UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE FÍSICA E ESCOLA DE ENGENHARIA
CURSO DE ENGENHARIA FÍSICA

LEONARDO MACHADO BARCELOS

Sistema para automatização do processo de limpeza dos dados de torres meteorológicas para estudos pré-construtivos, certificação do recurso eólico e da produção de energia de um parque eólico

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Engenharia Física

Orientador: Prof. Dr. Thiago Lopes Trugillo da Silveira

Porto Alegre
2024

AGRADECIMENTOS

Agradeço inicialmente aos meus pais Rosane e Miguel, pelo incondicional apoio ao longo de toda a minha trajetória acadêmica, possibilitando que eu fizesse o curso de Engenharia Física, apesar de todos os percalços emocionais e financeiros. Além disso, agradeço aos meus irmãos Vitor e Mylenna por sempre me encorajar a buscar os meus sonhos.

Agradeço a minha prima Mariane, por abrir os meus olhos e me incentivar a buscar uma área que fizesse sentido para mim, quando eu aos vinte e dois anos pensava ser velho demais para mudar de área.

Agradeço ao meu orientador, Prof. Dr. Thiago Lopes Trugillo da Silveira, por aceitar orientar um trabalho de Engenharia Física.

Agradeço a empresa DNV, por disponibilizar o software *WindFarmer: Analyst* para a marcação dos dados meteorológicos.

Por fim, Agradeço à Universidade Federal do Rio Grande do Sul (UFRGS), aqui inclusos o corpo docente e de técnicos administrativos, que possibilitaram a minha formação gratuita e de qualidade ao longo dos últimos seis anos.

RESUMO

O crescimento da população e o aumento na demanda por eletricidade, juntamente com os esforços para reduzir as emissões de carbono nas fontes de energia, proporcionam um cenário promissor para o desenvolvimento de tecnologias no setor de energias renováveis. Além disso, um fator determinante para acelerar a transição energética é a meta estabelecida no Acordo de Paris das Nações Unidas. Para limitar o aumento da temperatura média global a 1,5°C em relação aos níveis pré-industriais, é crucial alcançar emissão zero de gases do efeito estufa até 2050. Especificamente no setor de energia eólica, estudos pré-construtivos desempenham um papel fundamental no desenvolvimento de novos parques eólicos, abrangendo desde a fase de viabilidade técnica e econômica até a certificação do recurso eólico. Com o objetivo de facilitar o rápido avanço de projetos de energia eólica na fase pré-construtiva, este trabalho se propõe a desenvolver um sistema para a limpeza de dados provenientes de torres anemométricas. Esse sistema serve como uma ferramenta valiosa para complementar o trabalho dos Engenheiros Eólicos. Foram utilizados dados públicos de torre anemométrica situada na Dinamarca. Para marcação de períodos anômalos, foi desenvolvido um Protocolo com diretrizes sobre quando uma medida deve ser marcada como anômala. Avaliou-se cinco abordagens – PCA, *iForest*, LSTM, TranAD e MSCRED – junto com um algoritmo de força bruta, que implementa o Protocolo, sendo que a LSTM apresentou os melhores resultados. Foram avaliadas as métricas Precisão, Revocação e *F1-Score*. Além disso, foi adotado um pós-processamento dos resultados das abordagens avaliadas e métricas ajustadas também são apresentadas. Com isso, a LSTM apresentou um *F1-Score* médio, com pós-processamento, de 0,8.

Palavras-chave: Avaliação de produção de energia. energia eólica. torres meteorológicas. dados de vento.

System for automating the data cleaning process of meteorological towers for pre-construction studies, wind resource certification, and wind farm energy production

ABSTRACT

The growth of the population and the increase in demand for electricity, along with efforts to reduce carbon emissions in energy sources, provide a promising scenario for the development of technologies in the renewable energy sector. Additionally, a key factor in accelerating the energy transition is the goal set in the United Nations' Paris Agreement. To limit the increase in the global average temperature to 1.5°C above pre-industrial levels, it is crucial to achieve zero greenhouse gas emissions by 2050. Specifically in the wind energy sector, pre-construction studies play a fundamental role in the development of new wind farms, covering everything from the technical and economic feasibility phase to the certification of wind resources. With the aim of facilitating the rapid advancement of wind energy projects in the pre-construction phase, this work proposes to develop a system for cleaning data from anemometric masts. This system serves as a valuable tool to complement the work of Wind Engineers. Public data from anemometric mast located in Denmark were used. For labeling anomalous periods, a Protocol was developed with guidelines on when a measurement should be marked as anomalous. Five approaches – PCA, iForest, LSTM, TranAD, and MSCRED – were evaluated along with a brute-force algorithm that implements the Protocol, with LSTM showing the best results. Precision, Recall, and F1-Score metrics were evaluated. Additionally, post-processing of the results of the evaluated approaches was adopted, and adjusted metrics are also presented. Thus, LSTM showed an average F1-Score, with post-processing, of 0.8.

Keywords: Energy production assessment, wind energy, meteorological mast, wind data, wind resource certification.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABEEólica Associação Brasileira de Energia Eólica.

ABNT Associação Brasileira de Normas Técnicas.

AWEA *American Wind Energy Association*.

BPTT *Back Propagation Through Time*.

CFD *Computational Fluid Dynamics*.

CLA Camada Limite Atmosférica.

DNS *Direct Numerical Simulation*.

DTU *Technical University of Denmark*.

EPE Empresa de Pesquisa Energética.

ETO *Energy Transition Outlook*.

IEC *International Electrothecnical Commission*.

INMETRO Instituto Nacional de Metrologia, Normalização e Qualidade Industrial.

LES *Large Eddy Simulations*.

LIDAR *Light Detection and Ranging*.

LSTM *Long Short-Term Memory*.

MCP Medir, Correlacionar e Prever.

MEASNET *Network of European Measuring Institutes*.

MME Ministério de Minas e Energia.

MSCRED *Multi-Scale Convolutional Recurrent Encoder-Decoder*.

MTAD *Multivariate Time Series Anomaly Detection*.

PCA *Principal Component Analysis*.

PMMA Protocolo de Marcação de Medidas Anômalas.

RANS *Reynolds Average Navier Stokes*.

RNN *Recurrent Neural Network*.

SMD *Server Machine Dataset*.

SODAR *Sound Detection And Ranging*.

SWaT *Secure Water Treatment*.

TranAD *Transformer-based Anomaly Detection*.

WAsP *Wind Atlas Analysis and Application Program*.

SUMÁRIO

1 INTRODUÇÃO

O crescimento populacional e o progressivo aumento da demanda por energia elétrica, associado aos esforços de de-carbonização das matrizes energéticas faz com que o cenário das energias renováveis seja promissor (DNV, 2022). Além disso, outro fator determinante para a aceleração da transição energética é a emergência climática, que cada vez mais vem sendo uma realidade do presente, e não mais uma previsão para o futuro (HOEGH-GULDBERG et al., 2018).

