

PABLO NABIL VILLAR BOU ASSI

**PREDIÇÃO DO PREÇO SPOT DA ENERGIA
ELÉTRICA NO MERCADO LIVRE BRASILEIRO**

São Paulo
2022

PABLO NABIL VILLAR BOU ASSI

**PREDIÇÃO DO PREÇO SPOT DA ENERGIA
ELÉTRICA NO MERCADO LIVRE BRASILEIRO**

Trabalho apresentado à Escola Politécnica
da Universidade de São Paulo para ob-
tenção do Título de Engenheiro Mecatrônico.

São Paulo
2022

PABLO NABIL VILLAR BOU ASSI

**PREDIÇÃO DO PREÇO SPOT DA ENERGIA
ELÉTRICA NO MERCADO LIVRE BRASILEIRO**

Trabalho apresentado à Escola Politécnica
da Universidade de São Paulo para ob-
tenção do Título de Engenheiro Mecatrônico.

Orientador:

Marcos de Sales Guerra Tsuzuki

São Paulo
2022

Dedico este Trabalho de Conclusão de Curso a Nabil Kemil Bou Assi, meu avô que não pode concluir sua graduação em engenharia por conta de um trágico acidente. Espero encontrá-lo, pois Cristo verdadeiramente ressuscitou.

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus.

Agradeço à minha família. Todo sucesso que eu possa ter, devo ao meu pai Glauco, à minha mãe Marisol, e ao meu irmão Nicolás. Foram eles que sofreram as minhas dores, e comemoraram as minhas conquistas, nesta trajetória. Sem eles nada do que eu fiz até aqui, seria possível. Sou grato pelo carinho, incentivo e orações das minhas avós Neusa e Lola, pelo carinho e atenção que tive quando criança do meu avô Joaquim.

Muito obrigado, minha querida Heloísa Youssef, por estar comigo, pelo incentivo com doces palavras nos dias mais difíceis, por ser minha fonte principal de motivação ao longo deste ano.

Gostaria também de agradecer aos meus colegas politécnicos. E em especial, ao meu grande amigo Igor Casatti, quem sempre me ajudou quando precisei. Da primeira à última semana de Escola Politécnica, temos histórias a contar. E enquanto minha memória permitir, contarei estas histórias com muita alegria.

Agradeço ao meu professor Marcos Tsuzuki, por sua disposição e empenho em conduzir este trabalho, em um tema que eu tenho particular interesse. As sugestões e orientações feitas semanalmente, foram fundamentais para a conclusão desta monografia.

Por fim, este trabalho encerra minha formação como engenheiro mecatrônico pela Escola Politécnica da Universidade de São Paulo, sou grato à sociedade paulista por financiar esta instituição, e conseqüentemente minha graduação. Me comprometo a buscar sempre retribuir este privilégio que me foi oferecido. Agradeço a todos os professores e funcionários da Escola Politécnica, por seu empenho e excelência.

“Sempre faça tudo com muito amor e com muita fé em Deus, que um dia você chega lá. De alguma maneira você chega lá.”

-- Ayrton Senna

RESUMO

Este trabalho tem como objetivo principal a predição do Preço de Liquidação das Diferenças (PLD) para o submercado sudeste, em um intervalo crescente de 1 a 7 dias, por meio de modelos preditores disponíveis em bibliotecas de código aberto, a partir de dados públicos. O PLD é o preço spot (à vista, de curto prazo) da energia elétrica no Ambiente de Contratação Livre (ACL), ou mercado livre brasileiro de energia elétrica, onde os agentes do mercado (geradores, comercializadores e consumidores) podem firmar contratos particulares de compra e venda de energia elétrica. Para a predição do PLD, foram selecionadas inicialmente como variáveis preditoras: a própria série do PLD, proveniente da Câmara de Comercialização de Energia Elétrica (CCEE), além das séries de Energia Natural Afluente (ENA), Energia Armazenada (EAR) e carga, provenientes do Operador Nacional do Sistema (ONS). Com a revisão da literatura, observou-se a maior ocorrência e melhores resultados dos modelos ARIMA e Redes Neurais LSTM, para a predição de séries temporais em aplicações a sistemas elétricos. Como modelo comparativo referencial foi utilizado o método Naïve, dado pela replicação do último valor conhecido da série para as predições. As métricas de avaliação utilizadas foram Erro Absoluto Médio (MAE), Erro Percentual Absoluto Médio (MAPE) e Raiz Quadrada dos Erros Quadráticos Médios (RMSE). Não foi possível utilizar o modelo ARIMA para predição do PLD devido a não estacionariedade da série, em contraponto, foi demonstrada a boa capacidade preditiva deste modelo para uma série em que é possível obter estacionariedade como a série de EAR. A partir da análise das correlações entre as séries e da Análise de Componente Principal (PCA) houve a suspeita de que a série de carga pudesse ser eliminada sem perda da capacidade preditiva do modelo Rede Neural LSTM. Esta hipótese foi confirmada ao testar a predição de um dia à frente com e sem a variável de carga para entrada dos modelos: o teste em que a série de carga foi retirada obteve melhores resultados. Uma terceira tentativa de refino foi a correta eliminação dos períodos de saturação da série do PLD para o treinamento da rede, porém não houve melhoras dos resultados, indicando que este não é um tratamento necessário para o modelo. Nenhum destes três testes de predição de um dia à frente com Redes Neurais LSTM obteve melhores resultados que o método Naïve. Já para a predição de mais de um dia à frente, foram implementados modelos em dois métodos distintos, *single-step* e *multi-step*. O melhor método foi o *multi-step* que apesar de apresentar uma dinâmica crescente dos MAPEs em função do número de passos preditos, para mais de um dia à frente de predição obteve resultados melhores que o método Naïve. Ao término do trabalho, os objetivos foram atingidos.

Palavras-Chave – Mercado Livre de Energia Elétrica, Predição de Preço, Redes Neurais LSTM

ABSTRACT

The main objective of this project is the forecasting of the Settlement Price of Differences (PLD) for the Southeast submarket, in an increasing forecasting range of 1 to 7 days, through models available in open source libraries, using public data. The PLD is the spot price (short-term price) of electricity in the Environment of Free Contracting (ACL), or Brazilian free market of electric energy (also known as deregulated market), where the market agents (generators, traders and consumers) can sign power purchase agreements. For the PLD forecasting, it was initially selected as variables, the time series of the PLD, extracted from the Chamber of Electric Energy Commercialization (CCEE), the time series of the Natural Affluent Energy (ENA), the time series of Stored Energy (EAR), and the time series of the power consumption, extracted from the National System Operator (ONS). Through the literature review, it was observed the highest occurrence and better results of ARIMA models and LSTM Neural Networks, for the forecasting applications in electrical systems. The Naïve forecasting method was used as a comparative model, it is given by replicating the last known value of the time series as forecasted values. The evaluation measures selected were the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE). It was not possible to use the ARIMA model for PLD forecasting due to the non-stationarity of the series, in contrast, the forecasting capacity of this model demonstrated by the EAR forecasting, with some transformation of this time series it was possible to obtain stationarity. From the analysis of correlations and the Principal Component Analysis (PCA), there was a hypothesis that the power consumption time series could be eliminated without loss of forecasting capability of the LSTM Neural Network model. This hypothesis was confirmed by testing the one-day-ahead forecast with and without the power consumption variable as input: the test without the power consumption as variable obtained better results. A third refinement attempt was the correct elimination of saturation periods from the PLD time series for training, but there were no improvements in the results, indicating that this is not a necessary treatment for the model. None of these three prediction tests of a day ahead with LSTM Neural Networks performed better than the Naïve forecasting method. For the forecast of more than one day ahead, models were implemented in two different methods, single-step and multi-step. The best method was the multi-step method, which, despite of the MAPEs increasing trend while increasing number of forecasted steps, for more than one day ahead, obtained better results than the Naïve forecasting method. Finally, at the end of the work, the objectives were achieved.

Keywords – Deregulated Electricity Market, Price Forecasting, LSTM Neural Networks

LISTA DE FIGURAS

1	Matrizes elétricas mundial e brasileira	12
2	Decisão sob incerteza da operação centralizada de um sistema majoritaria- mente hidrotérmico	13
3	Número de documentos por ano relacionados a pesquisa "Electricity price forecasting" no portal <i>Scopus</i>	17
4	Representação de célula neural LSTM	27
5	Representação do funcionamento de rede neural LSTM	28
6	Média diária do PLD para o submercado Sudeste	30
7	a) Série da soma da Energia Natural Afluente para os 4 submercados do SIN b) Série da soma da Energia Armazenada para os 4 submercados do SIN c) Série da soma da Carga para os 4 submercados do SIN	31
8	a) Série da soma da Carga para os 4 submercados do SIN em março de 2021 b) Série da média móvel da soma da Carga para os 4 submercados do SIN	32
9	Matriz de correlação	32
10	Variância percentual acumulada para soma de cada componente principal com as componentes anteriores	33
11	Distribuição dos pontos em função das componentes principais	34
12	Gráfico da série do PLD logaritmicamente transformada e diferenciada	35
13	Gráfico da série do PLD transformada realizada e da predição com modelo ARIMA	36
14	Gráfico da série do PLD realizado e da predição com modelo ARIMA	36
15	Gráfico da série de EAR logaritmicamente transformada e diferenciada	37
16	Gráfico da série de EAR transformada realizada e da predição com modelo ARIMA	37
17	Gráfico da série de EAR realizada e da predição com modelo ARIMA	38

18	Gráfico do PLD realizado e das previsões em janela móvel modelo LSTM com 4 variáveis de entrada	39
19	Gráfico do PLD realizado e das previsões em janela móvel modelo LSTM com 3 variáveis de entrada	40
20	Esquema da função de eliminação das saturações	40
21	Gráfico do PLD realizado e das previsões em janela móvel modelo LSTM com 3 variáveis de entrada e eliminação das saturações do treinamento . .	41
22	Esquema dos conjuntos de entrada e saída para treinamento dos modelos <i>single-step</i>	42
23	a) Previsões em janela móvel modelo 2 b) Previsões em janela móvel modelo 3	42
24	a) Previsões em janela móvel modelo 4 b) Previsões em janela móvel modelo 5	43
25	a) Previsões em janela móvel modelo 6 b) Previsões em janela móvel modelo 7	43
26	Esquema da função de reorganização das previsões dos modelos do método <i>single-step</i>	44
27	Gráfico previsões método <i>single-step</i>	44
28	Esquema do conjunto de entrada e saída para treinamento do modelos <i>multi-step</i> para previsão de 7 dias à frente	45
29	Gráfico do PLD realizado e das previsões 1, 8 e 15 do modelo <i>multi-step</i> para 3 dias à frente	46
30	Gráfico do PLD realizado e das previsões 1, 7 e 13 do modelo <i>multi-step</i> para 5 dias à frente	46
31	Gráfico do PLD realizado e das previsões 1, 6 e 11 do modelo <i>multi-step</i> para 7 dias à frente	47
32	Erro médio percentual (MAPE) em função do passo de previsão para cada modelo	52

LISTA DE TABELAS

1	MAPE e SMAPE para as previsões de Aguiar, Oliveira e Souza	21
2	MAPE das previsões por LSTM de cada uma das séries de Santos	22
3	MAPE de cada modelo utilizado por Lopes	23
4	Dados publicados pela ONS e CCEE	25
5	Componentes principais da análise para 3 componentes	33
6	Períodos de saturação da média diária do PLD	35
7	Métricas de avaliação dos modelos <i>single-step</i>	43
8	Métricas de avaliação dos modelos <i>multi-step</i>	47
9	Comparação dos modelos para previsão de um dia à frente	50
10	Comparação do modelo <i>multi-step</i> para previsão de 3 dias à frente com método Naïve	50
11	Comparação do modelo <i>multi-step</i> para previsão de 5 dias à frente com método Naïve	50
12	Comparação do modelo <i>multi-step</i> para previsão de 7 dias à frente com método Naïve	51

SUMÁRIO

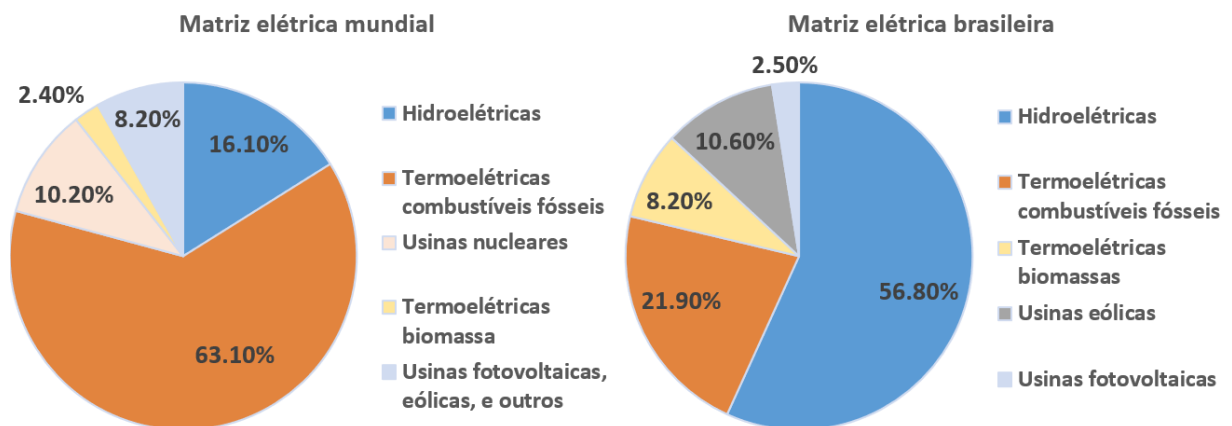
1	Introdução	12
1.1	Mercado livre brasileiro	13
1.2	Motivação	15
1.3	Objetivos gerais	16
2	Revisão da literatura	17
2.1	Frequência de publicações sobre o tema	17
2.2	Primeiros trabalhos	18
2.3	Anos recentes	19
2.4	Aplicações ao sistema elétrico brasileiro	21
2.5	Considerações	23
3	Metodologia	24
3.1	Requisitos	24
3.2	Seleção das variáveis	25
3.3	Modelos de predição	25
3.3.1	ARIMA	26
3.3.2	Redes Neurais LSTM	26
3.3.3	Método Naïve	29
3.3.4	Métricas de avaliação	29
4	Desenvolvimento e resultados	30
4.1	Tratamento, análise e seleção das variáveis	30
4.1.1	Análise de Componente Principal (PCA)	32
4.1.2	Saturações	34

4.2	Predição com modelo ARIMA	35
4.3	Predição do PLD um dia à frente com modelo LSTM	38
4.3.1	Resultado predição com 4 variáveis	39
4.3.2	Resultado predição com 3 variáveis	39
4.3.3	Resultado predição excluindo saturações do treinamento	40
4.4	Predição do PLD para mais de um dia à frente com modelo LSTM	41
4.4.1	Método <i>single-step</i>	42
4.4.2	Método <i>multi-step</i>	45
5	Discussão	48
5.1	Seleção de variáveis	48
5.2	Modelo preditivo	49
5.2.1	Comparação dos métodos	49
5.2.2	Dinâmica do erro médio percentual	52
5.3	Comparação com a literatura	53
6	Conclusões e Trabalhos Futuros	55
6.1	Trabalhos futuros	56
	Referências	58

1 INTRODUÇÃO

A definição do preço da energia elétrica é uma tarefa complexa no sistema elétrico brasileiro. Apesar da matriz elétrica brasileira estar em transformação, com o crescimento de novas fontes de energia renováveis, como a energia solar e eólica, predominantemente a geração elétrica brasileira ainda é hidráulica [1]. Esta é uma particularidade da matriz brasileira quando comparada com as matrizes dos demais países, como pode ser observado na Figura 1.

Figura 1: Matrizes elétricas mundial e brasileira



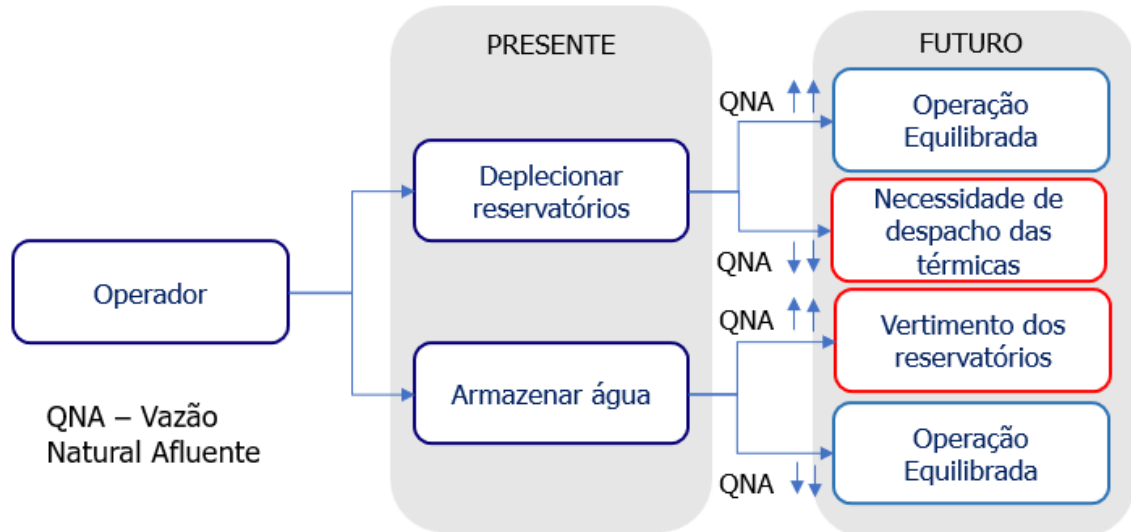
Fonte: adaptado de Empresa de Pesquisa Energética (EPE) [1].

