9 (part a)

### SBPO 2020 – LII Simpósio Brasileiro de Pesquisa Operacional

<https://proceedings.science/proceedings/100144/authors/462161>

**Title:** Forecasting hierarchical time series in power generation: modeling and analysis in Brazil

**Date:** 2-5 November 2020



## Chapter 9 (part b)

**Energies** - ISSN: 1996-1073

<https://doi.org/10.3390/en13143722>

**Title:** Forecasting hierarchical time series in power generation: modeling and analysis in Brazil

**Date:** 20 July 2020



Article

### Forecasting Hierarchical Time Series in Power Generation

Tiago Silveira Gontijo <sup>1,\*</sup> and Marcelo Azevedo Costa <sup>1,2</sup>

<sup>1</sup> Graduate Program in Industrial Engineering, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, Belo Horizonte 31270-901, MG, Brazil; azevedo@est.ufmg.br

<sup>2</sup> Department of Industrial Engineering, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, Belo Horizonte 31270-901, MG, Brazil

\* Correspondence: tsgontijo@hotmail.com

Received: 26 June 2020; Accepted: 15 July 2020; Published: 20 July 2020



**Abstract:** Academic attention is being paid to the study of hierarchical time series. Especially in the electrical sector, there are several applications in which information can be organized into a hierarchical structure. The present study analyzed hourly power generation in Brazil (2018–2020), grouped according to each of the electrical subsystems and their respective sources of generating energy. The objective was to calculate the accuracy of the main measures of aggregating and disaggregating the forecasts of the Autoregressive Integrated Moving Average (ARIMA) and Error, Trend, Seasonal (ETS) models. Specifically, the following hierarchical approaches were analyzed: (i) bottom-up (BU), (ii) top-down (TD), and (iii) optimal reconciliation. The optimal reconciliation models showed the best mean performance, considering the primary predictive windows. It was also found that energy forecasts in the South subsystem presented greater inaccuracy compared to the others, which signals the need for individualized models for this subsystem.

**Keywords:** power generation; electrical subsystems; time series

#### 1. Introduction

The advent of Industry 4.0 revolutionized factories worldwide, since it allowed the connectivity between measuring machines and the automation of companies, distributing the capacity to collect massive volumes of data [1]. In high-level data analysis, forecasting models allow the extraction of behavior patterns, as well as the prediction of future values for the collected data set [2].

In the above-mentioned scenario, the construction of predictive models is gaining prominence in the literature [3–5], since economic agents deal with uncertainty in multiple spheres and aim to achieve the best results using available resources [6]. Developing acceptably accurate models presents a meaningful challenge, as prediction is a technique that deals with risk and there will always be a fundamental error associated with it. The best model is the one that most adequately represents the phenomenon of interest.

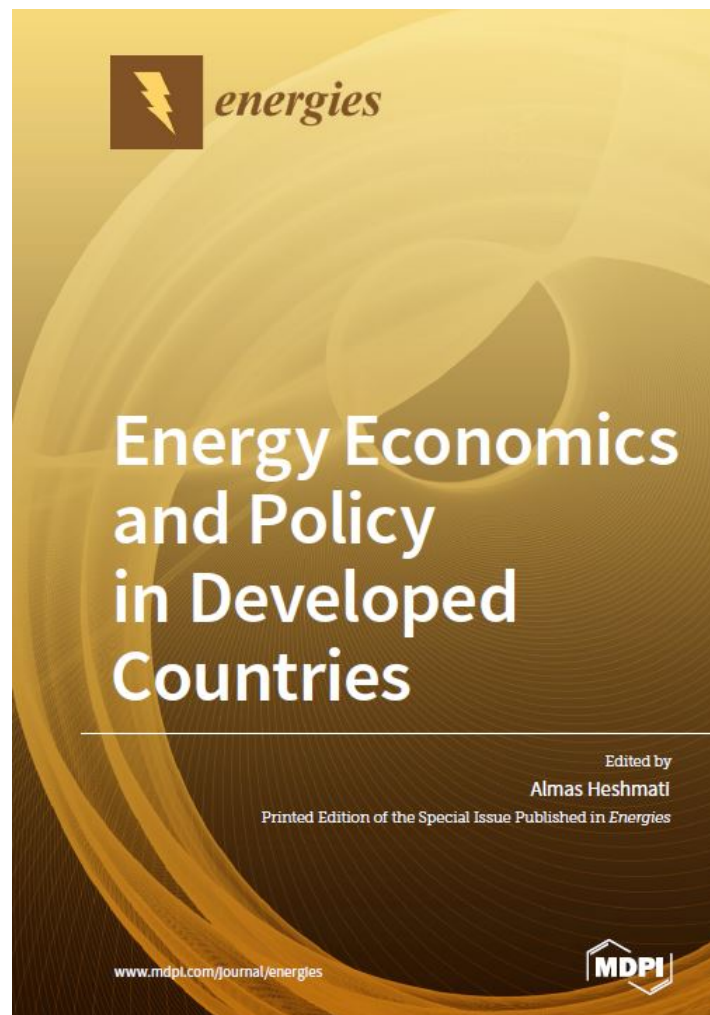
## Chapter 9 (part c)