Conforme o Acordo de Paris da Organização das Nações Unidas (UNFCCC, 2015), os 196 países signatários, incluindo o Brasil, se comprometem a contribuir com a redução da emissão de gases de efeito estufa no contexto do desenvolvimento sustentável. Essa redução objetiva manter o aumento na temperatura média global abaixo dos 2°C em comparação com níveis pré-industriais e limitar esse crescimento em 1,5°C, reduzindo as emissões pela metade até 2030 e zerando até 2050 (NIK; PERERA, 2020).

De acordo com DNV (2022) em seu *Energy Transition Outlook* (ETO), relatório com resultados de um modelo independente do sistema de energia mundial, devemos ter um aumento na demanda energética de pelo menos 13% até 2050. O relatório adiciona que, se continuarmos com a tendência de emissão que temos hoje, só atingiremos emissão zero ao final do século, conforme Figura 1.1.

Para que a crescente demanda energética seja atendida respeitando o Acordo de Paris e limitando o aumento da temperatura em 1,5°C, precisamos que setores de energias renováveis tripliquem. Ainda segundo DNV (2022):

> Energia solar fotovoltaica e eólica já são as formas mais baratas de geração de energia elétrica em vários lugares do mundo, e, em 2050, vão crescer vinte e dez vezes, respectivamente. Energia solar fotovoltaica contribuirá com 38% da eletricidade gerada e eólica 31%.

Com isso, temos um cenário propício para o aprimoramento e desenvolvimento de tecnologias e processos no setor de energia eólica. Segundo Veers et al. (2019), o futuro das ciências de energia eólica pode ser classificado em três grandes desafios: **(i)** melhor entendimento da física do escoamento atmosférico na zona crítica de operação de um parque eólico, **(ii)** materiais e sistemas dinâmicos de turbinas eólicas individuais, e **(iii)** otimização e controle de parques eólicos contendo centenas de aerogeradores individuais trabalhando sinergicamente dentro de um sistema de uma rede elétrica maior.

Hoje, de acordo com instruções da Empresa de Pesquisa Energética (EPE), órgão associado ao Ministério de Minas e Energia (MME) do Brasil responsável por estudos e pesquisas destinadas a subsidiar o planejamento do setor energético, todo projeto eólico que intenciona participar de leilões de energia elétrica precisa de uma certificação, tanto

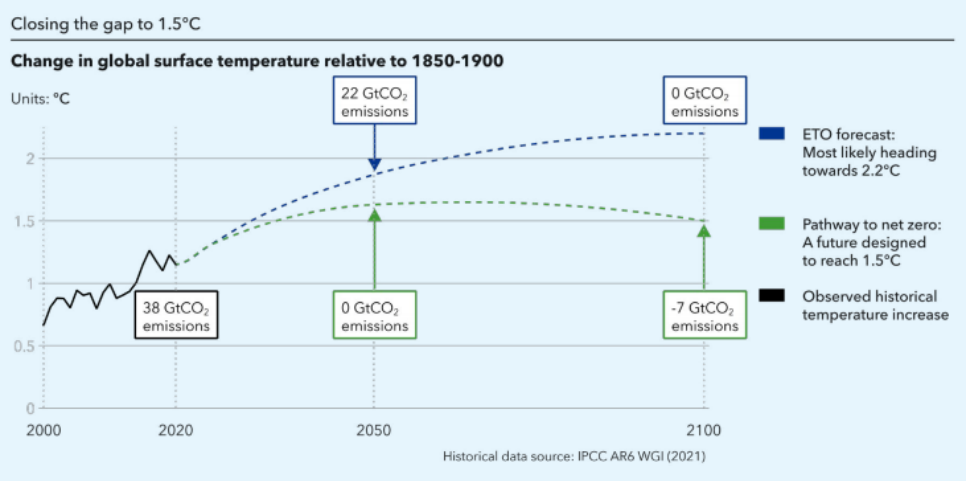


Figura 1.1 – Mudança na temperatura superficial terrestre em relação a níveis pré-industriais. Fonte: (DNV, 2022).

do recurso eólico quanto da produção de energia anual (EPE, 2021). Além disso, as instruções especificam que:

> Todos os procedimentos, critérios, normas e cálculos utilizados nas certificações deverão seguir as recomendações de entidades como a *International Electrothecnical Commission* (IEC), *Network of European Measuring Institutes* (MEASNET), *American Wind Energy Association* (AWEA), Associação Brasileira de Normas Técnicas (ABNT) e Instituto Nacional de Metrologia, Normalização e Qualidade Industrial (INMETRO), entre outras.

A EPE instrui que as estações de medições anemométricas precisam medir pelo menos duas alturas acima de 50 metros, por um período igual ou superior a três anos e uma taxa inferior a 10 porcento de perda de dados (EPE, 2021). Normalmente essas estações de medições são torres anemométricas, que são instaladas com diversos sensores como anemômetros, birutas, termômetros, barômetros, etc.

Analisar e entender como o escoamento acontece em uma região, através de uma estação de medição, é essencial na etapa pré-construtiva de um parque eólico, visto que a geração de energia está diretamente associada à velocidade de vento presente na região em estudo (BURTON et al., 2011). Segundo Custódio (2013), a relação da potência disponível no vento é dada por

$$P = \frac{1}{2} v^3 \rho A, \tag{1.1}$$

onde P é a potência $[W]$, v é a velocidade do vento $[m/s]$, ρ é a massa específica do ar $[Kg/m^3]$ e A é a área varrida pelas pás de um aerogerador $[m^2]$.

Para que seja conduzida uma certificação com qualidade e que seja representativa da região sendo analisada, é necessário que dados errôneos coletados na torre anemométrica sejam identificados e excluídos da série temporal de dados medidos, visto que um valor desviado na velocidade média final do vento da região será elevado ao cubo no

cálculo da energia (GALLON, 2015).

No presente trabalho, objetiva-se atacar o primeiro grande desafio proposto por Veers et al. (2019), através da proposição de um sistema ou processo que facilite e acelere a análise de dados meteorológicos de estações de medição anemométricas, mais especificamente as torres anemométricas. Atualmente, a análise dos dados de direção e velocidade de vento, assim como pressão atmosférica, temperatura e umidade relativa ainda é feita através de inspeção visual das séries temporais em alguns setores do mercado de energia eólica.

1.1 Definição do problema

Considerando que a demanda por energia eólica tem uma tendência de crescimento (DNV, 2022), e dado que a eficiência na análise de informações meteorológicas de torres anemométricas é chave para o rápido desenvolvimento dos projetos eólicos em sua etapa pré-construtiva, urge a criação de um sistema ou processo que torne o trabalho do engenheiro eólico mais automatizado.