No Brasil, o planejamento do despacho de energia deve levar em consideração as diferentes usinas do sistema, para tentar otimizar o uso dos recursos hídricos e térmicos. Para isso, diversas variáveis são importantes, como a interconexão de diferentes usinas hidrelétricas na mesma bacia (fazendo com que o despacho de uma usina, afete o reservatório de outra usina a jusante da primeira), a demanda (carga) do sistema, a vazão afluente, a disponibilidade dos reservatórios, entre outras variáveis [2].

A responsabilidade pelo planejamento e administração do Sistema Integrado Nacional (SIN), como é nomeado o sistema interligado de geração e transmissão de energia elétrica

brasileiro, é do Operador Nacional do Sistema Elétrico (ONS). O operador diariamente deve tomar a decisão entre deplecionar os reservatórios hídricos, ou poupá-los para o futuro [3], como ilustra a Figura 2.

Figura 2: Decisão sob incerteza da operação centralizada de um sistema majoritariamente hidrotérmico



Fonte: adaptado de Yasmin Monteiro Cyrillo [3].

Apesar de haver uma operação centralizada do SIN, efetivamente a geração, transmissão, comercialização e consumo da energia ocorre por meio de diversos agentes. Ao tratar do setor elétrico, a ideia mais comum é pensar no consumo residencial, que se dá por meio das distribuidoras, com preços e tarifas reguladas pelo Estado. Porém há também o mercado livre de energia elétrica, que será apresentado a seguir.

1.1 Mercado livre brasileiro

Abordando brevemente o histórico do setor elétrico nacional, no início do século XX, a demanda por energia elétrica no país ainda era inexpressiva, e concentrava-se nas capitais e cidades-polo atreladas a economia do café. Esse início do setor elétrico nacional, deu-se com a iniciativa privada, pela instalação da companhia canadense Light e da companhia norte-americana Amforp [4].

No final da década de 20, com a crise gerada pela quebra da bolsa de Nova York, a demanda por energia elétrica diminuiu, e o investimento no setor, realizado pelas empresas estrangeiras, perdeu a atratividade. Com isso, os governos estaduais e federal passaram a atuar expressivamente no setor, por meio de regulações e investimento direto, com a

criação de empresas estatais. Este modelo teve sucesso até a década de 70, com a crise do petróleo ocasionando a regulação das tarifas elétricas como política de contenção da inflação, e a desvalorização da moeda, a saúde financeira das empresas foi deteriorada [5].

Na década de 90, em vista da crise do setor elétrico, e da falta de capacidade de investimento do Estado, iniciou-se programas de privatizações, concessões e permissões no setor. Este processo foi iniciado antes mesmo da criação de órgãos governamentais para regulação e operação do sistema. Até o início dos anos 2000, não houve investimento suficiente para garantir o suprimento integral da demanda, somado a ausência de um operador centralizado, resultou em crises de oferta no início do milênio, o que levou o sistema elétrico nacional a uma nova reforma no final de 2003 [5]. Finalmente, após esta reforma e as evoluções regulatórias ao longo dos últimos anos, o mercado chegou ao estágio vigente.

Atualmente, há uma divisão do mercado de energia elétrica em Ambiente de Contratação Regulado (ACR) e Ambiente de Contratação Livres (ACL). No ACR (ou mercado cativo), os consumidores compram energia elétrica da distribuidora de sua região, sem liberdade de escolha ou poder de negociação, visto que a tarifa é regulada pela Agência Nacional de Energia Elétrica (ANEEL). No ACL (ou mercado livre), os consumidores de médio e grande porte, com uma demanda elétrica superior a 500 kW, que optaram por deixar o ACL e aderiram ao ACL, firmam contratos bilaterais com geradores para a compra direta de energia elétrica. Os preços e termos dos contratos são definidos com flexibilidade, de maneira particular.

Intuitivamente, por comparação a outros mercados, tende-se a pensar que o comprador consome fisicamente a energia que é gerada pelo próprio vendedor. Porém, devido a característica de integração e interligação do SIN, o que ocorre na prática é que o vendedor deve entregar o montante de energia vendido no ponto de transmissão ou distribuição definido pela ANEEL, e o consumidor passa a ter o direito de retirar de outro ponto (usualmente, distante do ponto de entrega do gerador) o montante de energia contratado.

Quando há diferenças entre os montantes contratados, e o que de fato foi consumido ou gerado, há a possibilidade de firmar um novo contrato, ou deixar esta diferença ser liquidada ao valor do Preço de Liquidação das Diferenças (PLD), que é tido como o preço à vista (spot) da energia elétrica no mercado livre. A determinação do PLD é de responsabilidade da Câmara de Comercialização de Energia Elétrica (CCEE).

Há uma divisão vigente em quatro submercados (Norte, Nordeste, Sudeste e Sul), eles são interligados com a possibilidade de transmissão elétrica de um submercado para

outro. Cada submercado tem seu próprio PLD, mas frequentemente coincidem no valor. A metodologia que determina o PLD é baseada em três modelos computacionais de otimização, em cadeia. São eles, NEWAVE, DECOMP e DESSEM, que são usados para o planejamento elétrico nacional [6].

Além da função prática de preço de liquidação das diferenças entre o contratado e o consumido ou gerado, o PLD é também um forte balizador dos contratos particulares de compra e venda de energia. No curto prazo, todo agente do ACL (gerador, comercializador, ou consumidor), tem a opção de comprar ou vender energia elétrica ao valor do PLD, e por este motivo serve como referencial para os contratos de curto prazo. Para o longo prazo, por segurança e confiabilidade energética, é exigido que os agentes do ACL tenham lastro, isto é, parte da compra de energia deve vir de um contrato com um agente com garantia física da geração [7], não podendo utilizar somente o mecanismo de liquidação a PLD. Porém ainda assim, no geral o nível de PLD atual também é referência para contratos de longo prazo, por representar uma expectativa futura dos perfis de oferta e demanda energética.

1.2 Motivação

O mercado livre de energia brasileiro está em crescimento. Em 2021, o número de agentes (geradores, distribuidores, comercializadores e consumidores) cresceu 14%, em comparação a 2020. Ao término de 2021, participavam do mercado 1.801 geradores produtores independentes, 456 comercializadores, 9.930 consumidores, além de outros agentes (geradores autoprodutores, geradores à título de serviço público e distribuidores), totalizando 12.240 agentes [8].

A predição do preço da energia é importante para todos os agentes. Os geradores precisam de estimativas de preço a longo prazo para avaliar investimentos, e de estimativas a curto prazo, para o caso de haver diferenças entre a energia contratualmente vendida, e a de fato gerada, podendo assim estruturar melhor sua decisão. Da mesma forma, os consumidores também precisam de estimativas a curto prazo, para caso haja diferenças entre a energia contratualmente comprada, e a de fato consumida. E também, para o próprio planejamento financeiro e de produção, especialmente no caso de consumidores em que a energia elétrica representa parte significativa dos custos variáveis de seu ramo de atividade. E para as comercializadoras, evidentemente, a predição do preço é fundamental. A sua atividade pode ser resumida em negociar a compra e a venda de energia elétrica.

Além disso, com o Projeto de Lei nº 414 de 2021 (PL414/2021), o crescimento do mercado livre de energia elétrica pode ser acelerado. Discute-se no Congresso Nacional a possibilidade da abertura do mercado cativo, para que mesmo consumidores de baixa tensão (e com demanda inferior a 500kW) possam aderir ao mercado livre, o que traria mais liberdade e concorrência ao setor. Este modelo já é adotado em muitos dos países integrantes da Organização para a Cooperação e Desenvolvimento Econômico (OCDE) [9]. Possíveis mudanças apresentadas pelo PL414/2021 trariam implicações tarifárias ao setor elétrico brasileiro, e portanto ainda devem ser objeto de estudo e discussões, mas caso o pequeno consumidor possa comprar energia elétrica com regras semelhantes às vigentes no ACL, será também extremamente útil a previsibilidade do preço para este consumidor.

1.3 Objetivos gerais

Pretende-se prever o valor do PLD para o submercado Sudeste (o maior e mais expressivo submercado do SIN) em intervalos crescentes de 1 a 7 dias, por meio de modelos computacionais, utilizando dados públicos relevantes para o sistema elétrico brasileiro. Como objetivos secundários, pretende-se estudar as influências de diferentes variáveis, e diferentes métodos na capacidade preditiva dos modelos a serem utilizados.

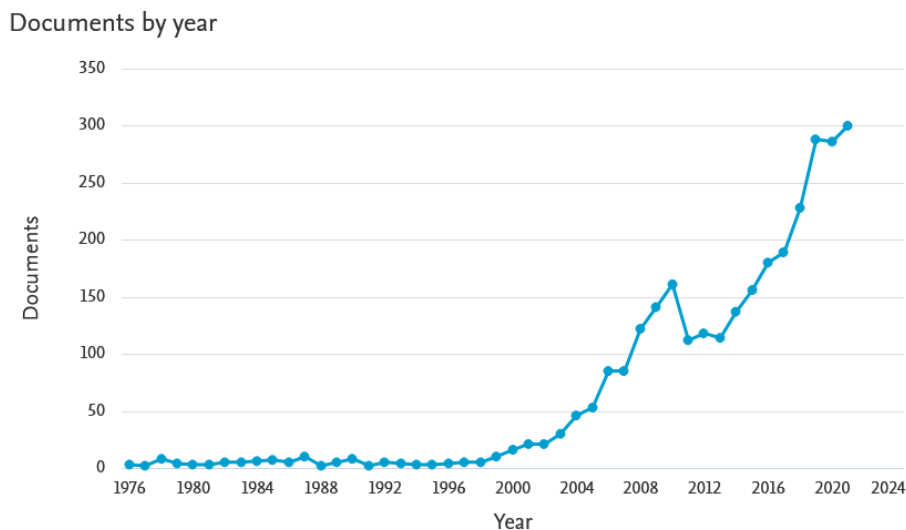
2 REVISÃO DA LITERATURA

A revisão da literatura foi dividida nas seções a seguir. Para demonstrar a relevância do tema, iniciou-se por meio de uma pesquisa do número de publicações por ano no portal *Scopus*, em seguida é feita uma breve contextualização do histórico dos primeiros trabalhos de predição no setor elétrico com diferentes métodos e modelos, depois trabalhos atuais sobre o tema, e por fim, trabalhos de predição de séries temporais no setor elétrico brasileiro.

2.1 Frequência de publicações sobre o tema

Com o desenvolvimento dos mercados livres de energia elétrica em diversos países (também chamados de mercados desregulados, ou mercados reestruturados), a predição dos preços de energia elétrica têm sido tema cada vez mais frequente em publicações. A Figura 3 mostra o número de documentos por ano, disponíveis no portal *Scopus*, relacionados a busca pelos termos "electricity price forecasting".

Figura 3: Número de documentos por ano relacionados a pesquisa "Electricity price forecasting" no portal *Scopus*



Fonte: adaptado de *Scopus* [10].

2.2 Primeiros trabalhos

O comportamento dos preços no setor energético é tema de estudo e pesquisa desde o século passado. Até a década de 1970, a maioria das publicações relacionadas ao tema tratavam de analisar os sistemas energéticos, e estimar a oferta e a demanda de combustíveis (como petróleo, gás natural, carvão mineral) e energia elétrica com base em modelos de planejamento energético e modelos econométricos [11]. Para o setor elétrico, como é sugerido pela Figura 3, e os autores observam em [12] e [13], o desenvolvimento da literatura de predição dos preços se dá a partir dos anos 2000.

Até a década de 1980, a maioria dos países adotava sistemas de geração, transmissão e distribuição monopolistas e altamente regulados, com pouca competição e liberdade de escolha para o consumidor [14], portanto a predição dos preços tinha uma aplicabilidade mais limitada em relação aos dias atuais, com mercados competitivos. Eram mais frequentes, publicações que focavam em estimar a demanda do consumo de energia elétrica, uma vez que esta variável é útil para o planejamento do crescimento dos sistemas elétricos nacionais e regionais, ainda em desenvolvimento.

Com base na publicação de Taylor de 1975 [15], que se propõe a revisar a literatura para artigos de demanda de energia elétrica, observa-se que a abordagem mais comum era modelar a demanda a partir da função demanda, clássica em estudos de econometria, em que a demanda de um bem, é função do preço deste bem, dos preços de demais bens, e da renda do consumidor. O autor consolida em uma tabela as estimativas de elasticidade preço da demanda e renda da demanda dos artigos revisados.

Em 2001, Hisao e Hong [16] publicaram um dos artigos pioneiros na aplicação de redes neurais à predição do preço da energia elétrica. O objetivo era prever os Preços Marginais Locais (Locational Marginal Prices, LMP) da PJM, uma organização regional de transmissão, responsável por coordenar a comercialização institucional (atacado) de energia elétrica em alguns estados do leste dos Estados Unidos. Vale esclarecer que nos Estados Unidos, diferentemente do Brasil, há mais de uma organização que coordena a comercialização de energia elétrica. Ainda assim, é possível considerar o LMP semelhante ao PLD no contexto do mercado brasileiro.

Assim como o PLD, o LMP tem discretização horária, a proposta de Hisao e Hong foi prever os 24 valores do LMP no dia seguinte ao dia utilizado como base. Para isso, utilizaram o modelo de rede neural recorrente, dando como entradas valores históricos do LMP, as cargas do sistema (demanda de energia), o nível de congestão do sistema

(categoricamente, o quão perto do limite de transmissão o sistema se encontra), e o fluxo de energia que entra e sai do sistema. Além disso, os autores perceberam que a série seguia determinado padrão em dias úteis, e outro padrão no sábado e no domingo, portanto os dados foram segregados em dias úteis, sábado e domingo.

Foram utilizados os dados horários disponíveis de abril a maio de 1999 (logo 1464 pontos), divididos em conjunto de treinamento, conjunto de teste e conjunto de validação. A principal métrica de avaliação dos modelos foi o Erro Percentual Absoluto Médio (MAPE), e os valores do MAPE foram 11,1% para dias úteis, 13,6% para sábados, e 16,2% para domingos. Os autores concluem que o novo método apresentado é eficiente para prever o preço da energia elétrica neste horizonte de dois dias à frente [16].

2.3 Anos recentes

Ghoddusi, Creamer, Rafizadeh (2019) [12] fazem uma revisão de aplicações de Machine Learning no mercado de energia. E especificamente para o setor de energia elétrica, as aplicações mais comuns são predição do preço e da demanda. Para a primeira aplicação, que é objetivo deste trabalho, os métodos mais usados são redes neurais, deep learning e métodos ensemble, que unem mais de um método em um único algoritmo.

Rajan, Chandrakala (2021) [17] propõe a predição do preço de energia elétrica no mercado indiano, para a região sul, por meio dos modelos estatísticos autorregressivo integrado de médias móveis (ARIMA) e autorregressivo generalizado condicional heterocedástico (GARCH). Os dados tem discretização diária, e para o conjunto de treinamento foi considerado os dados disponíveis de janeiro de 2019 a setembro de 2019, e as predições foram feitas em um horizonte de 3 meses. O MAPE do modelo ARIMA foi de 4,59% e do modelo GARCH de 63,67%.

Ugurlo, Oksus, Tas (2018) [13] aplicam 4 tipos de redes neurais para predição do preço horário de energia elétrica no mercado livre de energia da Turquia. Os modelos escolhidos foram redes multilayer perceptron, redes convolucionais, Long Short Term Memory (LSTM) e Gated Recurrent Unit (GRU), além destes, para efeito de comparação foram também aplicados modelos estatísticos como o ARIMA. As variáveis de entrada dos modelos foram 4 valores históricos do preço (atrasado 1, 23, 24, 48, 72, 168, 336 horas), predição da demanda de energia, demanda realizada atrasada 24 horas, temperatura e preço de balanceamento do mercado. Os dados de treinamento foram do período de janeiro de 2013 a dezembro de 2015, e foi feita a predição no horizonte de um dia (24 pontos),

para o ano de 2016. Os modelos foram aplicados recebendo 7 configurações diferentes de número de variáveis de entrada, desde somente 2 variáveis, até as 11 citadas. Para 4 variáveis, o modelo estatístico ARIMA obteve menor Erro Absoluto Médio (MAE) que a rede multilayer perceptron e a rede convolucional, mas LSTM e GRU obtiveram menores MAE. O menor MAE foi obtido pelo modelo GRU, com 11 variáveis de entrada. Vale destacar que com o aumento do número de variáveis de entrada, aumentaram os valores de MAE das redes multilayer perceptron e convolucionais, e diminuíram os valores de MAE das redes LSTM e GRU.