**MDPI Books:** ISBN 978-3-03943-246-2 (Hbk); ISBN 978-3-03943-247-9 (PDF)

<https://doi.org/10.3390/books978-3-03943-247-9>

**Title:** Energy Economics and Policy in Developed Countries - Chapter 3

**Date:** October 2020





# Annex

# ANNEX A – Author profile

## A.1 Brief curriculum vitae

Doctoral student, with a master, and bachelor's in production engineering. Bachelor's degree in economics and in administration. Ten years of experience in teaching, including academic management. I developed activities at UFOP, UFSJ, PUC-MG, UNIBH, UNA, and Izabela Hendrix. I worked at the Regulatory Agency for Water Supply in the state of Minas Gerais. I held internships in Brasília-DF: Ministry of Justice/Cade; Viçosa-MG: UFV Bioenergia Project, and Belo Horizonte-MG: Chistiano Ottoni Foundation. I have experience in data analysis and applied statistics, focusing on biostatistics, data science, renewable energies, and time series.

## A.2 Selected publications

### During the Doctorate - Journals Articles

- (1) 2023 - Dynamic Time Scan Forecasting: A Benchmark With M4 Competition Data (De Santis, Gontijo, & Costa, 2023).
- (2) 2022 - A Data-Driven Framework for Small Hydroelectric Plant Prognosis Using Tsfresh and Machine Learning Survival Models (de Santis, Gontijo, & Costa, 2022b).
- (3) 2022 - Condition-based maintenance in hydroelectric plants: A systematic literature review (de Santis, Gontijo, & Costa, 2022a).
- (4) 2022 - Robust stochastic frontier analysis applied to the Brazilian electricity distribution benchmarking method (Campos, Costa, Gontijo, & Lopes-Ahn, 2022).
- (5) 2022 - Efeitos da incorporação de julgamentos na avaliação da eficiência de clubes de futebol: uma abordagem por Data Envelopment Analysis (de Cássio Rodrigues, Gontijo, Gonçalves, & Pereira, 2022).
- (6) 2021 - Os determinantes da eficiência na Atenção Primária à Saúde dos municípios paulistas: um modelo georreferenciado (Gontijo & Reis, 2021).
- (7) 2021 - Eficiência do gasto público em atenção primária em saúde nos municípios do Rio de Janeiro, Brasil: escores robustos e seus determinantes (Rodrigues, Gontijo, & Gonçalves, 2021).
- (8) 2021 - Electricity price forecasting on electricity spot market: a case study based on the Brazilian Difference Settlement Price (Gontijo et al., 2021).
- (9) 2020 - Similarity search in electricity prices: An ultra-fast method for finding analogs (Gontijo, Costa, & de Santis, 2020).
- (10) 2020 - Forecasting hierarchical time series in power generation (Silveira Gontijo & Azevedo Costa, 2020).
- (11) 2020 - O valor do projeto de uma mina de ouro: uma análise comparativa pelos modelos de fluxo de caixa descontado e de opções reais (de Cássio Rodrigues, Gontijo, & de Almeida, 2020).