Em um processo visual e manual, a etapa de limpeza dos dados é onerosa e em casos onde hajam muitas torres e com campanhas de medição extensas, é razoável pensar que o tempo empregado nessa etapa do trabalho pode ser grande. Espera-se que uma análise mais automatizada auxilie o engenheiro eólico a reduzir o tempo de limpeza dos dados de torres anemométricas, e com isso possibilitar que mais projetos possam ser concluídos em menos tempo, contribuindo para o atendimento à crescente demanda por projetos eólicos.

Uma vez que os períodos de dados passíveis de exclusão são identificados, uma série de exclusão, com a classificação de cada tipo de exclusão pode ser gerada.

1.2 Objetivos e Contribuições

No presente trabalho objetiva-se a construção de um sistema que otimize o tempo de análise de dados anemométricos. Este sistema tem como principais objetivos específicos:

- Identificar períodos de medidas anômalas dos sensores;
- Exportar série temporal de classificação de exclusão.

O trabalho contribui tanto para área de análise de dados meteorológicos quanto para a área de pesquisa em energia eólica. As principais contribuições são:

- Definição de um protocolo de marcação para exclusão de dados meteorológicos utilizados em estudos eólicos pré-construtivos;
- Rótulos dos períodos anômalos de uma torre anemométrica situada na ilha de Kegnæs na Dinamarca (HANSEN; VASILJEVIC; SøRENSEN, 2021);
- Aplicação, avaliação e análise de modelos estatísticos e de aprendizado de máquina em séries temporais multivariadas de dados meteorológicos;

As contribuições deste trabalho podem ser acessadas em um repositório aberto do Github[1].

1.3 Organização dos Capítulos

Na sequência, será apresentada uma fundamentação teórica na Seção 2 detalhando: 1) os principais atores do mercado de energia eólica, evidenciando aqueles que podem se beneficiar do trabalho; 2) o recurso eólico e as suas principais características; e 3) as etapas de uma análise do recurso eólico e onde o trabalho se insere.

Na Seção 3, são analisados trabalhos relacionados à detecção de anomalias em séries temporais multivariadas, perpassando pelas principais técnicas, desde as convencionais até as que empregam modelos compostos com redes neurais profundas. Estes trabalhos servem de inspiração para a metodologia proposta, que é apresentada na Seção 4, onde são detalhados os passos tomados para marcação das medidas, preparação dos conjuntos de dados e avaliação dos modelos selecionados.

Por fim, os resultados e discussões são descritos na Seção 5. Nesta Seção, são apresentadas a métricas que possibilitam a comparação entre os modelos. As principais dificuldades encontradas ao longo da execução e as conclusões são discutidas na Seção 6.

[1] https://github.com/barcelosleo/met-mast-anomaly-detection

2 FUNDAMENTAÇÃO TEÓRICA

Como mencionado na Seção 1, de acordo com EPE (2021), todo empreendimento eólico que desejar participar dos leilões de energia elétrica precisa de uma certificação. Além disso, no mercado de energia eólica, é comum que projetos em sua etapa pré-construtiva sejam comprados e vendidos entre empresas desenvolvedoras, e para isso, a certificação é o documento que atesta o potencial do projeto.

A fim de que se entenda onde o trabalho proposto se enquadra em todo o cenário do mercado de energia eólica, serão feitas algumas definições a seguir.

2.1 Atores do mercado de energia eólica

No Brasil, existe a Associação Brasileira de Energia Eólica (ABEEólica), instituição sem fins lucrativos, que congrega e representa a indústria de energia eólica no país, incluindo empresas de toda a cadeia produtiva. Em seu relatório anual de 2021 (ABEEÓLICA, 2021), ela classifica os seus associados nas seguintes categorias:

1. Empreendedores, desenvolvedores e geradores;
2. Fabricantes de aerogeradores de grande porte;
3. Engenharia, consultoria e construção;
4. Fabricante de peças e componentes;
5. Fabricantes de pás eólicas;
6. Logística, montagem e transporte;
7. Comercializadores de energia;
8. Construção civil;
9. Federações;
10. Instituto de Pesquisa, Universidades e Centros de Estudo.

Tendo isso em mente, o presente trabalho tende a focar nos atores que lidam com as etapas iniciais de um projeto eólico, a exemplo dos Atores 1, 2 e 3. Os Atores 1 precisam estudar o recurso eólico das regiões onde têm interesse em implantar um parque eólico; os Atores 2 precisam entender os extremos climáticos da região para a qual pretendem instalar um parque, a fim de fornecer as máquinas mais adequadas; e os Atores 3 podem prestar serviços de análise do recurso eólico mais especializados.

2.2 Recurso eólico

De acordo com Burton et al. (2011), do ponto de vista da energia eólica, a característica notável do vento é a sua variabilidade. O vento é altamente variável, tanto espacialmente quanto temporalmente. A importância dessa variabilidade no contexto de geração de energia elétrica é amplificada pela relação cúbica da velocidade do vento com a energia potencial disponível (BURTON et al., 2011).

Em larga escala, a variabilidade espacial descreve o fato de que há diferentes regiões climáticas no mundo, onde em algumas venta mais (BURTON et al., 2011). Burton et al. (2011) adiciona que essas regiões são diferenciadas principalmente pela latitude, que interfere na quantidade de insolação que essa região recebe. Dentro de cada região climática, há grande variação em uma menor escala, principalmente causada pela geografia local – a proporção de terra e mar, o tamanho de massas de terra, e a presença de montanhas ou planícies por exemplo (BURTON et al., 2011). Mais especificamente, de acordo com Freire (2012), a região atmosférica afetada pela interação física e térmica com a superfície terrestre é definida como Camada Limite Atmosférica (CLA).

2.2.1 Efeitos de topografia e rugosidade

Segundo Burton et al. (2011), em larga escala espacial, os ventos são originados principalmente pela irradiação solar, onde regiões próximas a do equador recebem maior intensidade solar, e com a rotação da terra essas massas de ar aquecido se movimentam. O ar quente circula até a atmosfera e desce para regiões mais frias (BURTON et al., 2011).

Custódio (2013) ainda adiciona que, o movimento de larga escala do ar é fortemente influenciado pelas forças de Coriolis causadas pela rotação da Terra. Isso forma um padrão de circulação de escala global, como pode ser visto na Figura 2.1.

Devido a não-uniformidade da superfície terrestre, com padrões de massas de terra e massas de oceanos, esse padrão de circulação global é perturbado em escalas menores, como numa escala continental. Essas interações geram variações altamente complexas e resultam em um cenário caótico de escoamento do ar (BURTON et al., 2011). Segundo Custódio (2013), as escalas espaciais podem ser classificadas conforme a Tabela 2.1.