Tschora et al. (2022) [18] estudaram a aplicação de diferentes modelos de machine learning para a predição de um dia à frente, do preço da energia elétrica com discretização horária nos mercados da França, Alemanha e Bélgica. Os modelos testados foram Random Forrest, Support Vector Machine (SVM), redes neurais convolucionais, e redes neurais profundas. As variáveis preditoras utilizadas foram os preços de energia dos 3 países, além dos preços de outros países como Espanha e Suíça ao considerar as diferentes correlações, além de predições de carga e geração intermitente, e dos preços de gás natural da França. Para cada país houve dois conjuntos distintos de seleção de variáveis preditoras. As principais conclusões dos autores são a respeito das influências de cada variável na capacidade preditiva dos modelos, para cada mercado há variáveis mais ou menos relevantes, e a respeito das desigualdades de desempenho dos modelos preditivos, Random Forrest e redes neurais convolucionais poderiam ser descartadas para esta aplicação, as redes neurais profundas e SVM obtiveram os melhores desempenhos.

Miletic et al. (2022) [19] desenvolveram um trabalho cujo objetivo era a predição do preço para um dia à frente no mercado acoplado europeu (SDAC), composto por 27 países. Foram utilizadas redes neurais LSTM como principal modelo, e as variáveis de entrada foram somente a própria série histórica dos preços, e séries decorrentes da original (como médias móveis, máximos e mínimos). Os testes foram realizados para mais de uma divisão da amostra de treinamento e teste. Os autores concluem que as características da série de preços está em mudança após o período pandêmico, e que isso prejudicou a acurácia dos modelos a depender da amostra de treinamento selecionada, além disso, que os modelos se beneficiam de determinadas variáveis endógenas, como máximo e mínimo de semanas anteriores. E por fim, indicam que os modelos poderiam ser refinados com o acréscimo de variáveis exógenas, como outras séries temporais relacionadas ao mercado elétrico europeu.

2.4 Aplicações ao sistema elétrico brasileiro

Nesta subseção, foram selecionados somente artigos que propõe previsões com base em dados do sistema brasileiro de energia elétrica.

Aguiar, Oliveira, Souza (2021) [20] em seu artigo expõem os resultados das previsões da série temporal de carga horária do SIN. Os modelos escolhidos para o trabalho foram as variações de Holt-Winters (aditivo, Multiplicativo e Sazonal Duplo) e ARIMA. Especificamente, o Holt-Winters Sazonal Duplo foi desenvolvido para ser capaz de modelar tanto a sazonalidade diária quanto a sazonalidade semanal, características dos dados de carga. A amostra de treinamento foram os dados de carga horária de janeiro a junho de 2019, e a previsão foi feita no horizonte de 24 horas. Duas das métricas usadas para avaliar os desempenhos foram o MAPE e o Erro Médio Absoluto Simétrico Percentual (SMAPE), os valores são expostos na Tabela 1.

Tabela 1: MAPE e SMAPE para as previsões de Aguiar, Oliveira e Souza

Modelo	MAPE	SMAPE
Holt- Winters Aditivo	10,22%	10,86%
Holt- Winters Multiplicativo	9,57%	10,12%
Holt- Winters Sazonal Duplo	5,94%	6,17%
ARIMA	13,2%	14,4%

Fonte: adaptado de Aguiar, Oliveira, Souza [20].

Gontijo, Costa, Santis (2021) [21] fazem um estudo comparativo entre o método Dynamic Time Scan Forecasting (DTSF, implementado na linguagem R inicialmente para a previsão de energia eólica) e outros diversos modelos estatísticos de séries temporais, para prever o PLD. A amostra escolhida foi o PLD semanal de janeiro de 2006 a maio de 2019 para o submercado sudeste, em patamar de carga pesada. O modelo que obteve o menor SMAPE foi o DTSF, com 20,72%.

Nascimento, Olivi, Lima (2018) [22] propõem-se a prever o PLD semanal exclusivamente a partir do histórico da própria série temporal. Para isso, utilizam o modelo de Rede Neural Autorregressiva, que recebe n_p entradas atrasada da própria série e n_q entradas referentes às saídas temporais, além das médias e desvios padrões destes dados. Selecionaram os dados do PLD de 132 semanas de 2017, e utilizaram as primeiras 120 semanas para treinamento e as demais para validação. Foram preditos os valores em um intervalo temporal de 4, 8 e 12 semanas. A principal métrica de avaliação do modelo foi o MAE, o que impede a comparação do erro com demais artigos revisados. Porém, os

autores concluem que o modelo é capaz de generalizar o comportamento da série temporal do PLD, e que pode basear tomadas de decisões a respeito da contratação de energia a curto prazo.

Santos (2019) [23] em seu trabalho prediz o PLD semanal para cada um dos 4 submercados e cada um dos patamares de carga (leve, média, pesada). Para isso, utiliza o modelo de rede neural Long Short Term Memory (LSTM), capaz de considerar memória de longo prazo ao tratar uma série temporal. De todos os artigos revisados para predição do PLD, este é o que leva o maior número de variáveis de entradas: são consideradas as séries históricas do PLD, da carga de energia ativa (demanda em MW médios), da demanda máxima, da Energia Natural Afluente (ENA), da Energia Armazenada (EAR), da geração de energia hidroelétrica, da geração de energia termoeletrica, dos intercâmbios de energia (entre os subsistemas), e dos intercâmbios de energia internacionais. A amostra de treinamento foram os dados de junho de 2001 a junho de 2015, e a amostra de validação de julho 2015 a dezembro de 2018, e as predições foram feitas no horizonte de um passo a frente (uma semana). Os MAPEs obtidos para cada uma das séries estão expressos na Tabela 2.

Tabela 2: MAPE das predições por LSTM de cada uma das séries de Santos

-	Patamar de carga		
Subsistema	Leve	Média	Pesada
Norte	9,26%	11,19%	14,31%
Nordeste	10,34%	8,59%	9,11%
Sudeste	7,73%	6,21%	6,18%
Sul	11,04%	10,97%	11,00%

Fonte: adaptado de Santos [23].

Lopes (2022) [24] utiliza modelos de Machine Learning para prever o PLD horário do submercado Sul pelos próximos 7 dias seguintes a uma data base, a partir dos dados dos últimos 7 dias do próprio PLD, da ENA (energia natural afluente por Bacia), da EAR (energia armazenada por bacia) e da PO (precipitação observada). Os dados hidrológicos têm discretização diária, então ajustou-se os dados para discretização horária, repetindo os valores diários 24 vezes. A amostra escolhida foram os dados do início de janeiro de 2020 ao final de julho de 2021. Foram escolhidos 4 modelos, regressão linear simples, random forest, gradient boosting e redes neurais convolucionais. Os MAPEs são expostos na Tabela 3.

Tabela 3: MAPE de cada modelo utilizado por Lopes

Modelo	MAPE
Regressão Linear	38,77%
Random Forest	27,59%
Gradient Boosting	28,39%
Redes Neurais	24,70%

Fonte: adaptado de Lopes [24].

2.5 Considerações

A respeito dos modelos preditivos, observa-se que os trabalhos aplicaram diversos modelos e métodos, e há conclusões específicas para cada aplicação. Porém, de maneira geral, as redes neurais, especialmente LSTM, obtiveram bom desempenho. Também é observada, frequentemente, a aplicação de modelos estatísticos como ARIMA, para predições de séries temporais no setor elétrico.

Com base na revisão da literatura, é possível inferir que para o problema de predição do preço da energia elétrica é mais adequada uma abordagem multivariável, isto é, com mais de uma variável preditiva, além da própria série a ser predita. Aplicando ao sistema elétrico brasileiro, as principais variáveis a serem analisadas são a série histórica do PLD, as séries hidrológicas e de carga. De acordo com o levantamento apresentado na Reunião Ampla com Agentes de 2020, coordenada pela Empresa de Pesquisa Energética (EPE) [25], as variações do PLD são devidas principalmente à hidrologia, ao nível de armazenamento dos reservatórios, e ao patamar de carga do sistema.

3 METODOLOGIA

O desenvolvimento prático do trabalho se dá pela elaboração dos algoritmos baseados nos modelos de predição, testes e análise dos resultados. Para todos os modelos considerados, os dados selecionados devem ser divididos em amostra de treinamento e amostra de teste, além disso, todos terão as mesmas métricas de avaliação. Os testes serão realizados com intervalos de predição crescentes, para que seja possível observar a dinâmica dos erros de predição com o aumento do intervalo. Na revisão da literatura realizada, não foi possível encontrar, para esta aplicação (PLD), predições com intervalo maior de 7 dias e discretização diária, os artigos revisados que utilizaram um intervalo de predição maior do que 7 dias, utilizaram discretização semanal (intervalos de predição de 1 a 3 passos).

A definição dos parâmetros e hiperparâmetros dos modelos, inicialmente se dá por meio da revisão da literatura e análise dos dados, e posteriormente por ciclos iterativos de testes, análise dos resultados, e refinamento dos algoritmos.

3.1 Requisitos

Os requisitos definidos para o projeto são listados abaixo.

- Os algoritmos devem ser implementados em linguagem que possibilite a importação de bibliotecas de código aberto, onde estejam implementados modelos de predição para séries temporais;
- Os dados utilizados devem ser públicos, disponibilizados por órgãos e instituições brasileiras;
- Os algoritmos devem receber os dados de maneira parametrizada, de forma que seja possível alterar os intervalos temporais de análise, treinamento e predição;
- Os modelos devem ser avaliados sob as mesmas métricas para efeito comparativo;
- Os resultados das predições devem ser expressos de maneira gráfica.

3.2 Seleção das variáveis

Para a seleção das variáveis, inicialmente é feito o tratamento dos dados utilizando a biblioteca *Pandas* (biblioteca de código aberto *Python*). As variáveis que são publicadas originalmente com discretização horária, foram tratadas para serem utilizadas com discretização diária, visto que a maior parte das variáveis têm esta discretização.

A seguir, são expostas, na Tabela 4, as variáveis previamente selecionadas para análise.

Tabela 4: Dados publicados pela ONS e CCEE

Dados	Descrição	Unidade
Preço de Liquidação das Diferenças (PLD)	Preço spot da energia elétrica no mercado livre brasileiro (principal objeto de estudo do trabalho)	R\$/MWh
Energia Natural Afluente (ENA)	Energia produzível pelas usinas, é calculada pelo produto das vazões naturais aos reservatórios com as produtividades a 65% dos volumes úteis	MW médio
Energia Armazenada (EAR)	Energia associada ao volume de água disponível nos reservatórios que pode ser convertido em geração na própria usina e em todas as usinas à jusante na cascata	MW mês
Carga de energia	Representam o perfil de consumo de energia elétrica com discretização diária	MW médio

Fonte: adaptado de CCEE [26] e ONS [27].

Os dados, acima descritos, são publicados para cada um dos 4 submercados em que o sistema elétrico brasileiro se divide (Norte, Nordeste, Sudeste, Sul), logo são ao menos 4 séries temporais (e portanto 4 variáveis) para cada classe. E em vista da característica de interligação dos subsistemas, foram também definidas como variáveis as somas de ENA, EAR e Carga de todos os subsistemas.

3.3 Modelos de predição

Foram definidos os modelos a serem testados, por meio da revisão da literatura e da própria fundamentação teórica de cada modelo.

3.3.1 ARIMA

O modelo Autorregressivo Integrado de Médias Móveis (ARIMA) utiliza a autocorrelação da série temporal, ou seja a correlação dos valores da série com os valores defasados no tempo (*lags*) para prever valores futuros [17].

São definidos 3 parâmetros (p,i,q) para a construção do modelo: p é o número de termos autorregressivos, d é o número de diferenciações, e q o termo para médias móveis. Estes parâmetros são definidos com base na autocorrelação e autocorrelação parcial da série, e também de maneira iterativa de forma a minimizar o erro de predição.

Este modelo requer que a série temporal seja estacionária para média e variância. Para isto, em alguns casos, é possível o tratamento por meio de diferenciação da série e escala logarítmica.

3.3.2 Redes Neurais LSTM

Redes neurais LSTM são um caso particular de redes neurais recorrentes com *gates* (portas), que controlam a memória das células da rede através dos passos temporais. Este tipo de rede foi criado devido ao problema de dissipação do gradiente (*vanishing gradients problem*), que faz com que as redes neurais recorrentes simples, não sejam capazes de preservar memória através dos passos temporais [28].

As redes neurais recorrentes simples são dadas por:

$$s_t = g(s_{t-1}W^s + x_tW^x + b)$$

$$y_t = s_t.$$

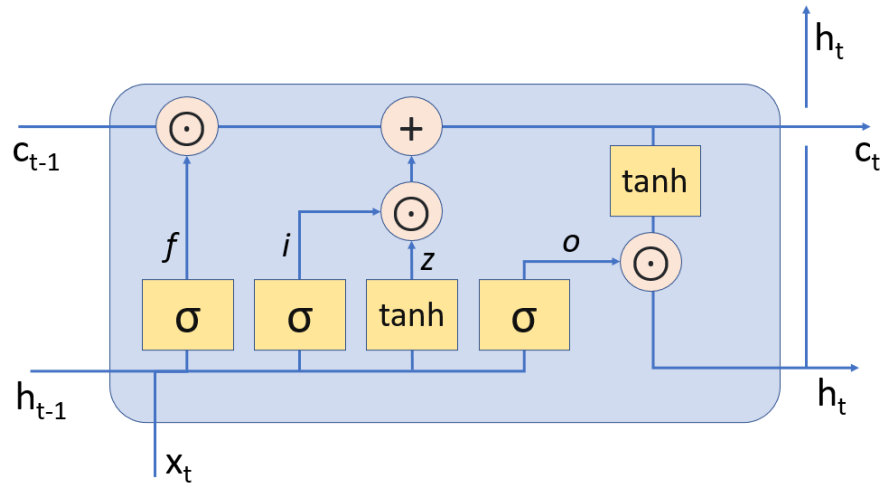
Em que s_t é o vetor estado da célula no instante t , e y_t é o vetor saída da célula no instante t . A função de ativação da célula é g (normalmente tangente hiperbólica ou ReLu (*rectified linear activation unit*), W^s é o peso matricial do estado da célula, W^x é o peso matricial da entrada da célula, e b o viés (*bias*). O índice subscrito $t - 1$, indica o valor de determinado termo no passo temporal anterior.

Observa-se que a formulação matemática de redes neurais recorrentes simples não trazem nenhuma possibilidade de conservação dos pesos matriciais ao longo dos avanços temporais. Como solução a este problema, as redes neurais LSTM são dadas por:

$$\begin{aligned}
s_t &= [c_t; h_t] \\
c_t &= f \odot c_{t-1} + i \odot z \\
h_t &= o \odot \tanh(c_t) \\
i &= \sigma(x_t W^{xi} + h_{t-1} W^{hi}) \\
f &= \sigma(x_t W^{xf} + h_{t-1} W^{hf}) \\
o &= \sigma(x_t W^{xo} + h_{t-1} W^{ho}) \\
z &= \tanh(x_t W^{xz} + h_{t-1} W^{hz}).
\end{aligned}$$

Em que s_t é o estado da célula no instante t , dado por dois vetores: componente de memória c_t , e componente de estado oculto h_t , também no instante t . O candidato à atualização z , é dado pela combinação linear do vetor de entrada x_t e do vetor componente de estado oculto do passo temporal anterior h_{t-1} . passados por uma função de ativação tangente hiperbólica. Há três gates: entrada i , esquecimento f , e saída o . Os gates são dados também pela combinação linear de x_t e h_{t-1} porém passados por uma função de ativação sigmoide, com pesos matriciais W^x para x_t e W^h para h_{t-1} , distintintos para cada gate. A componente de memória é atualizada por meio da soma dos produtos de elemento por elemento dos vetores (ou produto de Hadamart, representado por " $\odot$ ") de f , c_{t-1} e i , z . E a componente de estado oculto h_t é dada pelo produto de Hadamart de o e a tangente hiperbólica de c_t . A célula LSTM é esquematizada na Figura 4.