- (13) 2020 Analysis of olive oil market volatility using the arch and garch techniques (Gontijo, de Cássio Rodrigues, De Muylder, Falce, & Pereira, 2020).
- (13) 2019 - A Bibliometric study on Industry 4.0 (Gontijo & Alves, 2019).
- (14) 2019 - Measuring the technical and scale efficiency of Rio de Janeiro samba schools: a DEA approach (de Cássio Rodrigues, Gontijo, & De Muylder, 2019).
- (15) 2019 - A two-stage DEA model to evaluate the efficiency of countries at the Rio 2016 Olympic Games (de Cássio Rodrigues, Gonçalves, & Gontijo, 2019).
- (16) 2019 - Incorporando julgamentos de especialistas em educação na avaliação da eficiência de cursos de graduação: uma abordagem por data envelopment analysis (de Cássio Rodrigues & Gontijo, 2019).

### **Before the Doctorate - Journals Articles**

- (17) 2018 - Um estudo sobre as causas que geram a indisponibilidade no processo de fabricação de peças automotivas (de Azevedo, Gontijo, Victor, de Souza, & de Oliveira, 2018).
- (18) 2018 - Eficiência das unidades do CEFET-MG: uma avaliação por data envelopment analysis (de Cássio Rodrigues, De Muylder, & Gontijo, 2018).
- (19) 2018 - Aplicativos educacionais na Engenharia de Produção: o caso do Enade Nota 10 (de Cássio Rodrigues, de Azevedo, et al., 2018).
- (20) 2018 - O impacto da política nacional de resíduos sólidos nas publicações científicas sobre Logística Reversa (Dutra, De Souza, Gontijo, de Cássio Rodrigues, & de Matos Andrade, 2018).
- (21) 2018 - Um Estudo Sobre a Geração de Relatórios de Manutenção Através do Uso de Smartphones (de Azevedo et al., 2018).
- (22) 2018 - Estudo de caso em uma trefilaria: proposta de redução da perda de maior representatividade (de Barros, Teixeira, & Gontijo, 2018).
- (23) 2018 - Incorporating managed preferences in the evaluation of public organizations efficiency: a DEA approach (Gontijo, De Muylder, & de Cássio Rodrigues, 2018).
- (24) 2017 - Consumo industrial de energia elétrica: um estudo comparativo entre métodos preditivos (Gontijo, Rodrigues, de Cássio Rodrigues, da Silva, & de Azevedo, 2017).
- (25) 2017 - Um estudo de caso sobre o impacto das restrições médicas nos custos ergonômicos escolares de um município (Gontijo, dos Santos, Franco, & de Azevedo, 2017).
- (26) 2017 - Otimização da produtividade através da redução do tempo de setup em terminais de cartão de crédito (E. G. Santos, Calipo, & Gontijo, 2017).
- (27) 2017 - Habilidades, competências e o perfil do profissional de Engenharia de Produção no sudeste brasileiro (de Azevedo & Gontijo, 2017).
- (28) 2017 - Avaliação da eficiência na construção civil: um estudo de caso em uma obra da região metropolitana de Belo Horizonte (Paulino, Mendonça, de Azevedo, Gontijo, & Casagrande, 2017).
- (29) 2017 - Gestão de estoque em uma microempresa do ramo alimentício: Comparação entre a Curva ABC e o Método XYZ (Catarino, Santos, Gontijo, & Rodrigues, 2017).
- (30) 2016 - A eficiência no processo de impressão em uma indústria de embalagens plásticas da região metropolitana de Belo Horizonte (Duarte Filho, de Souza, & Gontijo, 2016).
- (31) 2015 - A utilização de ferramentas da qualidade em uma empresa de manutenção de equipamentos eletromédicos (da Silva, de Matos, & Gontijo, 2015).
- (32) 2011 - Análise da volatilidade do retorno da commodity dendê: 1980-2008 (Gontijo, Fernandes, & Saraiva, 2011).