De acordo com Burton et al. (2011), em uma menor escala espacial, ao longo de colinas e montanhas, há um aumento na velocidade do vento, devido ao aumento na altitude. Isso se deve ao fato de esses acidentes geográficos projetarem o escoamento

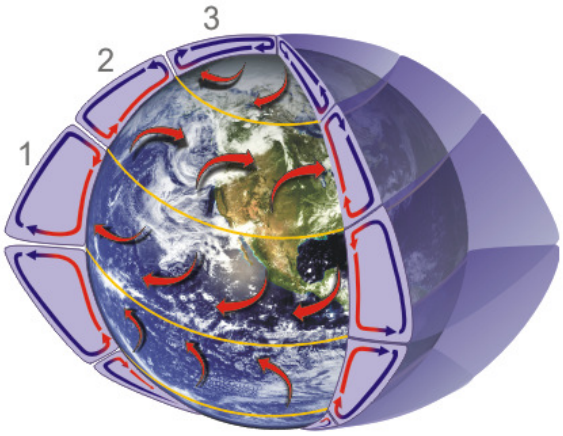


Figura 2.1 – Padrão de circulação global. Fonte: (NOAA, 2023).

Tabela 2.1 – Escalas espaciais dos movimentos atmosféricos.

Escala	Comprimento [km]	Exemplos
Circulação global	1.000 a 40.000	ventos de movimento, correntes de vento (ventos alísios)
Escala sinóptica	100 a 5.000	ciclones, anticiclones, furacão
Mesoescala	1 a 100	tornados, temporais, brisas
Microescala	<1	turbulências, rajas

para regiões de mais alta velocidade de vento (BURTON et al., 2011). Além disso, em locais mais altos, acontece a compressão das linhas de fluxo do vento e, com isso, há uma aceleração do vento, de acordo com o princípio de Bernoulli (SUN; SUN, 2015). Como podemos ver na Figura 2.2, a medida que o vento escoa sobre um cume, por exemplo, as linhas de fluxo tendem a ser comprimidas em seu topo, e isso faz com que a velocidade no topo seja mais alta.

Ademais, fatores como a cobertura vegetal, presença de cidades ou vilarejos, lagos ou rios, etc. também influenciam no escoamento atmosférico (BURTON et al., 2011). Esses fatores compõem a rugosidade superficial do terreno, e de acordo com Gallon (2015) ela é definida pelo tamanho e pela distribuição desses.

Assim, a influência da cobertura e a topografia da superfície vão influenciar o movimento das camadas de ar na CLA, gerando um perfil vertical do vento (BURTON et al., 2011). As camadas mais baixas de ar normalmente possuem menores velocidades, devido

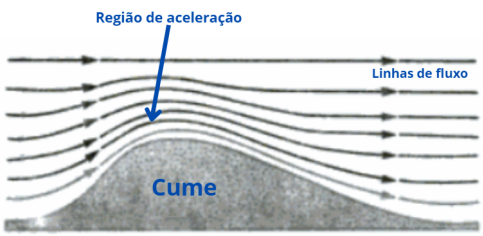


Figura 2.2 – Linhas de fluxo de vento sobre um cume. Fonte: Adaptado de (ELDRIDGE, 1980).

a forte interação com a topografia e os elementos de rugosidade, e vão aumentando gradativamente com o aumento da altitude (NFAOUI, 2012). A intensidade da variação vertical do vento é caracterizada como um cisalhamento, ou *wind shear*, e pode ser descrita por uma lei de potência (BURTON et al., 2011):

$$U_H = U_h \left(\frac{H}{h} \right)^{\alpha} \tag{2.1}$$

onde U_H é a velocidade do vento em uma camada superior, U_h é a velocidade do vento em uma camada inferior, H é a altura da camada superior, h é a altura da camada inferior e α é o coeficiente de cisalhamento ou *shear*.

2.2.2 Estratificação térmica e estabilidade atmosférica

Conforme discutido na Seção 2.2, o Sol é o principal ator no que diz respeito ao movimento das massas de ar. No entanto, a principal troca de calor acontece com a superfície planetária, visto que a composição de gases atmosféricos é transparente à maior parte do espectro de luz visível (BARRIATTO, 2018). A superfície da Terra absorve a radiação solar e se aquece, e parte dessa radiação é re-emitida para as camadas atmosféricas mais baixas, no espectro do infra-vermelho, aquecendo vapor de água, dióxido de carbono e outros gases que compõem a atmosfera (BARRIATTO, 2018).

Conforme Sucevic and Djurisic (2012), a estabilidade atmosférica pode ser definida como a acentuação ou atenuação nos movimentos verticais de vento. A estratificação térmica dentro da CLA vai definir a classificação da estabilidade atmosférica que pode ser (SUCEVIC; DJURISIC, 2012):

- **Camada Limite Instável ou Convectiva (CLC):** tipicamente acontece durante o dia, quando a radiação solar predomina aquecendo o solo e as camadas de ar mais próximas ao solo são aquecidas de baixo para cima, e se caracteriza pelo movimento de massas aquecidas que sobem da superfície para camadas mais altas da CLA. Curva vermelha na Figura 2.3;
- **Camada Limite Estável (CLE):** normalmente ocorre durante a noite, quando a irradiação do solo predomina e as camadas mais inferiores de ar resfriam e a densidade do ar aumenta. Neste caso, as movimentações verticais de massa de ar são atenuadas. Curva preta na Figura 2.3;
- **Camada Limite Neutra (CLN):** geralmente acontece em dias nublados e/ou com

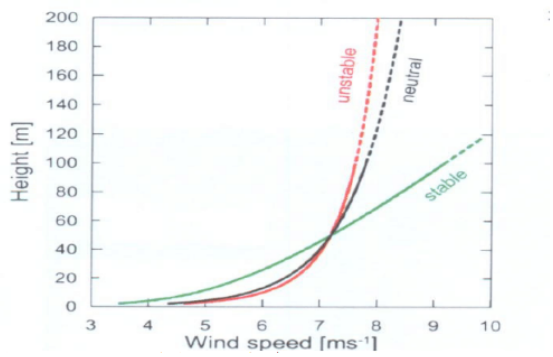


Figura 2.3 – Influência da estabilidade atmosférica no perfil vertical de velocidade do vento. Fonte: (SUCEVIC; DJURISIC, 2012).

ventos fortes, e com isso as movimentações verticais do ar não são nem acentuados e nem atenuados. Além disso, o fluxo vertical de calor é quase zero. Neste caso, os efeitos de estratificação térmica são menos significantes. Curva verde na Figura 2.3.