Figura 4: Representação de célula neural LSTM



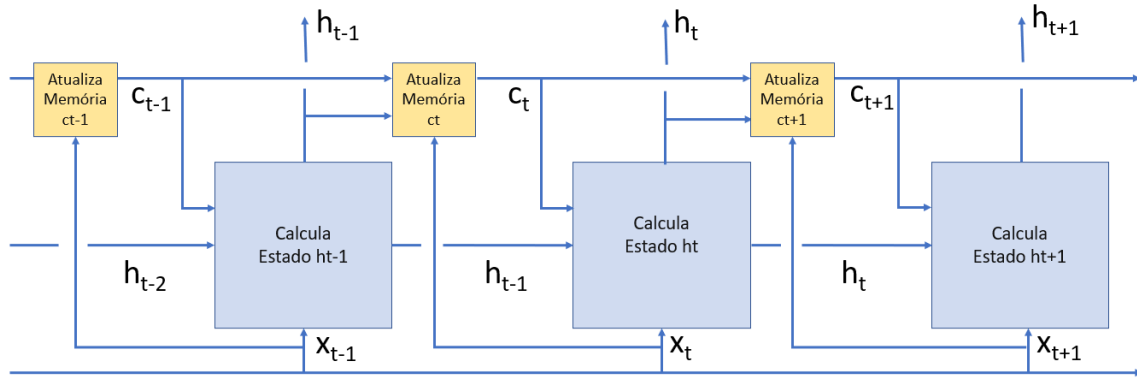
Fonte: adaptado de Guilehrme Santos [23] e Yaov Goldberg [28].

Efetivamente, a memória c_t é atualizada por meio do gate de esquecimento f , que

controla quanto do estado de memória anterior permanecerá, e do gate de entrada i , que controla quanto da nova entrada será adicionada à memória. E a atualização do estado oculto h_t , que é também a saída da célula y_t , ocorre por meio do controle do gate de saída o , aplicado à memória c_t .

Uma outra boa representação esquemática de redes neurais LSTM é adaptada do livro de François Chollet [29], exposta na Figura 5. De maneira simplificada, é indicado o funcionamento da rede em uma divisão de duas subcélulas, responsáveis por cada um dos vetores c_t e h_t , que compõe o estado s_i . Uma das subcélulas atualiza o vetor de memória c_t e a outra sub célula o estado h_t .

Figura 5: Representação do funcionamento de rede neural LSTM



Fonte: adaptado de François Chollet [29].

Por esta capacidade de guardar informações através da evolução dos passos temporais, as Redes Neurais LSTM são vastamente aplicadas para predição de séries temporais e apresentam no geral grande vantagem na capacidade preditiva em relação a outros modelos de predição [23], por este motivo será o modelo principal de predição do trabalho.

Para a definição da arquitetura dos modelos de rede neural LSTM a serem utilizados, não há uma regra objetiva. O número de neurônios e camadas da rede são definidos inicialmente baseado na literatura, e em seguida por meio de testes iterativos buscando o melhor resultado. O mesmo ocorre para definição de hiperparâmetros de treinamento. O *batch size*, tamanho do lote da amostra de treinamento, e *epochs*, número de ciclos completos de treinamento, são definidos em listas de possíveis valores, e passados ao modelo iterativamente.

3.3.3 Método Naïve

Para efeito comparativo, serão implementadas previsões sob o método Naïve. Este método consiste na ideia de que um preditor referencial, de uma série temporal, para os valores y_{t+n} dos n passos sucessivos ao passo t , é o próprio valor y_t do passo t . Conforme enunciado na equação abaixo.

$$y_{t+n,predito} = y_{t,realizado}$$

Como indicado por François Chollet [29], este método simples é comumente utilizado como referencial e em muitas aplicações, seu desempenho não consegue ser superado por modelos mais complexos.

3.3.4 Métricas de avaliação

Para que seja possível avaliar os resultados obtidos, e compará-los com a literatura e também realizar a comparação entre os modelos implementados, são definidas 3 métricas de avaliação frequentemente utilizadas para avaliação de previsões de séries temporais: Erro Absoluto Médio (MAE), Erro Percentual Absoluto Médio (MAPE), Raiz Quadrada dos Erros Quadráticos Médios (RMSE). As métricas são dadas por:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_{i,predito} - y_{i,realizado}|$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y_{i,predito} - y_{i,realizado}|}{y_{i,realizado}} \cdot 100\%$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{i,predito} - y_{i,realizado})^2}.$$

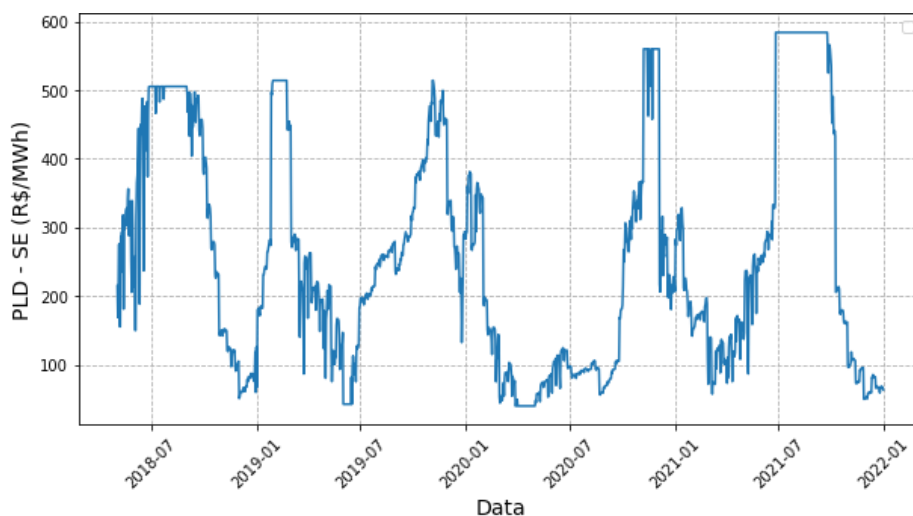
4 DESENVOLVIMENTO E RESULTADOS

Para o desenvolvimento do trabalho inicialmente foram analisadas todas as séries temporais disponíveis relacionadas às variáveis expostas na seção 3.2. Os testes dos modelos foram realizados com diferentes parâmetros e métodos, buscando a obtenção dos melhores resultados em termos das métricas de avaliação definidas. Os *notebooks* de desenvolvimento do projeto estão disponíveis em repositório *GitHub* [30].

4.1 Tratamento, análise e seleção das variáveis

Os dados do PLD são disponibilizados em base horária no site da CCEE, em formato "csv". A fim de trabalhar com todos os dados em discretização diária, os dados foram importados para *dataframe* e feito a média diária para cada um dos submercados. A série de interesse é o PLD do submercado Sudeste. O intervalo disponível é a partir de 02/05/2018, e foram selecionados os dados até 31/12/2021. A série é exposta na Figura 6.

Figura 6: Média diária do PLD para o submercado Sudeste

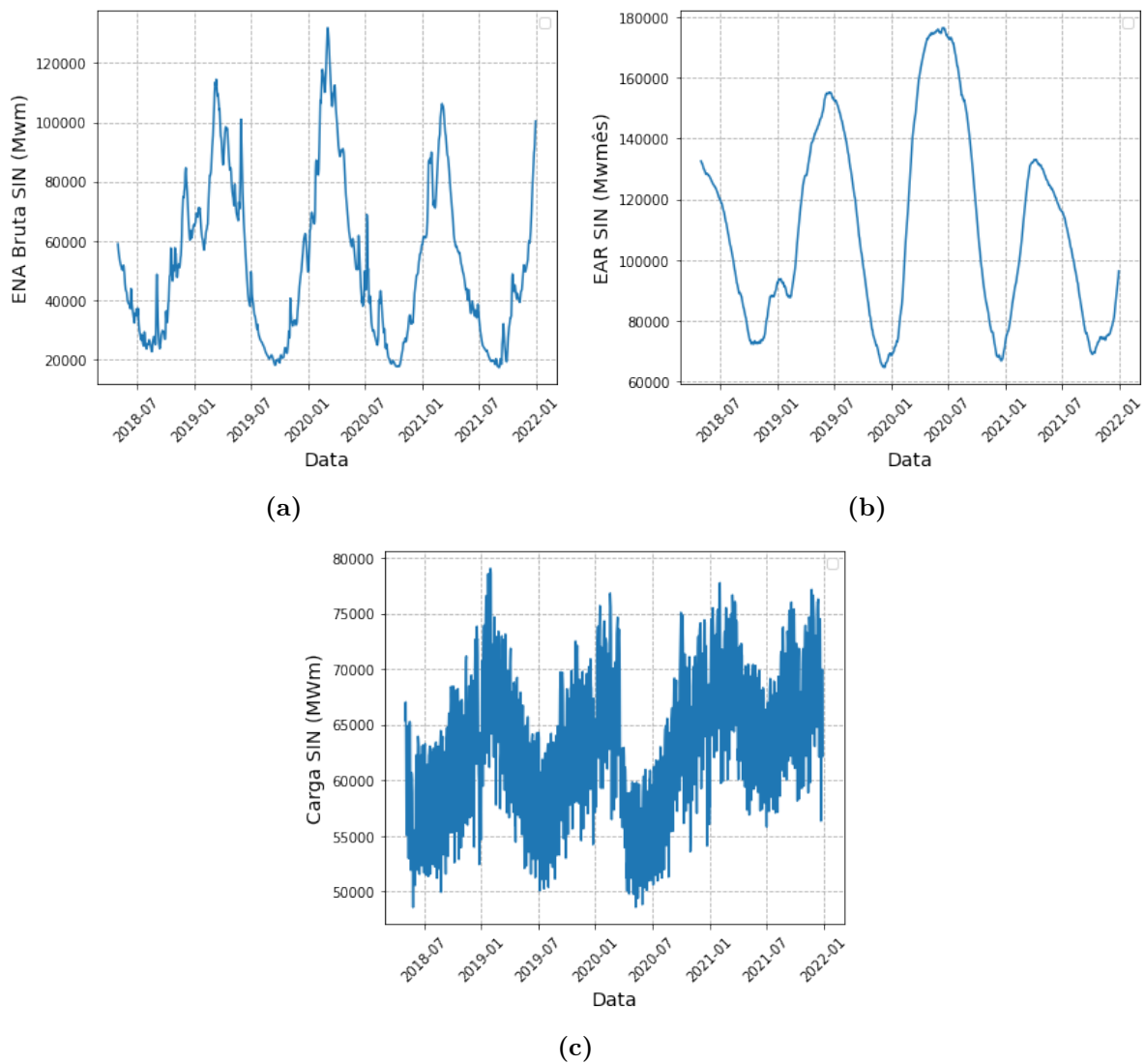


Fonte: elaborado pelo próprio autor, a partir de dados da CCEE [26].

Já as variáveis hidrológicas e de carga são disponibilizadas pelo ONS, também em

formato "csv" para cada um dos subsistemas. As diferentes séries foram analisadas e as mais relevantes foram selecionadas baseando-se nas correlações das séries, e também na revisão de literatura que indica causalidade das variações do PLD em função de outras variáveis [25]. Desta forma, foram selecionadas, a princípio, a série da soma de ENA Bruta e Energia Armazenada para os 4 subsistemas, e a série da soma de carga dos 4 subsistema. A seguir, estas séries são plotadas nas Figuras 7a, 7b e 7c.

Figura 7: a) Série da soma da Energia Natural Afluente para os 4 submercados do SIN b) Série da soma da Energia Armazenada para os 4 submercados do SIN c) Série da soma da Carga para os 4 submercados do SIN

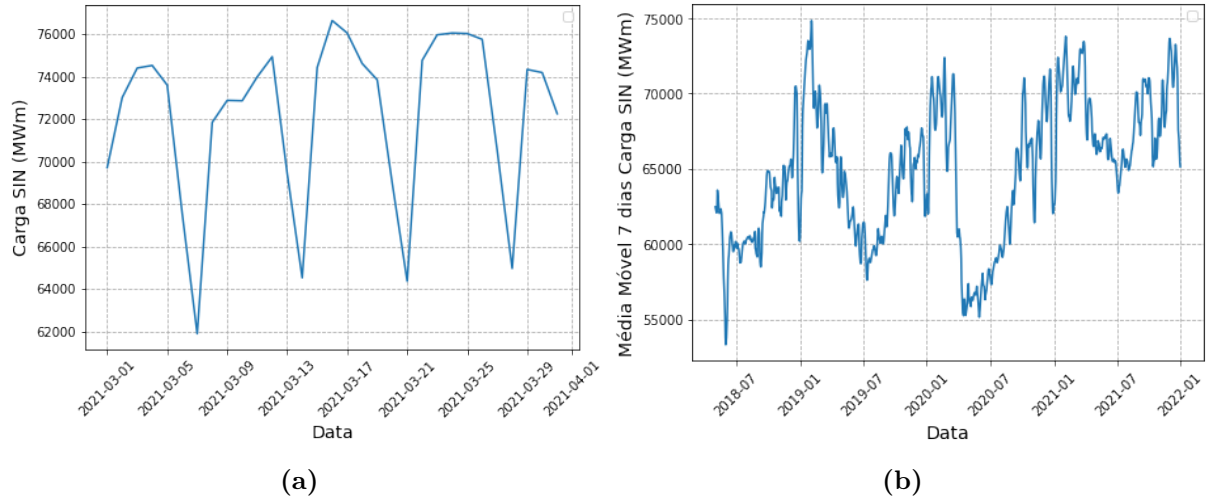


Fonte: a), b) e c) elaborado pelo próprio autor, a partir de dados do ONS [27].

Para a série da soma de carga dos 4 subsistemas, diferentemente das séries das demais variáveis, é observada sazonalidade semanal. Os níveis de carga seguem padrões distintos em dias úteis e em dias não úteis (finais de semana), conforme ilustrado pela Figura 8a que

expressa a série no mês de março de 2021. Para mitigar este efeito sazonal, foi utilizada como variável a média móvel de 7 dias da série da soma de carga dos 4 subsistemas, expressa na Figura 8b.

Figura 8: a) Série da soma da Carga para os 4 submercados do SIN em março de 2021 b) Série da média móvel da soma da Carga para os 4 submercados do SIN



Fonte: a), b) elaborado pelo próprio autor, a partir de dados do ONS [27].

Por fim, foram selecionadas como possíveis variáveis para a predição da série do PLD, a série da soma de ENA para os 4 subsistemas, a série da soma de EAR para os 4 subsistemas, e a série da média móvel de 7 dias para a soma de carga dos 4 subsistemas. A matriz de correlação entre essas variáveis é exposta na Figura 9.

Figura 9: Matriz de correlação

	SE - PLD Médio (R\$/MWh)	ENA Bruta (MWm)	EAR Verif. (MWmês)	Média Móvel 7d Carga (MWm)
SE - PLD Médio (R\$/MWh)	1.000	-0.520	-0.462	0.112
ENA Bruta (MWm)	-0.520	1.000	0.175	0.257
EAR Verif. (MWmês)	-0.462	0.175	1.000	-0.585
Média Móvel 7d Carga (MWm)	0.112	0.257	-0.585	1.000

Fonte: elaborado pelo próprio autor.

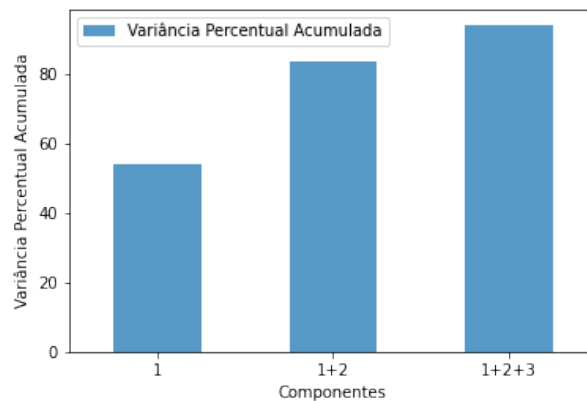
4.1.1 Análise de Componente Principal (PCA)

A fim de poder analisar a variabilidade da amostra multivariável, e a interação entre as diferentes variáveis, foi realizada a Análise de Componentes Principais (PCA). Esta

análise consiste na transformação linear ortogonal do conjunto de dados multivariável, em um novo sistema com variáveis não correlacionadas (componentes principais) [31], de maneira que a primeira componente principal contenha a maior proporção da variância total da amostra, e a segunda componente, a segunda maior proporção da variância total, assim até chegar ao total da variância ao somar todas as componentes.

Aplicando PCA aos dados, utilizando 3 componentes principais (CP), a variância explicada (porcentagem da variância total, contida na componente) pela CP 1 foi de 54,0%, pela CP2 de 29,7%, e pela CP3 10,6% . As 3 componentes principais são capazes de explicar 94,3% da variância total da amostra. A seguir é expressa graficamente, na Figura 10, a variância percentual acumulada por componente (isto é, a soma da proporção de variância total contida cumulativamente em cada componente, considerando componentes anteriores), e os pesos de cada variável para cada uma das componentes, na Tabela 5.

Figura 10: Variância percentual acumulada para soma de cada componente principal com as componentes anteriores



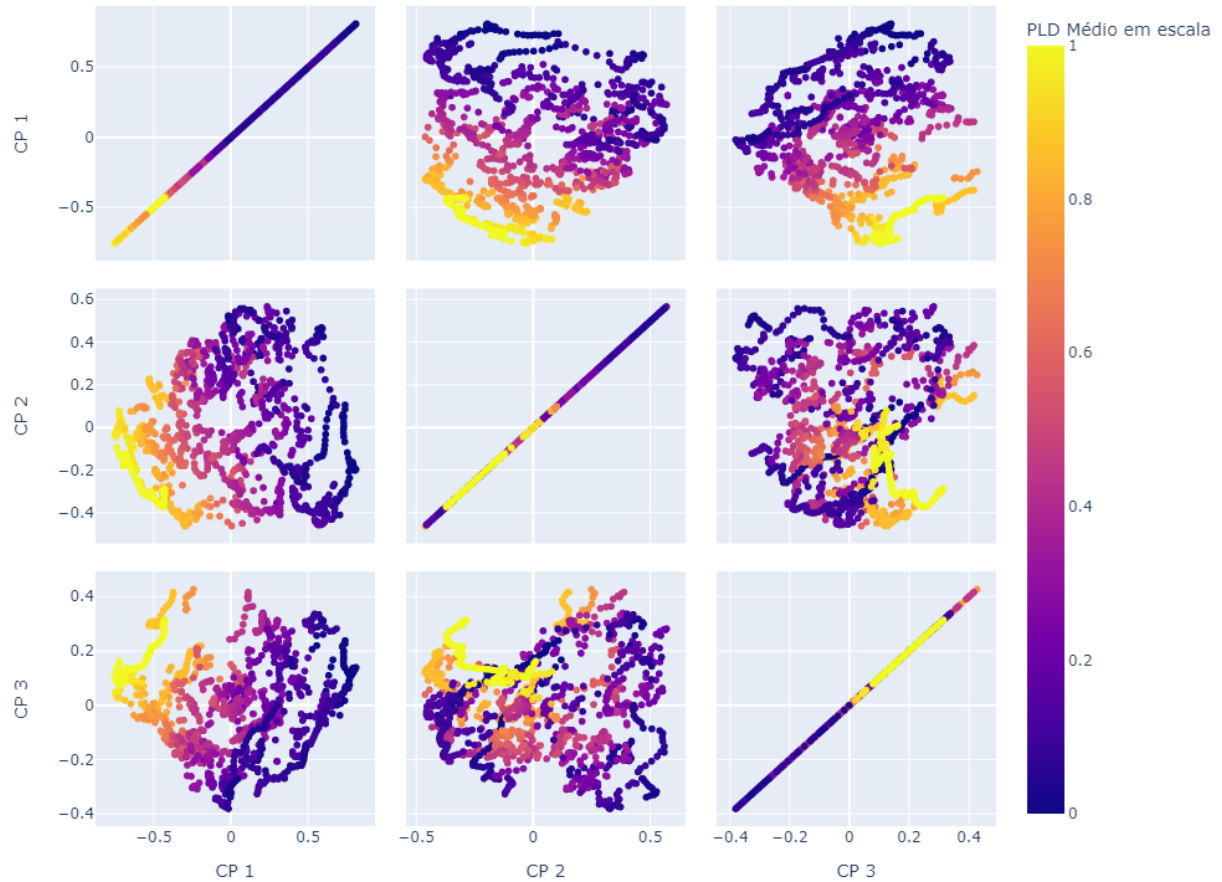
Fonte: elaborado pelo próprio autor.

Tabela 5: Componentes principais da análise para 3 componentes

Variável	CP 1	CP 2	CP 3
PLD	-0,70	-0,37	0,62
ENA	0,32	0,54	0,65
EAR	0,60	-0,50	0,42
Média Móvel Carga	-0,21	0,57	0,13

Fonte: elaborado pelo próprio autor.

Foi plotado também o gráfico da distribuição dos pontos, rotulados em cores de acordo com o nível do PLD em escala (de 0 a 1), em função das componentes principais nos eixo, exposto na Figura 11.

Figura 11: Distribuição dos pontos em função das componentes principais

Fonte: elaborado pelo próprio autor.

Observa-se que a Média Móvel da Carga é a variável com menor peso para a componente principal 1. Além disso, da seção anterior, é notada correlação relevante entre a série de EAR e a série de média de carga. Preliminarmente, esta correlação pode ser explicada pelo funcionamento do SIN. O aumento da demanda do sistema elétrico brasileiro, faz com que o ONS tenha que despachar usinas hidroelétricas, deplecionando os reservatórios.

4.1.2 Saturações

De acordo com a Resolução Normativa ANEEL nº 858/19 o PLD é limitado anualmente por valores máximos e mínimos. Logo, observa-se na série temporal da média diária do PLD, saturações dos dados. A fim do estudo da influência das saturações na capacidade preditiva dos modelos, os 6 períodos em que ocorrem foram identificados e expostos na Tabela 6.

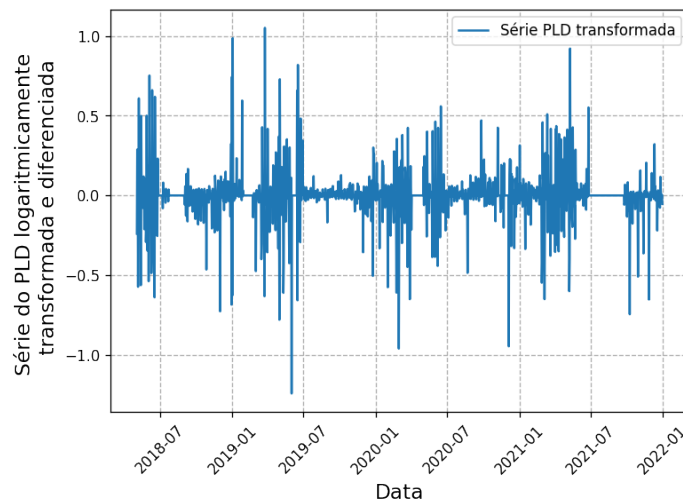
Tabela 6: Períodos de saturação da média diária do PLD

Data de início	Data de término
01/06/2018	01/10/2018
26/01/2019	23/02/2019
01/06/2019	14/06/2019
28/03/2020	02/05/2020
06/11/2020	05/12/2020
24/06/2021	06/10/2021

Fonte: elaborado pelo próprio autor.

4.2 Predição com modelo ARIMA

O modelo de predição ARIMA parte do princípio de que a série predita é estacionária em variância e média. É perceptível que a série do PLD não é estacionária. Uma estratégia comum para o tratamento das séries a serem preditas pelo modelo ARIMA, é a criação de uma nova série do logaritmo dos valores, em busca da estacionariedade da variância, e em seguida a derivada desta nova série, em busca da estacionariedade da média. Ao realizar estas transformações na série do PLD, é obtida a nova série abaixo, expressa na Figura 12.

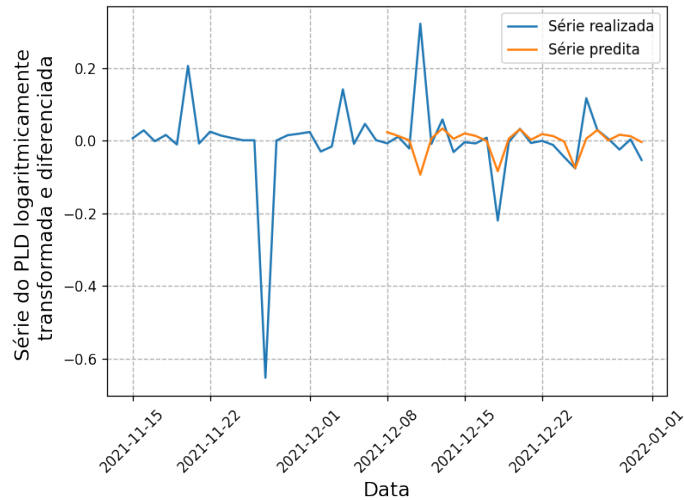
Figura 12: Gráfico da série do PLD logaritmicamente transformada e diferenciada

Fonte: elaborado pelo próprio autor.

Observa-se que mesmo após as transformações, a série não é estacionária. Há períodos de maior variância, e períodos de menor variância (como as saturações). Desta forma, os resultados de um modelo ARIMA implementado para esta série, não terá boa capacidade preditiva para mudanças de tendência. Para confirmar esta incapacidade do modelo

ARIMA, houve uma tentativa de implementação, com os parâmetros $p=7$, $d=1$ e $q=7$. Para o teste, foi escolhido a janela de 08/12/2021 a 31/12/2021. O resultado é exposto na Figura 13.

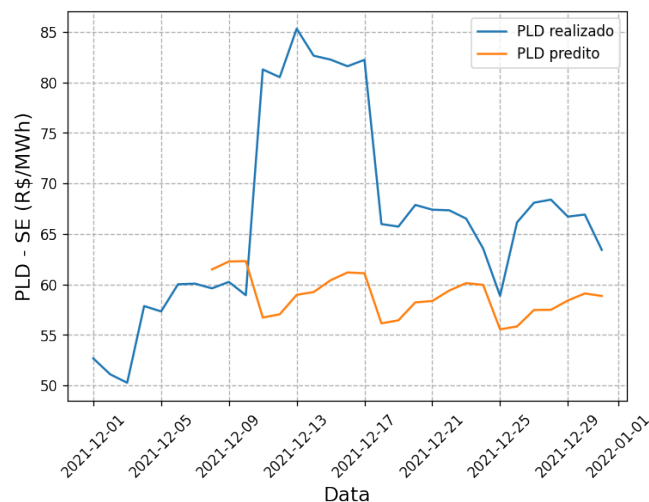
Figura 13: Gráfico da série do PLD transformada realizada e da predição com modelo ARIMA



Fonte: elaborado pelo próprio autor.

Aplicando as transformações inversas, é obtida a série original novamente, e as predições, expressas graficamente na Figura 14.

Figura 14: Gráfico da série do PLD realizado e da predição com modelo ARIMA



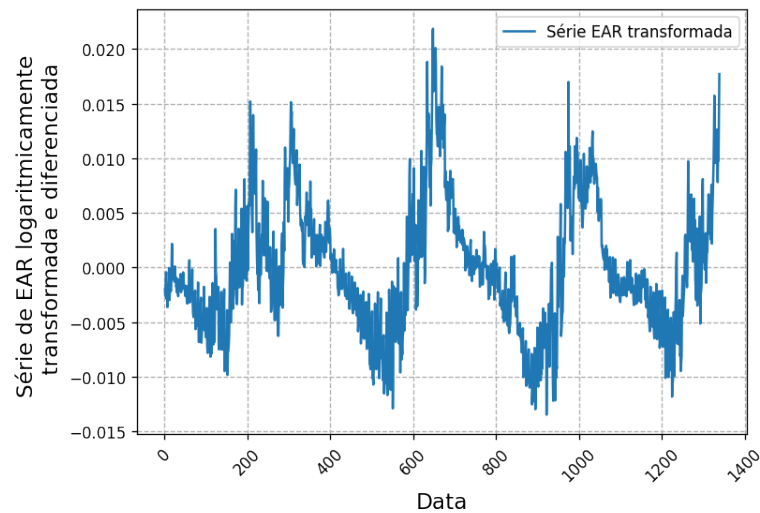
Fonte: elaborado pelo próprio autor.

As métricas de avaliação obtidas para predição de 7 dias à frente foram MAE de R\$ 15,01, MAPE de 18,70 % e RMSE de 18,56. Conclui-se que em vista da característica de não estacionariedade da série do PLD, mesmo após transformações, o modelo ARIMA não

é um bom preditor para esta aplicação.

Como contraponto, em vista da aplicação de ARIMA para outras séries e mercados de outros países, é feito o teste da implementação de um modelo ARIMA, para uma série que seja possível obter estacionariedade. Para isso, é implementado um modelo ARIMA para predição da série de EAR. Aplicando as mesmas transformações para busca de estacionariedade da série, é obtida a série exposta na Figura 15.

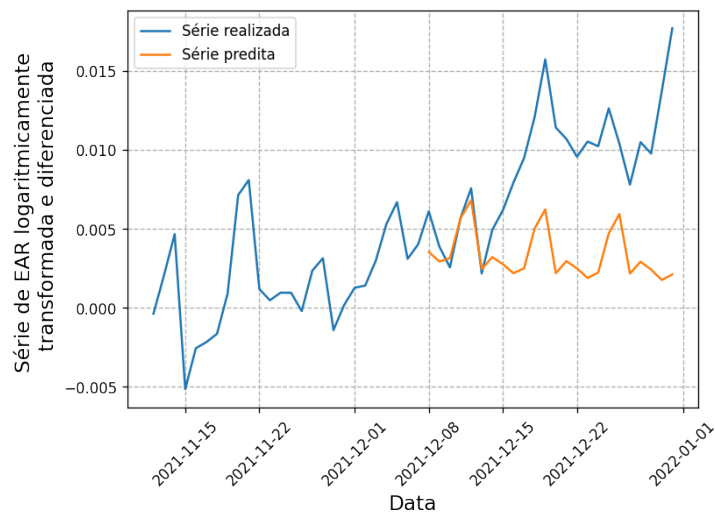
Figura 15: Gráfico da série de EAR logaritmicamente transformada e diferenciada



Fonte: elaborado pelo próprio autor.

Checada a estacionariedade da série, foi implementado um modelo ARIMA com os parâmetros $p=7$, $d=1$ e $q=7$. Para o teste, foi escolhido também a janela de 08/12/2021 a 31/12/2021.

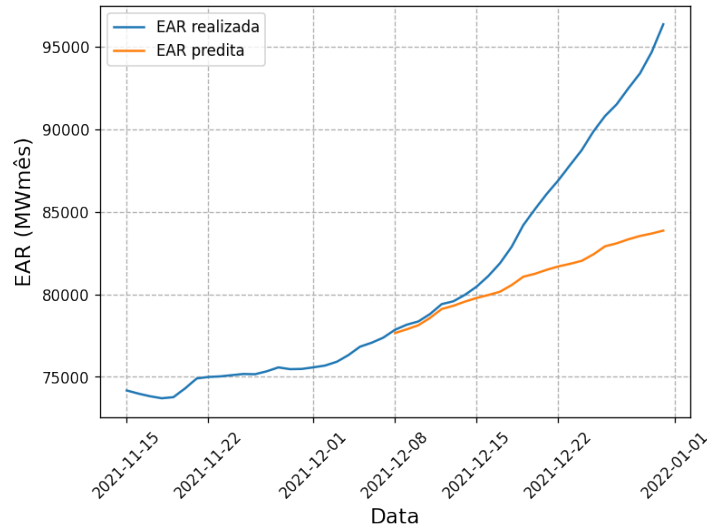
Figura 16: Gráfico da série de EAR transformada realizada e da predição com modelo ARIMA



Fonte: elaborado pelo próprio autor.

E finalmente, realizando as transformações inversas, é obtida a série original, e as previsões.

Figura 17: Gráfico da série de EAR realizada e da previsão com modelo ARIMA



Fonte: elaborado pelo próprio autor.

Para predição de 7 dias à frente foi obtido MAE de 273,92 MWmês, MAPE de 0,35% e RMSE de 281,19. Já considerando 24 dias à frente, foi obtido MAE de 4314,01 MWmês, MAPE de 4,78% e RMSE de 5810,44.

Assim, fica demonstrada a boa capacidade preditiva do modelo ARIMA para séries estacionárias, e a impossibilidade do uso do modelo para a predição do PLD, pela característica de não estacionariedade. Nas subseções seguintes serão abordados o desenvolvimento e resultados das previsões com o modelo de rede neural LSTM, modelo principal de enfoque do presente trabalho.

4.3 Predição do PLD um dia à frente com modelo LSTM

Inicialmente, foi implementado um modelo de rede neural LSTM, para predição de somente um dia à frente. A rede foi construída com uma camada de entrada, uma camada com 100 neurônios LSTM, uma camada Dense de 50 neurônios, e por fim uma última Dense unitária de saída.

A amostra de treinamento foi definida de 02/05/2018 a 30/11/2021, e a amostra de teste de 01/12/2021 a 31/12/2021. A predição de um dia à frente foi realizada com base nos últimos 7 dias, logo os conjuntos de entrada e saída continham 7 pontos e 1 ponto,

respectivamente. Como já abordado na subseção 3.3.2, os hiperparâmetros dos modelos foram definidos de maneira iterativa, buscando minimizar o erro, e a convergência dos treinamentos.

Os treinamentos e testes foram realizados com os dados organizados de 3 maneiras distintas: conjunto completo com as 4 variáveis expostas na seção anterior; conjunto com apenas 3 variáveis, em que é excluída a variável de carga; e conjunto com 3 variáveis sem os períodos de saturação da série do PLD.

4.3.1 Resultado predição com 4 variáveis

Para o modelo com 4 variáveis de entrada, o melhor resultado, em função dos diferentes hiperparâmetros de treinamento da rede (*epochs*, *batchsize*), obteve MAE de R\$4,76, MAPE de 6,78% e RMSE de 6,49. É exposto o gráfico na Figura 18.

Figura 18: Gráfico do PLD realizado e das predições em janela móvel modelo LSTM com 4 variáveis de entrada



Fonte: elaborado pelo próprio autor.