2.2.3 Variabilidade temporal do vento

Em determinada região, a variabilidade temporal de larga escala pode ser um pouco difícil de prever (BURTON et al., 2011), sendo essas escalas variações de um ano para o outro ou ao longo de décadas. Ainda segundo Burton et al. (2011), em escalas menores do que um ano, as variações já são bem mais previsíveis.

Temos evidências de que a velocidade do vento em qualquer localização particular, pode estar sujeita a lentas variações de longo prazo. Palutikof, Guo and Halliday (1991) demonstraram tendências que estão associadas a variações de temperatura de longo prazo. Além disso, temos correlação do aumento da temperatura global com a atividade humana (ALEXIADIS, 2007) e, sem dúvida, isso mudará o clima e consequentemente os ventos nas próximas décadas.

Também, essas tendências de mudança de longo prazo podem estar associadas a fenômenos globais de clima, como no caso do *el niño*, erupção vulcânicas, etc. (BURTON et al., 2011). Burton et al. (2011) adiciona que, essas variações tendem a adicionar incerteza na previsão do recurso eólico de uma determinada região.

2.3 Etapas de uma análise do recurso eólico e estimativa de geração de energia

Tendo em mente todas as variabilidades do vento, é importante identificar as especificidades climáticas e meteorológicas do vento de uma determinada região para que se obtenha uma estimativa do recurso eólico de longo prazo. De acordo com Custódio (2013), a principal metodologia de estimativa do potencial eólico é baseada em medições de vento realizadas no local em estudo. Para isso, é necessário que sejam feitas campanhas de medição do vento, através da instalação de estações de medição como torres anemométricas, *Light Detection and Ranging* (LIDAR) ou *Sound Detection And Ranging* (SODAR) e o tempo de medição deve ser longo o suficiente para cobrir as variações meteorológicas na região (EPE, 2021). Quanto maior o período de medições, menores serão as incertezas no comportamento do vento no local (CUSTÓDIO, 2013).

Fatores como a topografia, rugosidade e a estratificação térmica vão influenciar o escoamento, e o número de pontos de medição e a distribuição deles ao longo do *site* vão definir o quão bem as variabilidades espaciais em Microescala serão identificadas (CUSTÓDIO, 2013). Dependendo da complexidade do *site*, mais ou menos estações de medições são necessárias, em posições que sejam representativas das futuras turbinas eólicas. Fatores que influenciam na complexidade de um *site* podem ser, por exemplo, a topografia, a presença de florestas e os efeitos de mesoescala (BURTON et al., 2011).

É razoável pensar que terrenos de baixa complexidade topográfica, ou seja, com pouca variabilidade na declividade, normalmente possuem um escoamento mais simples, e, com isso, com poucos pontos de medição é possível prever as condições climáticas na posição das turbinas . Já terrenos mais complexos, como aqueles com morros, montanhas e declividades altas já necessitam de um número maior de pontos de medição, visto que o escoamento se dará de forma mais complexa no *site*.

Entender o regime de escoamento local e os efeitos físicos presentes na região são as etapas iniciais de uma certificação. Conforme podemos ver na Figura 2.4, os resultados da análise dos dados meteorológicos são utilizados nas etapas seguintes de reconstrução, extrapolação e cálculos de energia. O presente trabalho objetiva atuar na etapa indicada em vermelho:

Figura 2.4 – Etapas de uma certificação. Fonte: o autor.

Figura 2.5 – Torre anemométrica treliçada. Fonte: (3DOTENERGY, 2023).

2.3.1 Análise e limpeza de dados meteorológicos

Uma vez que as medições meteorológicas tenham sido realizadas, dados como velocidade e direção do vento, temperatura ambiente, umidade relativa do ar, pressão atmosférica, entre outros devem ser inicialmente tratados (GALLON, 2015).

Em torres anemométricas, as medidas são feitas através de sensores especializados (CUSTÓDIO, 2013): 1) Velocidade: normalmente medida através de anemômetros de copo; 2) Direção: medida através de uma biruta ou *wind vane*; 3) Temperatura e Umidade Relativa: através de um termo-higrômetro; e 4) Pressão: através de um barômetro. Os sensores que realizam as medições ficam expostos às condições climáticas por períodos prolongados, geralmente por pelo menos um ano, e podem apresentar comportamentos anômalos, falhas e degradação. Comumente, torres anemométricas medem, no mínimo, velocidade e direção do vento. Além disso, essas medições podem acontecer em diversas alturas, conforme Figura 2.5, para identificação do perfil vertical de velocidade do vento e/ou identificação de efeitos de estratificação da temperatura.

Já no caso de sensoriamento remoto, temos os SODAR e LIDAR. No SODAR, o vento é medido remotamente por ultrassom e apresenta baixa precisão (CUSTÓDIO, 2013). Já o LIDAR, que possui um custo elevado e apresenta uma melhor precisão, mede o vento por meio de um laser: o vento carrega partículas microscópicas chamadas de

aerossóis, que refletem o laser com um deslocamento devido ao efeito Doppler, proporcional à velocidade do vento (CUSTÓDIO, 2013). Custódio (2013) ainda adiciona que, normalmente, medidas por sensoriamento remoto são utilizadas de forma complementar às medições em torres anemométricas.

Em qualquer tipo de estação de medição, na etapa de tratamento, os dados errôneos devem ser excluídos para que somente a variabilidade do vento seja captada (GALLON, 2015). Como resultado dessa etapa, obtém-se séries temporais dos dados meteorológicos representativos da região e do período de medição.

2.3.2 Reconstrução e extrapolação para o longo prazo

Conforme dito anteriormente, os sensores estão passíveis de falha e, quando isso acontece, os dados errôneos são excluídos. Apesar disso, é comum que os sensores de medição de velocidade e direção tenham redundâncias, a mesma altura ou ao longo de várias alturas da torre. Com isso, utilizando algum método de reconstrução, a exemplo do Medir, Correlacionar e Prever (MCP), é possível sintetizar os períodos de dados faltantes (GALLON, 2015).

Os métodos de MCP são os mais comuns e bem aceitos na área de energia eólica, onde utiliza-se uma fonte de referência, com uma ampla cobertura de dados e representativa da localização de uma fonte alvo, para estimar o recurso eólico desta fonte alvo no longo prazo. De acordo com Carta, Velázquez and Cabrera (2013), no processo de MCP, busca-se quantificar a relação entre a fonte e o alvo no período em que ambas possuem medições concorrentes.

Para estabelecer a relação entre referência e alvo existem diversas metodologias, e uma delas pode ser uma reta de correlação, dada pela seguinte equação:

$$y = ax + b, \tag{2.2}$$

onde a é a inclinação da reta, b é o deslocamento da reta, x e y são respectivamente as variáveis que utilizamos como referência e a que queremos prever que, no contexto de energia eólica, é a direção ou a velocidade do vento. Para estimar os coeficientes a e b, pode-se considerar a técnica de mínimos quadrados baseando-se no período de dados concorrente entre os dados de referência e o alvo (CARTA; VELÁZQUEZ; CABRERA, 2013).

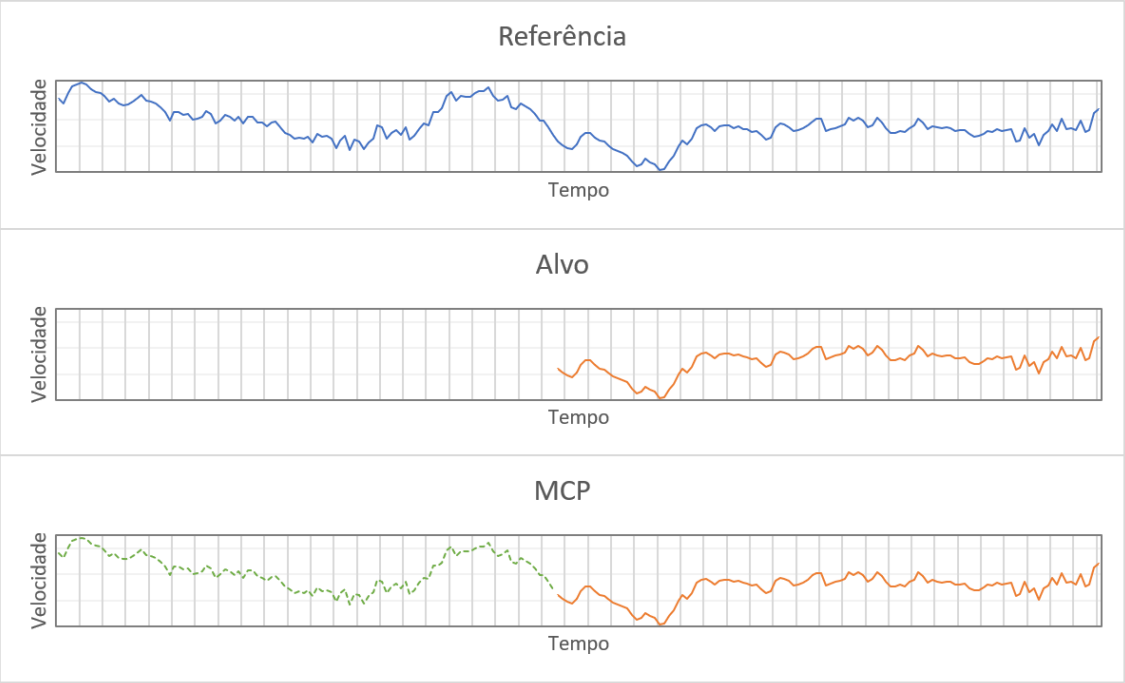


Figura 2.6 – Exemplificação do método de MCP, onde um alvo é reconstruído com uma referência, desde que expostos ao mesmo recurso eólico. Fonte: o autor.

Assim que estabelecida a relação entre referência e alvo, utiliza-se da Equação (2.2), com os devidos parâmetros calculados, para a sintetização de dados do alvo com base na referência, para o período de dados faltantes. Uma vez que os dados foram limpos e reconstruídos, eles serão representativos apenas do período de medição.

Na Figura 2.6, temos o exemplo de uma fonte fictícia, que possui uma cobertura grande em comparação com o alvo. A técnica de MCP, irá gerar uma série temporal sintetizada (em verde pontilhado) de maior cobertura para o a fonte alvo.

Comumente, as campanhas de medição têm apenas alguns anos de dados e, para que as velocidades de vento sejam representativas de um período similar ao da vida útil de um parque eólico, necessita-se extrapolar temporalmente essas velocidades. Assim, as tendências de mudança de longo prazo serão de alguma forma incorporadas à análise. Para isso, novamente pode-se utilizar da técnica de MCP (GALLON, 2015).

2.3.3 Extrapolação espacial do recurso eólico

Assim que se tem o recurso eólico de longo prazo, é necessário extrapolá-lo para a posição das turbinas do projeto. Para extrapolar verticalmente, pode-se utilizar da Equação (2.1) para estimar o perfil vertical do vento na altura de medição das torres e, com base nisso, estimar o recurso eólico na altura do rotor da turbina.

Já para estimar o recurso eólico na posição das turbinas, ou seja, extrapolar horizontalmente, é necessário utilizar algum modelo de escoamento atmosférico. Um dos modelos mais difundidos é o *Wind Atlas Analysis and Application Program* (WAsP) (ASTRUP; LARSEN, 1999), que é baseado em um modelo linear simplificado, e fornece informações com níveis de incerteza aceitáveis para terrenos com topografia simples. Para terrenos de maior complexidade, o WAsP costuma cometer erros visto que a linearização das equações governantes do escoamento deixem escapar dinâmicas que acontecem nessas topografias. Assim, nos casos de topografia complexa, métodos numéricos não lineares em *Computational Fluid Dynamics* (CFD) podem ser aplicados para diminuir a incerteza na predição da velocidade em cada ponto de interesse. Este modelo se divide em três categorias principais (RODRIGUEZ, 2019):

- ***Direct Numerical Simulation* (DNS):** onde não há qualquer tipo de modelagem para simplificar as as equações governantes do escoamento e todas as equações para as escalas temporais e espaciais são resolvidas. Este tipo de simulação demanda extremo poder computacional, e normalmente não é utilizada em escalas como as dos parques eólicos.
- ***Large Eddy Simulations* (LES):** as escalas espaciais e temporais, abaixo de um certo valor e dependendo do problema a ser simulado, são modeladas e as maiores são resolvidas diretamente, sem modelagem. Apesar de ser cada vez mais utilizada, ainda não é uma opção viável para escalas como as dos parques eólicos.
- ***Reynolds Average Navier Stokes* (RANS):** nessa categoria de simulações, todo o escoamento é simplificado usando as suas componentes médias de pressão e velocidade do vento. Esta abordagem é especialmente utilizada para aplicações comerciais com o fim de simular parques eólicos, e demanda poder computacional um pouco menor que as anteriores.

2.3.4 Estimativa da geração de energia

Por fim, assim que o recurso eólico foi extrapolado para a posição e altura do rotor das turbinas eólicas do projeto, realiza-se o cálculo da energia gerada por cada turbina (CUSTÓDIO, 2013). Para isso, utiliza-se da curva de potência disponibilizada pelo fabricante do aerogerador (BURTON et al., 2011).