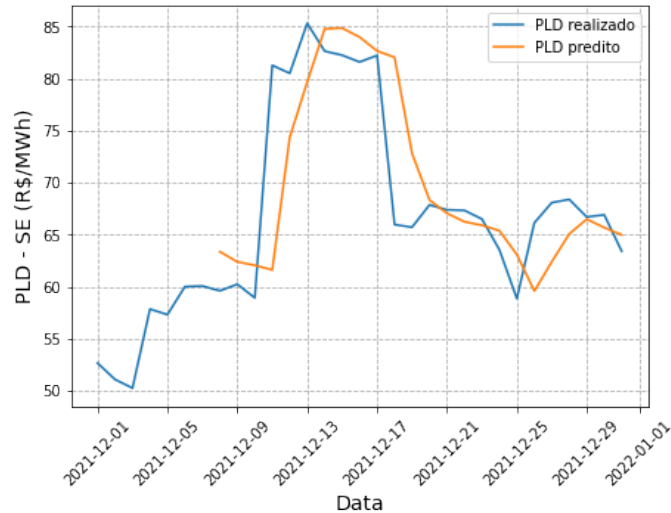
4.3.2 Resultado predição com 3 variáveis

O treinamento e teste do modelo foi realizado também com somente 3 variáveis (PLD, ENA e EAR). As motivações para a exclusão da série de carga foram a menor correlação com a série do PLD e o menor peso atribuído a esta variável para componente principal 1 (como exposto na subseção 4.1.1).

O melhor resultado obteve MAE de R\$4,10, MAPE de 5,81% e RMSE de 6,20. É

exposto o gráfico na Figura 19.

Figura 19: Gráfico do PLD realizado e das previsões em janela móvel modelo LSTM com 3 variáveis de entrada

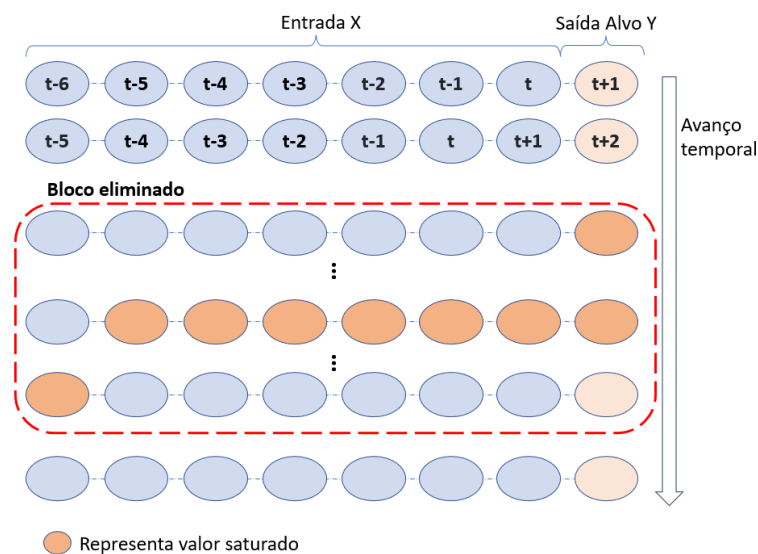


Fonte: elaborado pelo próprio autor.

4.3.3 Resultado predição excluindo saturações do treinamento

A fim de observar a performance dos modelos excluindo os períodos de saturação, foi criada uma função para eliminar corretamente a formação dos conjuntos de entrada e saída do treinamento, conforme esquematizado na Figura 20.

Figura 20: Esquema da função de eliminação das saturações



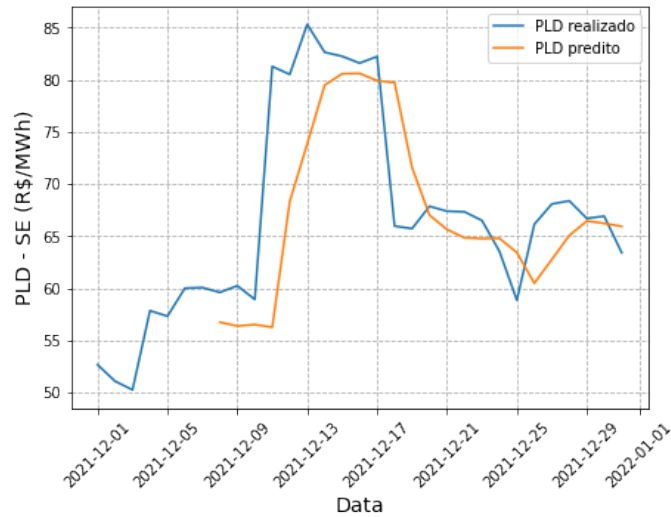
Fonte: elaborado pelo próprio autor.

Devido à característica de autocorrelação das séries, não basta somente excluir os

dados dos períodos de saturação, é necessário garantir que não haja dias não consecutivos, no mesmo conjunto de treinamento. Para isso, é necessário eliminar da amostra os blocos em que a saturação ocorre na saída do conjunto de treinamento, continuamente até a obtenção de um novo conjunto de entrada e saída que não haja nenhum valor saturado, como indicado na Figura 20.

Feito a eliminação correta das saturações, e treinada a rede, o melhor resultado obtido, exposto graficamente na Figura 21 teve MAE de R\$4,83, MAPE de 6,70% e RMSE de 7,36.

Figura 21: Gráfico do PLD realizado e das previsões em janela móvel modelo LSTM com 3 variáveis de entrada e eliminação das saturações do treinamento



Fonte: elaborado pelo próprio autor.

4.4 Predição do PLD para mais de um dia à frente com modelo LSTM

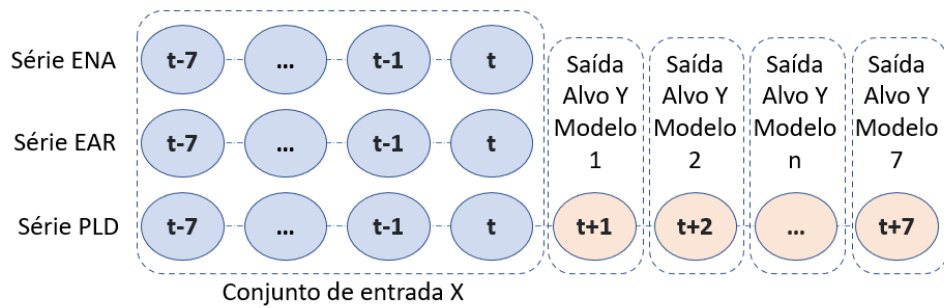
A predição da série do PLD para mais de um dia à frente, foi realizada com dois métodos distintos. O primeiro, a partir de modelos de redes neurais para predição *single-step*, em que as redes têm saída unitária, porém utilizando para o treinamento, conjuntos de entrada com saídas deslocadas no tempo. E o segundo método, modelos de redes neurais para predição *multi-step*, em que as redes têm mais de uma célula de saída.

Em vista dos resultados da seção anterior, que confirmaram a possibilidade da eliminação da carga como variável sem que houvesse perda de capacidade preditiva, como foi sugerido a partir da análise das variáveis, os modelos foram criados utilizando 3 variáveis: série do PLD, série de ENA e série de EAR.

4.4.1 Método *single-step*

Para o método *single-step*, foi usada uma arquitetura de rede exatamente como a da seção anterior: uma camada de entrada, uma camada LSTM com 100 neurônios, uma camada Dense com 50 neurônios, e uma camada de saída unitária. Foram utilizados 7 modelos distintos, para a predição de 7 dias à frente. O modelo para predição de um dia à frente, é o mesmo da subseção 4.3.2. Para os demais 6 dias à frente, foram implementados modelos de mesma arquitetura de rede, com os conjuntos de treinamento, com mesma entrada (contendo 3 variáveis em 7 dias), porém saída alvo distinta, deslocada no tempo, conforme ilustra a Figura 22.

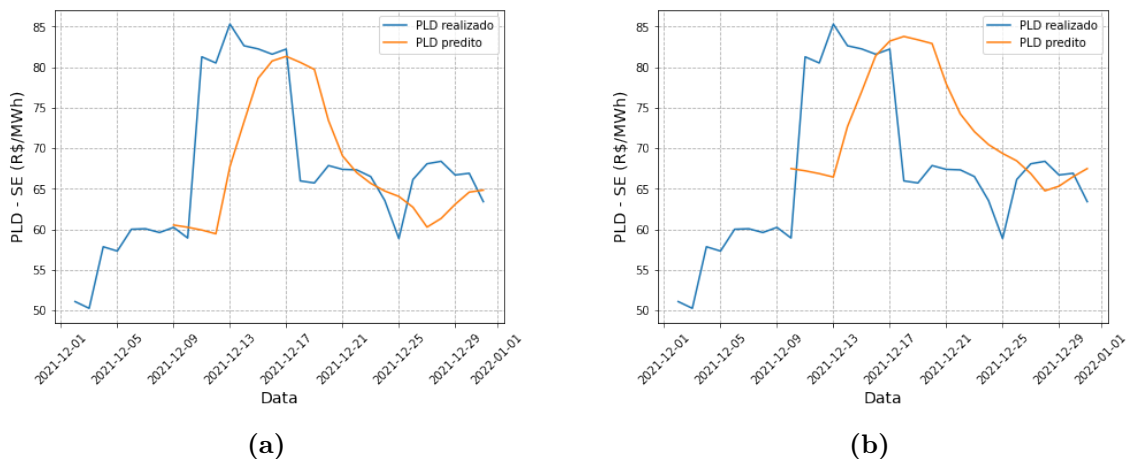
Figura 22: Esquema dos conjuntos de entrada e saída para treinamento dos modelos *single-step*



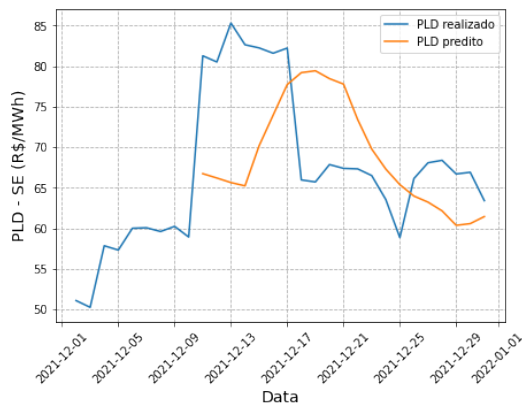
Fonte: elaborado pelo próprio autor.

Para os 6 novos modelos a amostra de treinamento foi definida de 02/05/2018 a 30/11/2021, e a amostra de teste de 01/12/2021 a 31/12/2021. Os melhores resultados em função dos hiperparâmetros de treinamento, para cada modelo, são expostos na Tabela 7 e graficamente nas Figuras 23a, 23b, 24a, 24b, 25a, 25b.

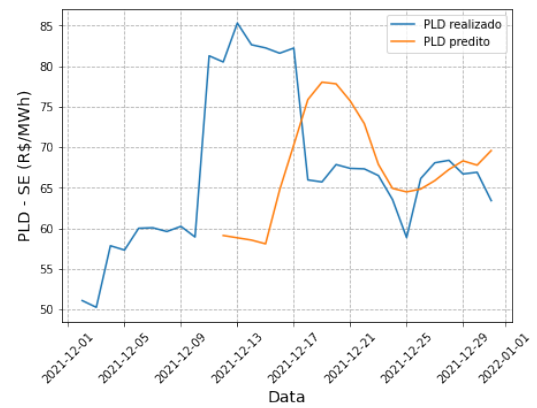
Figura 23: a) Predições em janela móvel modelo 2 b) Predições em janela móvel modelo 3



Fonte: a), b) elaborado pelo próprio autor.

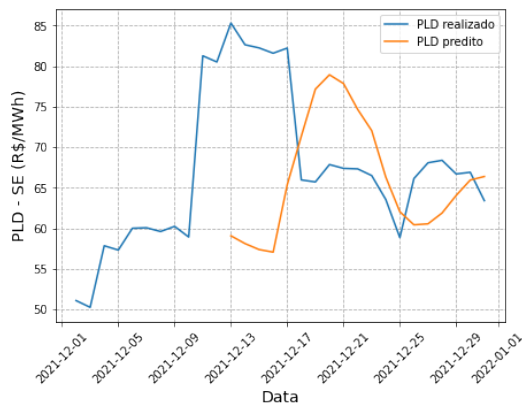
Figura 24: a) Predições em janela móvel modelo 4 b) Predições em janela móvel modelo 5

(a)



(b)

Fonte: a), b) elaborado pelo próprio autor.

Figura 25: a) Predições em janela móvel modelo 6 b) Predições em janela móvel modelo 7

(a)



(b)

Fonte: a), b) elaborado pelo próprio autor.

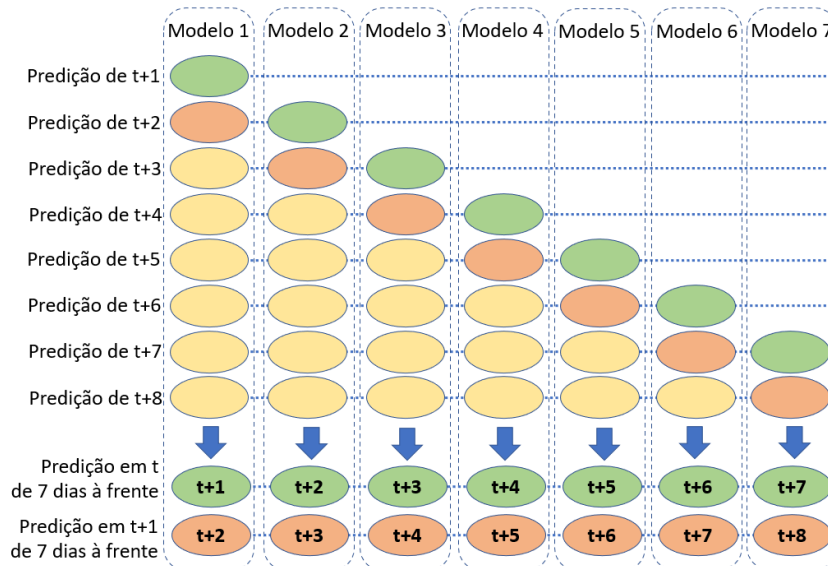
Tabela 7: Métricas de avaliação dos modelos *single-step*

Modelo	MAE (R\$)	MAPE %	RMSE
1	4,10	5,81	6,20
2	6,32	8,62	9,16
3	7,97	11,34	9,96
4	8,84	12,08	10,16
5	9,63	12,72	12,81
6	10,55	14,05	13,45
7	10,76	15,19	12,43

Fonte: elaborado pelo próprio autor.

A utilidade deste método consiste em prever os 7 dias à frente fornecendo a mesma entrada para os modelos. Para isso, foi necessário implementar uma função que recebe a saída dos 7 modelos, e reorganiza os dados sequencialmente. A Figura 26 ilustra esta reorganização.

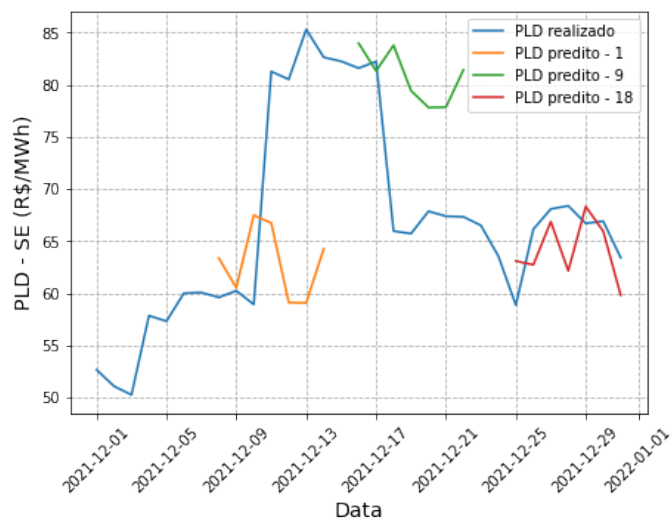
Figura 26: Esquema da função de reorganização das previsões dos modelos do método *single-step*



Fonte: elaborado pelo próprio autor.

Feita esta reorganização das previsões, são obtidos 18 previsões sequenciais para a janela completa de teste. O gráfico contido na Figura 27 expressa a primeira, a nona e a última previsão realizada pelos modelos, sob o método *single-step*.