Como podemos ver na Figura 2.7, para cada velocidade de vento, há um uma

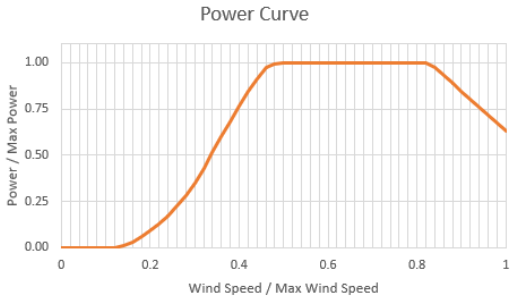


Figura 2.7 – Exemplo de uma curva de potência. Fonte: o autor.

potência de geração. Além disso, é possível verificar que só há geração de energia a partir de uma determinada velocidade de vento, chamada *cut-in*. Com isso, é possível estimar a produção de energia a nível de turbina, e uma vez que somadas as gerações de todas as turbinas, tem-se uma estimativa da produção do parque eólico.

3 TRABALHOS RELACIONADOS

Nesta seção, trabalhos relacionados são avaliados a fim de propor uma metodologia para a detecção de anomalias em séries temporais multivariadas de torres meteorológicas. São considerados trabalhos que buscam detectar anomalias em dados de séries temporais.

Belay et al. (2023) faz uma extensa revisão sobre o estado da arte em detecção de anomalias em séries temporais multivariadas. No trabalho dele, são apresentados treze modelos promissores, e uma avaliação numérica do desempenho de cada modelo em dois conjuntos de dados públicos, detalhando vantagens e deficiências de cada abordagem. Os conjuntos de dados avaliados por Belay et al. (2023) foram o *Server Machine Dataset* (SMD), com 5 semanas de dados de 33 métricas de 28 servidores de uma grande companhia de internet (SU et al., 2019); e o *Secure Water Treatment* (SWaT) com 11 dias de dados de testes de tratamento de água operacional que simula o processo físico e o sistema de controle de uma grande estação de tratamento de água moderna em uma cidade grande (GOH et al., 2017).

Uma coleção de medições de múltiplos sensores em instantes de tempo discretos pode ser definida como uma série temporal multivariada (BELAY et al., 2023). Tais séries temporais, podem ser representadas em uma forma matricial. Mais especificamente, para um sistema com K sensores, Belay et al. (2023) define um estado de sistema em um instante de tempo discreto n como

$$\vec{x}[n] = (x_1[n], x_2[n], \ldots, x_K[n])^\top, \tag{3.1}$$

onde o operador $(\cdot)^\top$ denota uma transposição. Com isso, uma série temporal multivariada com K sensores e N instantes de tempo pode ser representada por uma matriz

$$\mathbf{X} = (\vec{x}[1], \vec{x}[2], \ldots, \vec{x}[N]) = \begin{pmatrix} x_1[1] & x_1[2] & \ldots & x_1[N] \\ x_2[1] & x_2[2] & & x_2[N] \\ \vdots & \vdots & \ddots & \vdots \\ x_K[1] & x_K[2] & \ldots & x_K[N] \end{pmatrix}, \tag{3.2}$$

onde o (n, k)-*ésimo* $x_k[n]$ representa a medida coletada pelo k-*ésimo* sensor no n-*ésimo* instante de tempo.

De acordo com Choi et al. (2021), para séries temporais multivariadas, a relação

entre as medidas através do domínio do tempo e ao longo dos sensores é mais complexa do que a análise individual de cada sensor. Belay et al. (2023) adiciona que métodos de detecção de anomalias para séries temporais multivariadas devem levar em conta todos os sensores simultaneamente.

Segundo Belay et al. (2023), em abordagens não-supervisionadas de detecção de anomalias em séries temporais multivariadas, assume-se a disponibilidade de uma amostra $\mathbf{X_{treino}} \in \mathbb{R}^{K \times N}$ contendo medidas em condições normais grande o suficiente para capturar o comportamento não-anômalo dos sensores. Neste trabalho, no entanto, são utilizados conjuntos de treino que possuem anomalias. Belay et al. (2023) ainda adiciona que, para avaliar a performance dessas abordagens, assume-se a disponibilidade de uma amostra $\mathbf{X_{teste}} \in \mathbb{R}^{K \times M}$, com $M \ll N$, contendo medições tanto normais quanto anormais.

A seguir, serão descritos alguns modelos tradicionais e o estado da arte para detecção de anomalias em séries temporais multivariadas. Foram avaliados cinco modelos, divididos entre técnicas convencionais, baseados em redes neurais e modelos compostos.

3.1 Técnicas convencionais

Modelos como PCA e *Isolation Forest* são alguns exemplos de técnicas convencionais utilizadas na detecção de anomalias em séries temporais multivariadas. A seguir, são detalhados estes modelos e como eles se aplicam à detecção de anomalias.

3.1.1 PCA

De acordo com Belay et al. (2023), a técnica de PCA é usualmente utilizada para reduzir dimensionalidade. Segundo Shyu et al. (2003), PCA preocupa-se em explicar a estrutura de variância-covariância de um conjunto de variáveis através de um novo conjunto de variáveis que são função das originais.

Ainda de acordo com Shyu et al. (2003), as componentes principais são combinações lineares de *K* variáveis randômicas X_1, X_2, ..., X_K com três importantes propriedades:

- As componentes principais não são correlacionadas;
- A primeira componente principal tem a maior variância, a segunda componente

principal tem a segunda maior variância, e assim por diante;

- A variação total em todas as componentes principais combinadas é igual a variação nas variáveis originais.

As componentes principais podem ser obtidas facilmente através da análise dos autovalores e autovetores da matriz de covariância, ou da matriz de correlação das variáveis X_1, X_2, ..., X_K (JOLLIFFE, 1986). Segundo Shyu et al. (2003), a análise componentes principais a partir da matriz de correlação ou a partir da matriz covariância normalmente não são a mesma coisa. Shyu et al. (2003) ainda adiciona que, se algumas das variáveis são muito maiores em magnitude do que as outras, elas receberão altos pesos no cálculo das componentes principais. Por esta razão, se as variáveis são medidas em escalas com grandes diferenças, é melhor performar PCA na matriz de correlação (SHYU et al., 2003).

Seja $\mathbf{C} \in \mathbb{R}^{K \times K}$ uma matriz de correlação computada a partir de N observações de K variáveis randômicas. Se $(\lambda_1, \vec{e_1})$, $(\lambda_2, \vec{e_2})$, ..., $(\lambda_K, \vec{e_K})$ são os K pares de autovalores e autovetores de $\mathbf{C}$, com $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_K \geq 0$, então a i-*ésima* componente principal de um vetor de observações $\vec{x}[n]$ é definida pela Equação (3.3) (SHYU et al., 2003):

$$y_i = \vec{e_i}^{\top}, \vec{z}, \quad i = 1, 2, \ldots, K, \tag{3.3}$$

onde $\vec{e_i}$ é o i-*ésimo* autovetor e $\vec{z}$ o vetor estandardizado de obervações.