Figura 27: Gráfico previsões método *single-step*



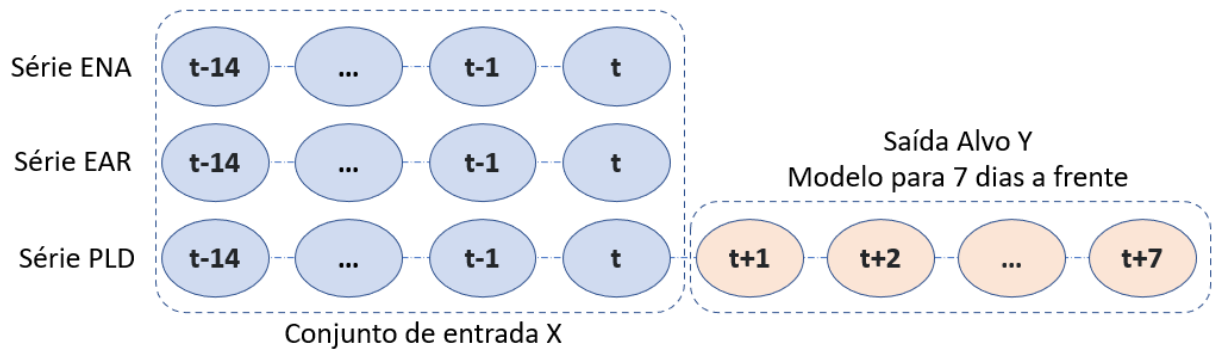
Fonte: elaborado pelo próprio autor.

As métricas obtidas para as previsões reorganizadas desta forma foram MAE de R\$9,21, MAPE de 12,52% e RMSE de 11,68.

4.4.2 Método *multi-step*

Para este outro método, foram implementados modelos com saídas múltiplas para cada dia de previsão. A arquitetura de rede utilizada é semelhante ao do método *single-step*, diferindo somente na saída: uma camada de entrada, uma camada LSTM com 100 neurônios, uma camada Dense com 50 neurônios, e uma camada de saída com n neurônios, em que n representa o número de dias preditos. Foram implementados 3 modelos, para a previsão de 3, 5 e 7 dias à frente. O conjunto de treinamento, com entrada e saída, foi dividido conforme demonstrado na Figura 28, tomando o modelo para previsão de 7 dias como exemplo.

Figura 28: Esquema do conjunto de entrada e saída para treinamento do modelos *multi-step* para previsão de 7 dias a frente

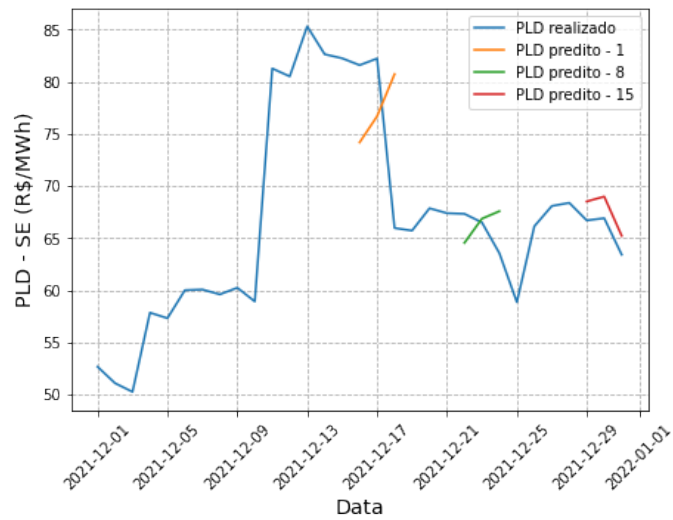


Fonte: elaborado pelo próprio autor.

A amostra de treinamento foi também definida de 02/05/2018 a 30/11/2021, e a amostra de teste de 01/12/2021 a 31/12/2021. Porém utilizando 14 dias para a previsão de 3, 5 e 7 dias à frente, para cada modelo.

O melhor modelo para a previsão de 3 dias à frente, foi testado para todos os possíveis períodos de previsão na amostra de teste, totalizando 15 testes. Na Figura 29, são expressos graficamente o primeiro, o oitavo e o último teste.

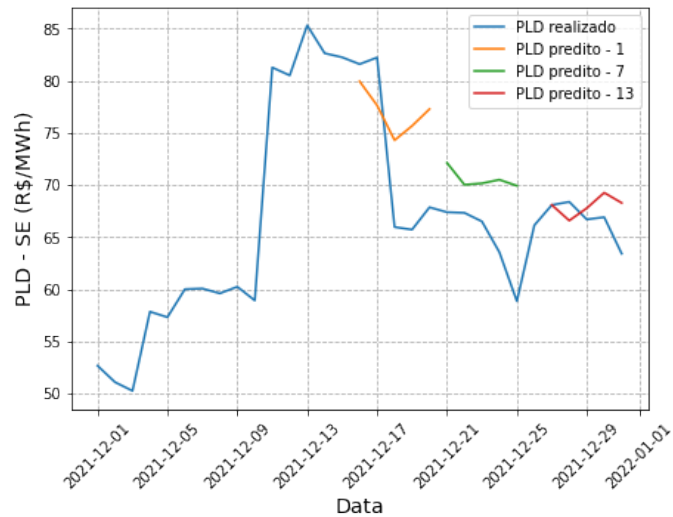
Figura 29: Gráfico do PLD realizado e das previsões 1, 8 e 15 do modelo *multi-step* para 3 dias à frente



Fonte: elaborado pelo próprio autor.

Já para a previsão de 5 dias à frente, o total de testes possíveis realizados foi de 13 testes. Na Figura 30, são expressos graficamente o primeiro, o sétimo e o último teste.

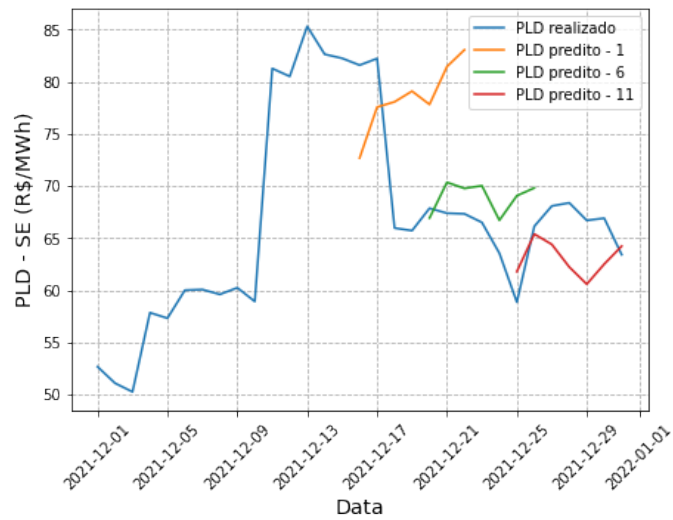
Figura 30: Gráfico do PLD realizado e das previsões 1, 7 e 13 do modelo *multi-step* para 5 dias à frente



Fonte: elaborado pelo próprio autor.

E por fim, para a previsão de 7 dias à frente, foram realizados os de 11 possíveis testes. Na Figura 31, são expressos graficamente o primeiro, o sexto e o último teste.

Figura 31: Gráfico do PLD realizado e das previsões 1, 6 e 11 do modelo *multi-step* para 7 dias à frente



Fonte: elaborado pelo próprio autor.

Os desempenhos dos modelos nos termos das métricas de avaliação são expostos na Tabela 8.

Tabela 8: Métricas de avaliação dos modelos *multi-step*

Modelo	MAE (R\$)	MAPE %	RMSE
Modelo 3 - predição de 3 dias à frente	5,02	7,37	6,31
Modelo 5 - predição de 5 dias à frente	5,44	8,26	6,60
Modelo 7 - predição de 7 dias à frente	7,00	10,48	8,40

Fonte: elaborado pelo próprio autor.

5 DISCUSSÃO

Após o desenvolvimento e teste dos modelos, é possível avaliar quais foram as variáveis mais expressivas para a predição do PLD, quais foram os melhores métodos, e como foi o desempenho dos modelos no geral. Para a avaliação do desempenho, são consideradas duas comparações: a primeira o método referencial Naïve, que consiste na repetição do último valor conhecido da série como predição, e a segunda a comparação com resultados reportados na literatura para a mesma aplicação.

5.1 Seleção de variáveis

Por meio da análise da correlação entre as variáveis, da análise de componente principal (PCA), e dos testes de predição de um dia à frente, ficou evidente que para a predição da série do PLD, além da própria série, as variáveis hidrológicas são as mais importantes.

A característica predominantemente hidroelétrica do sistema de geração elétrica brasileiro, faz com que a carga seja menos determinante para o preço da energia, comparando a mercados livres de energia de outros países [23]. Isso porque o ONS e os órgãos públicos de planejamento do sistema elétrico, são capazes de operar o sistema de modo a manter os níveis dos reservatórios em patamares adequados para absorver as variações de demanda (carga).

A correlação da série de Energia Armazenada e da série de carga, pode ser interpretada por dois fatores principais. O primeiro fator é o mais lógico, caso haja decremento dos níveis de carga do SIN o ONS contém o despacho de usinas hidroelétricas e isso faz com que os níveis dos reservatórios aumentem, esta dinâmica fica evidente no início do período pandêmico no primeiro semestre de 2020 em que os níveis de carga foram abaixo das médias anuais para o período, e o inverso ocorreu com o nível dos reservatório. O segundo fator, se deve a sazonalidade anual de carga e ENA, a carga é correlacionada com a temperatura, logo os maiores níveis de carga ocorrem no verão, e o período chuvoso brasileiro também ocorre predominantemente no verão (níveis mais elevados de ENA), isso faz com que o

ONS busque preservar os reservatórios para o término do período chuvoso, fazendo com que obtenham os níveis mais altos. Desta forma, a série de carga e de EAR ficam com sazonalidades anuais inversas: no verão a carga está em patamar alto e os reservatórios em níveis baixos, e o inverso ocorre no inverno.

Do ponto de vista metodológico, no geral variáveis altamente correlacionadas prejudicam a capacidade preditiva dos modelos. Para a aplicação de predição do PLD, houve uma melhora do desempenho do modelo LSTM ao retirar a série de carga do conjunto de treinamento, mantendo todas as demais variáveis e parâmetros iguais: a mudança em termo de MAPE foi de 6,78% para 5,81%, e em termo de RMSE de 6,49 para 6,20.

5.2 Modelo preditivo

Como foi levantado por meio da revisão da literatura, o modelo de redes neurais LSTM, apresenta boa capacidade preditiva de séries temporais. Especificamente para a série do PLD, a possibilidade de operar com séries com baixa estacionariedade da média e da variância, possibilitaram que o modelo funcionasse sem a necessidade da remoção das saturações. Como foi mostrado na subseção 4.3.3, a remoção das saturações do treinamento não trouxe melhora ao desempenho do modelo.

Esta capacidade de trabalhar com séries sem estacionariedade é uma grande benefício quando comparado a outros modelos como o ARIMA. Foram indicados bons resultados destes modelos em aplicações de séries temporais relacionadas a sistemas elétricos. Por esta razão, houve a tentativa de implementação deste modelo, apesar de na revisão da literatura, não ser encontrado aplicação de ARIMA para predição do preço no mercado elétrico brasileiro.

5.2.1 Comparação dos métodos

Para as predições de um dia à frente, nenhum dos modelos implementados foi capaz de bater o método referencial Naïve, conforme exposto na Tabela 9.

Tabela 9: Comparação dos modelos para predição de um dia à frente

Modelos	MAE (R\$)	MAPE %	RMSE
Método Naïve	3,21	4,53	6,14
Modelo com 4 variáveis de entrada	4,76	6,78	6,49
Modelo com 3 variáveis	4,10	5,81	6,20
Modelo com 3 variáveis e exclusão das saturações	4,83	6,70	7,36

Fonte: elaborado pelo próprio autor.

Observa-se que apesar de nenhum modelo ter superado o método Naïve, o melhor modelo (LSTM com 3 variáveis de entrada), obteve um RMSE em linha com o RMSE obtido por Naïve. Isto indica que apesar da diferença em termo de MAE e MAPE, o método Naïve apresenta pontos de predição com grandes erros, comparativamente ao melhor modelo LSTM. E como indicado por François Chollet [29], para muitas aplicações, modelos mais complexos não são capazes de apresentar melhor desempenho na predição de um passo à frente em relação a este método referencial.

Já para as predições de mais de um dia à frente, alguns dos modelos implementados são capazes de superar o método Naïve. As comparações para 3 dias à frente são expressas na Tabela 10, para 5 dias à frente na Tabela 11 e para 7 dias à frente na Tabela 12.

Tabela 10: Comparação do modelo *multi-step* para predição de 3 dias à frente com método Naïve

Modelos	MAE (R\$)	MAPE %	RMSE
Método Naïve	5,78	7,95	9,43
Modelo <i>multi-step</i> para predição de 3 dias à frente	5,02	7,37	6,31

Fonte: elaborado pelo próprio autor.

Tabela 11: Comparação do modelo *multi-step* para predição de 5 dias à frente com método Naïve

Modelos	MAE (R\$)	MAPE %	RMSE
Método Naïve	7,60	10,41	11,31
Modelo <i>multi-step</i> para predição de 5 dias à frente	5,44	8,26	6,60

Fonte: elaborado pelo próprio autor.

Tabela 12: Comparação do modelo *multi-step* para predição de 7 dias à frente com método Naïve

Modelos	MAE (R\$)	MAPE %	RMSE
Método Naïve	8,73	11,99	12,31
Método <i>single-step</i> para predição de 7 dias à frente	8,86	12,09	11,38
Modelo <i>multi-step</i> para predição de 7 dias à frente	7,00	10,48	8,40

Fonte: elaborado pelo próprio autor.

Todos os modelos *multi-step* foram capazes de obter métricas melhores que as predições utilizando o método Naïve. Já para o método *single-step*, as predições foram piores em termos de MAE e MAPE, e melhores em termo de RMSE. Isso pode ser explicado pelo fato da repetição do último valor como predição ocasionar em alguns pontos erros de maior valor absoluto, o que aumenta consideravelmente o RMSE.

Para o método *single-step*, nota-se na seção 4.4.1 que quanto maior a distância entre o passo t (último valor do conjunto de entrada), e o passo $t + n$ (saída alvo, valor a ser predito), maior a defasagem da série do PLD realizada e a série predita em janela móvel. Os modelos não foram eficientes em antecipar movimentações da curva do PLD. Isto pode ser explicado pelo fato da diminuição da correlação (ou autocorrelação) de um passo de uma série temporal, com o passo anterior, a medida que a distância do passo anterior aumenta.

Além disso, comparando os dois métodos para predição de mais de um dia à frente, a principal diferença é que as células neuronais das saídas dos modelos *single-step*, que predizem cada passo temporal, recebem entradas diferentes, enquanto no método *multi-step*, em um mesmo modelo, as células neuronais de saída, para cada um dos passos, recebem exatamente a mesma entrada (por estarem na mesma rede). Efetivamente, isso faz com que os processos de treinamento das redes sejam completamente distintos.

No caso do método *single-step*, o *backpropagation* (processo de atualização dos pesos matriciais das células neuronais na direção da saída para entrada) irá ocorrer buscando minimizar o erro da predição de somente um passo (saída unitária), e por conta da diminuição da autocorrelação em função da distância do passo a ser predito, faz sentido que a otimização dos pesos matriciais durante o treinamento, resultem em uma rede que produza um resultado atrasado em relação à série realizada.

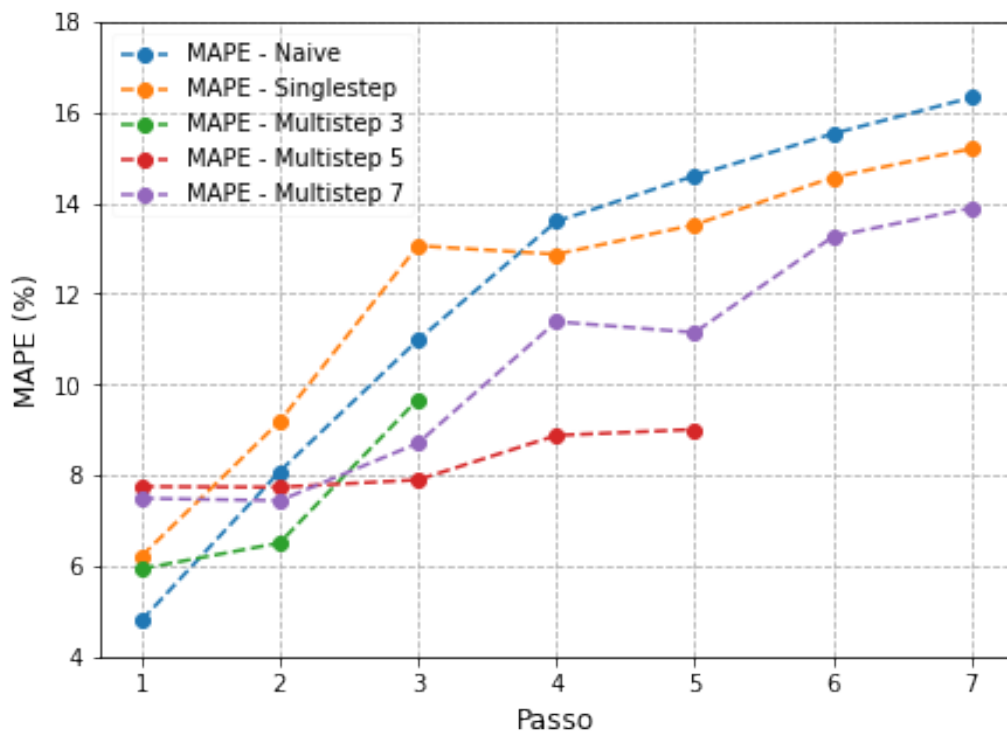
Já para o método *multi-step*, o *backpropagation* irá ocorrer buscando minimizar o erro da predição de mais de um passo temporal em conjunto, e somente na camada de saída é que cada célula neuronal tem seus pesos matriciais atualizados buscando a minimização do

erro para cada passo. Esta diferença proporcionou um melhor desempenho a este método.

5.2.2 Dinâmica do erro médio percentual

Para a predição de mais de um dia à frente, o método *multi-step* obteve melhores resultados do que o método *single-step*, e esta vantagem fica mais evidente ao analisar o erro médio percentual em função dos passos preditos. Na Figura 32 é exposto o gráfico do MAPE em função do passo de predição para os diferentes modelos.

Figura 32: Erro médio percentual (MAPE) em função do passo de predição para cada modelo



Fonte: elaborado pelo próprio autor.

Como é esperado, todos os modelos apresentam uma tendência de aumento do erro médio percentual em função do passo predito. Porém a partir deste gráfico, fica explícito que a grande vantagem dos modelos sobre o método Naïve é obtido com o aumento dos passos de predição. Para o método *single-step*, esta vantagem só é obtida a partir do quarto passo de predição, e a curva fica próxima a curva do método Naïve, por isso no MAPE total dos 7 dias, obteve um resultado pior. E para o método *single-step*, os 3 modelos apresentaram MAPEs menores que o método Naïve já a partir do segundo dia de predição.

5.3 Comparação com a literatura

Retornando à revisão da literatura, nota-se que a predição de séries temporais é um tema frequentemente abordado no setor elétrico de diversos países. Dos trabalhos revisados, há três trabalhos que tinham como objetivo a predição da série do PLD, por meio de modelos clássicos de machine learning. Estes trabalhos são as publicações de Nascimento, Olivi e Lima [22], Santos [23] e Lopes [24]. Dadas as algumas diferenças, é possível a comparação dos resultados com os trabalhos de Santos e Lopes. Com o trabalho de Nascimento, Olivi e Lima, não é possível comparação dos resultados, pois os erros foram expressos em termos de MAE somente.

Santos [23] em seu trabalho, utiliza redes neurais LSTM para a predição do PLD. As séries de PLD utilizadas, tinham discretização semanal, e ainda apresentavam o modelo antigo da CCEE, em que os valores do PLD eram dados em patamares de carga, e correspondiam aproximadamente as médias semanais dos modelo atual de divulgação do PLD. As predições foram feitas para um passo à frente. Considerando todas as séries, os modelos de Santos obtiveram um MAPE de 9,66%. Dada a discretização semanal, há duas possíveis comparações com este trabalho: primeiro a comparação considerando o número de passos preditos, e segundo a comparação considerando o intervalo temporal predito. Comparando os resultados da predição de um passo à frente, o melhor modelo da seção 4.3 obteve MAPE de 5,81%, o que é um resultado consideravelmente melhor. Já comparando o intervalo temporal de predição de 7 dias, o melhor modelo e método da seção 4.4 obteve MAPE de 10,48%, o que é um valor pior. Infere-se que esta diferença ocorra por conta da discretização, é esperado que modelos de predição que utilizem discretização semanal com valores médios semanais, apresentem melhor acurácia. Desta forma, mesmo considerando como comparação o intervalo temporal de predição, o resultado obtido é considerado como um resultado em linha com o de Santos.

Lopes [24] usa 4 diferentes modelos de machine learning para predição da série do PLD para o submercado Sul, para 7 dias à frente. O melhor modelo foi rede neural convolucional com MAPE de 24,70%. Como mencionado, o melhor modelo para predição de 7 dias à frente obteve MAPE de 10,48%. Esta expressiva diferença pode ser explicadas por alguns fatores: Lopes treinou a rede com dados de somente 12 meses, portanto há menos informação temporal no treinamento; no trabalho de Lopes foi utilizado discretização horária, e neste trabalho discretização diária com a média horária de cada dia, logo, seguindo a lógica do parágrafo anterior, a discretização adotada neste trabalho favoreceria a acurácia; e por fim, há a diferença do modelo de rede neural utilizada, pela formulação

teórica das redes, infere-se que uma rede neural LSTM tenha melhor capacidade preditiva que uma rede neural puramente convolucional.

Em relação a todos os trabalhos analisados, de previsões de séries temporais aplicadas ao sistema elétrico brasileiro, não há uma exposição do erro em função do passo predito, como foi exposto na subseção 5.2.2. Esta análise se mostrou fundamental para o entendimento do comportamento do erro e da vantagem de cada método e modelo em relação a um método básico referencial (Naïve), porém não será possível a comparação desta análise, com outras publicações da literatura.

Em síntese, o trabalho obteve resultados em linha ou melhores que os trabalhos comparáveis da literatura, ao se ponderar os diferentes métodos. Com o desenvolvimento do mercado livre de energia elétrica brasileiro, publicações com estes temas deverão se tornar cada vez mais frequentes, para possibilidade de comparação entre os métodos de cada trabalho, sugere-se a exposição das métricas de avaliação tanto em função do intervalo temporal predito, como em função dos passos de predição. Estas duas informações devem facilitar comparações de séries com discretizações diferentes.

6 CONCLUSÕES E TRABALHOS FUTUROS

O objetivo principal do trabalho foi atingido, a predição do PLD em um intervalo crescente de 1 a 7 dias. Os requisitos de implementação foram cumpridos, os resultados expressos graficamente, e os modelos foram comparados por meio das métricas de avaliação. Ainda foi possível analisar as diferentes variáveis elegidas a serem preditoras da série do PLD.

Sobre a análise das variáveis, por meio das correlações, do PCA e da comparação dos testes para predição de um dia à frente com a variável de carga e sem ela, ficou evidente a maior importância da Energia Natural Afluente e da Energia Armazenada, para predição do PLD, em detrimento da série de carga. O caráter ainda predominantemente hidráulico da geração elétrica do sistema brasileiro, implica em maior relevância das variáveis hidrológicas para a predição do preço da energia elétrica. Além disso, este fato pode ser explicado pelo bom planejamento de segurança energética nacional, realizada pelos órgãos governamentais como ONS, EPE e ANEEL, que faz com que os níveis dos reservatórios sejam capazes de absorver as variações nos patamares de carga do sistema.

A respeito dos modelos preditores, conclui-se que o modelo ARIMA não é um bom preditor para a série do PLD. Os modelos de redes neurais LSTM ganham vantagem sobre o ARIMA, por terem maior capacidade de modelar o comportamento não estacionário da série. Através do sistema de *gates* das células neuronais das redes LSTM, há a possibilidade de atribuir pesos diferentes às entradas vindas das séries temporais em intervalos distintos.

Para predição de um dia à frente o melhor preditor encontrado foi o método Naïve (último valor realizado da série), porém de 2 a 7 dias alguns modelos apresentaram resultados melhores que este método. Apesar do erro médio percentual crescer com o aumento do número de passos preditos, como apresentado na subseção 5.2.2, o MAPE do método Naïve cresce a uma taxa maior que para os modelos LSTM *multi-step*. Assim, os modelos LSTM *multi-step* são úteis para a predição de mais de um dia à frente, sendo que para a predição do primeiro dia à frente, o melhor preditor encontrado foi o método Naïve.

Por fim, a partir dos trabalhos comparáveis apresentados na revisão da literatura, consideradas as diferenças apresentadas na seção 5.3, o trabalho obteve resultados de predição do PLD em linha ou melhores que os reportados nestes trabalhos. Um ponto que traria melhor comparabilidade dos trabalhos, seria a exposição do MAPE por passo predito, como apresentado também na subseção 5.2.2. Ainda há muitas extensões de estudos a serem exploradas em trabalhos futuros, algumas indicações serão feitas no capítulo seguinte.

6.1 Trabalhos futuros

Em vista da complexidade do setor elétrico brasileiro, há muitas possibilidades de aprofundamento e continuidade deste trabalho, neste capítulo final serão sugeridas algumas delas.

Inicialmente, vale destacar que a característica de predominância hidrelétrica da matriz de geração brasileira não é imutável, em vista do maior rigor nas licenças ambientais dos projetos de geração, e das políticas de incentivo a investimentos em energia eólica e solar, a matriz energética brasileira está em mudança. Portanto, trabalhos futuros de predição do preço da energia elétrica no mercado brasileiro devem sempre reavaliar a seleção das variáveis preditoras.

No cenário atual, como o preço da energia elétrica é correlacionado com dados hidrológicos, um estudo útil para um trabalho futuro, seria a possibilidade do refino dos modelos preditivos, com o acréscimo de variáveis de entrada de previsão meteorológica. Para o desenvolvimento deste trabalho, as principais fontes de dados foram a CCEE e o ONS, porém foram visitadas também as bases de dados de institutos como a Agência Nacional de Águas e Saneamento (ANA), Instituto Nacional de Meteorologia (INMET) e Instituto Nacional de Pesquisas Espaciais (INPE). Em nenhum dos institutos, foi encontrado uma base de dados do histórico de previsões meteorológicas, somente bases de dados realizados (observados) das estações meteorológicas. No caso do INPE, há histórico dos boletins técnicos onde constam as previsões, esta poderia ser uma fonte para a criação de uma base de dados, ou caso no futuro seja publicado uma base de dados histórica das previsões, por um destes institutos, poderia ser utilizada no refino dos modelos.

Pensando na possibilidade da criação de um sistema automatizado de predição do PLD, como os dados de entrada dos modelos são provenientes de mais de uma instituição, seria útil a implementação de uma base dados única, sob metodologia ETL (*extract, transform,*

load). Uma vantagem na implementação desta base seria a possibilidade do agendamento da aquisição e tratamento dos dados das diversas fontes, permitindo que as variáveis de entrada do modelo fossem atualizadas e transformadas rotineiramente. Outro benefício seria a maior robustez trazida ao processo, uma vez que os dados seriam obtidos pelos modelos de uma base única, e não haveria manipulação humana na aquisição destes dados, evitando erros.

Finalmente, uma última indicação de trabalho futuro, é o desenvolvimento de um sistema, como um web site ou aplicativo, em que houvesse a publicação das previsões do PLD. Caso o PL414/2021 avance, ampliando o número de consumidores elegíveis a opção pelo mercado livre, passaria a ser mais demandado tal sistema. Atualmente, há algumas plataformas como o *Pool de Denergia* do Dcide [32], em que é vendido o acesso aos dados das expectativas consolidadas, de diversos agentes, sobre o preço futuro da energia elétrica. Um sistema como o proposto, seria uma alternativa ou complementação a esses.

REFERÊNCIAS

- [1] EPE. *Balanço Energético Nacional 2022*. 2022. Disponível em: <https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-675/topico-631/BEN_S%C3%ADntese_2022_PT.pdf>.
- [2] SANTOS, T. N. et al. Hourly pricing and day-ahead dispatch setting in brazil: The dessem model. *Electric Power Systems Research*, Elsevier, 2020.
- [3] CYRILLO, Y. M. *Avaliação do modelo PVARm interconfigurações para geração de cenários de ENA no planejamento da operação de médio prazo*. Dissertação (Mestrado) — Pontifícia Universidade Católica do Rio de Janeiro, 2018.
- [4] PINHEIRO, C. Guia do mercado livre de energia para consumidores livres e especiais. Universidade Federal do Ceará, 2018.
- [5] WALVIS, A. *Avaliação das reformas recentes no setor elétrico brasileiro e sua relação com o desenvolvimento do mercado livre de energia*. Dissertação (Mestrado) — Fundação Getúlio Vargas, 2014.
- [6] Conceitos de Preços - CCEE. Disponível em: <<https://www.ccee.org.br/web/guest/precos/conceitos-precos>>.
- [7] EPE. Lastro & Energia. *Comitê de Implantação da Modernização do Setor Elétrico*, 2019. Disponível em: <<https://www.epe.gov.br/sites-pt/areas-de-atuacao/energia-eletrica/Documents/Guia%20de%20Perguntas%20%28Lastro%20e%20Energia%29-03.03.2021.pdf>>.
- [8] CCEE. Relatório anual de administração 2021. 2022. Disponível em: <<https://www.ccee.org.br/relatoriodeadministracao/relatorio-anual-administracao-ccee-2021.pdf>>.
- [9] GUTIERREZ, M. Reformando o setor elétrico brasileiro: uma aproximação com os países da ocde. *IPEA*, 2022.
- [10] Analysis search results - Electricity price forecasting. 2022. Disponível em: <<https://www.scopus.com/term/analyzer.uri?sid=80448a04a760021bf8b0dc60b1dab47f&origin=resultslist&src=s&s=TITLE-ABS-KEY%28electricity+price+forecasting%29&sort=plf-f&sdt=b&sot=b&sl=44&count=3127&analyzeResults=Analyze+results&txGid=0f9aed2fc982c6b8b199c32b58de3092>>.
- [11] HOFFMAN, K. C.; WOOD, D. O. Energy system modeling and forecasting. *Annual Review of Energy*, 1976.
- [12] GHODDUSI, H.; CREAMER, G. G.; RAFIZADEH, N. Machine learning in energy economics and finance: A review. *Energy Economics*, North-Holland, 2019.

- [13] UGURLU, U.; OKSUZ, I.; TAS, O. Electricity price forecasting using recurrent neural networks. *MDPI Energies*, 2018.
- [14] TUTTLE, D. P. et al. The history and evolution of the u.s. electricity industry. *The University of Texas*, 2016.
- [15] TAYLOR, L. D. The demand for electricity: a survey. *The Bell Journal of Economics*, 1975.
- [16] HONG, Y. Y.; HSIAO, C. Y. Locational marginal price forecasting in deregulated electric markets using a recurrent neural network. *Proceedings of the IEEE Power Engineering Society Transmission and Distribution Conference*, v. 2, 2001.
- [17] RAJAN, P.; CHANDRAKALA, K. R. V. Statistical model approach of electricity price forecasting for indian electricity market. *Proceedings of the IEEE Madras Section International Conference 2021, MASCON 2021*, Institute of Electrical and Electronics Engineers Inc., 2021.
- [18] TSCHORA, L. et al. Electricity price forecasting on the day-ahead market using machine learning. *Applied Energy*, Elsevier, 2022.
- [19] MILETIC, M. et al. Day-ahead electricity price forecasting using LSTM networks. *2022 7th International Conference on Smart and Sustainable Technologies*, 2022.
- [20] AGUILAR, S.; OLIVEIRA, D.; SOUZA, R. C. Previsão da curva de carga horária do sistema interligado nacional (SIN) utilizando métodos de suavização exponencial. *LIII Simpósio Brasileiro de Pesquisa Operacional*, 2021.
- [21] GONTIJO, T. S.; COSTA, M. A.; SANTIS, R. B. D. Electricity price forecasting on electricity spot market: A case study based on the brazilian difference settlement price. In: . [S.l.]: EDP Sciences, 2021. v. 239.
- [22] NASCIMENTO, F. B. D.; OLIVI, L. R.; LIMA, L. H. L. Multi-step forecasting of the brazilian energy market price using autoregressive neural networks. *SBSE 2018 - 7th Brazilian Electrical Systems Symposium*, Institute of Electrical and Electronics Engineers Inc., 2018.
- [23] SANTOS, G. *Uma aplicação de redes neurais recorrentes do tipo LSTM à previsão dos preços de curto prazo do mercado de energia elétrica brasileiro*. Dissertação (Mestrado) — Fundação Getúlio Vargas, São Paulo, 2019.
- [24] LOPES, G. H. A. Predição do preço de liquidação das diferenças do submercado sul a partir de variáveis hidrológicas. Universidade Federal de Santa Catarina, 2022.
- [25] EPE. *Reunião Ampla com os Agentes CPAMP*. 2020. Disponível em: <<https://abraceel.com.br/wp-content/uploads/post/2020/06/Webinar-GT-Metodologia.pdf>>.
- [26] CCEE. *Painel de Preços*. Disponível em: <<https://www.ccee.org.br/web/guest/precos/painel-precos>>.
- [27] ONS Dados Abertos. Disponível em: <<https://dados.ons.org.br/>>.

- [28] HIRST, G.; GOLDBERG, Y. *Neural Network Methods for Natural Language Processing*. [S.l.]: Morgan & Claypool Publishers, 2017.
- [29] CHOLLET, F. *Deep Learning with Python*. [S.l.]: Manning Publications Co., 2018.
- [30] ASSI, P. N. V. B. *Repositório dos códigos de desenvolvimento do projeto*. Disponível em: <<https://github.com/Pablo-Nabil/predicaoPLD>>.
- [31] BARBER, D. *Bayesian Reasoning and Machine Learning*. [S.l.]: UCL, 2020.
- [32] DCIDE. *Pool Denergia*. Disponível em: <<https://www.dcide.com.br/produtos/pool-denergia/>>.