Segundo Belay et al. (2023), anomalias em séries temporais multivariadas podem ser identificadas utilizando PCA através do cálculo da distância de um ponto das componentes principais. Belay et al. (2023) adiciona, que um *score* de anomalia pode ser atribuído a esse ponto baseado nessa distância.

3.1.2 *Isolation Forest*

Liu, Ting and Zhou (2008) propõem um método baseado em árvores que objetiva isolar anomalias. Para isso, o método proposto por tais autores se aproveita de duas propriedades quantitativas de anomalias: 1) elas são a minoria, consistindo de apenas algumas instâncias; e 2) elas possuem valores bem diferentes daquelas que são consideradas instâncias normais.

Segundo Liu, Ting and Zhou (2008), em seu método, uma estrutura de árvore é construída, e os pontos anômalos são isolados próximos à raiz, devido a sua susceptibili-

dade a ser isolado, enquanto que pontos normais são isolados em locais distantes da raiz. Essa característica de isolação das árvores é a base do método para detectar anomalias. Cada árvore é chamada de *Isolation Tree* ou *iTree*.

Isolation Forest ou *iForest* se refere a um *ensamble* de *iTrees* para um determinado conjunto de dados. São considerados anomalias aqueles pontos que tiverem um caminho médio pequeno nas *iTrees* (LIU; TING; ZHOU, 2008).

De acordo com Liu, Ting and Zhou (2008), os dados são aleatoriamente particionados até que todas as instâncias sejam isoladas, formando uma árvore randômica. Ainda segundo Liu, Ting and Zhou (2008), esse particionamento produz caminhos notavelmente menores para anomalias de maneira que: 1) quanto menor o número de anomalias, menor o número de particionamentos; e 2) instâncias com valores-atributo distinguíveis são mais suscetíveis a serem separados mais cedo no processo de particionamento.

Dado um conjunto de observações $\mathbf{X} \in \mathbb{R}^{K \times N}$, com K variáveis e N observações, uma *iTree* pode ser construída recursivamente selecionando-se randomicamente uma variável q, das K variáveis, e um valor p, entre o valor máximo e mínimo da variável selecionada (LIU; TING; ZHOU, 2008). Com base em q e p, $\mathbf{X}$ é particionada.

Segundo Liu, Ting and Zhou (2008), uma maneira de detectar anomalias é ordenar os pontos de acordo com a distância até a raiz. A distância até a raiz é definida como o número de vértices pelos quais se passa até chegar ao ponto para o qual se está calculando (LIU; TING; ZHOU, 2008). A partir dessa distância, é possível se estabelecer um *score* de anomalia, já que os pontos anômalos estarão mais próximos à raiz, enquanto pontos normais estarão mais profundos na *iTree*.

3.2 Modelos baseados em redes neurais profundas

Recentemente, técnicas baseadas em *deep learning* têm melhorado a detecção de anomalias em conjuntos de dados multidimensionais (BELAY et al., 2023). Segundo Pang et al. (2021), essas abordagens são capazes de modelar complexas e altamente não-lineares inter-relações entre múltiplos sensores e são capazes de capturar correlações temporais eficientemente.

> Uma abordagem comum para detecção de anomalias em séries temporais com redes neurais profundas é o uso do conceitos de regressão, para prever um ou mais valores futuros baseados em valores passados e, então, a partir do erro de predição, determinar se um ponto é anômalo ou não (BELAY et al., 2023).

Long Short-Term Memory (LSTM) é uma das abordagens baseadas em redes neu-

rais recorrentes, *Recurrent Neural Network* (RNN), que evitam problemas como o de *vanishing gradients* quando modelando dependências temporais de longo prazo (CHUNG et al., 2014). Esses modelos performam bem em tarefas de detecção de anomalias devido às suas capacidades de predição e modelagem de correlação temporal (MALHOTRA et al., 2015).

3.2.1 LSTM

Redes neurais recorrentes são sistemas dinâmicos, que tem um estado interno a cada passo de tempo da classificação (STAUDEMEYER; MORRIS, 2019). Ainda segundo Staudemeyer and Morris (2019), isso se deve às conexões circulares entre neurônios de camadas superiores e inferiores e conexões opcionais de *auto-feedback*. Essas conexões de *feedback* permitem que as RNN propaguem dados de eventos anteriores para etapas de processamento atuais, construindo assim uma memória de eventos de séries temporais.

De acordo com Belay et al. (2023), as RNNs são uma extensão de redes neurais *feed-forward*, com uma memória interna. Usualmente, sua arquitetura é composta de uma camada de entrada, uma camada escondida e uma camada de saída (BELAY et al., 2023). No entanto, diferentemente das redes neurais *feed-forward*, o estado da camada escondida muda ao longo do tempo (BELAY et al., 2023). Belay et al. (2023) ainda adiciona que os neurônios dessa camada escondida não estão conectados apenas aos neurônios das camadas de entrada e saída, mas também com os outros neurônios da própria camada.

Mais especificamente, a camada de entrada com K neurônios recebe uma sequência de vetores $(\ldots, \vec{x}[t-1], \vec{x}[t], \vec{x}[t+1], \ldots)$ e as unidades de entrada são conectadas à camada escondida com M unidades escondidas $\vec{h}[t] = (h_1, h_2, \ldots, h_M)^\top$ através de uma matriz de pesos $\mathbf{W}_{ih}$ (BELAY et al., 2023). De acordo com Belay et al. (2023), a relação recursiva na camada escondida (responsável pelo efeito de memória) é descrita pela Equação (3.4):

$$\vec{h}[t] = f_h(\mathbf{W}_{ih}\vec{x}[t] + \mathbf{W}_{hh}\vec{h}[t-1] + \vec{b}_h), \tag{3.4}$$

onde $f_h(\cdot)$ é a função de ativação da camada escondida, $\mathbf{W}_{hh}$ é a matriz de pesos definindo a conexão entre o estado escondido atual e anterior e $\vec{b}_h$ é o vetor de viés na camada escondida (BELAY et al., 2023).

O estado escondido no instante de tempo t é uma função da entrada atual de dados

e do estado escondido no tempo $t-1$ (BELAY et al., 2023). Com isso, a camada de saída com L unidades $\vec{y}[t] = (y_1, y_2, \ldots, y_L)^\top$ é definida como:

$$\vec{y}[t] = f_o(\mathbf{W}_{ho}\vec{h}[t] + \vec{b}_o), \tag{3.5}$$

onde $f_o(\cdot)$ é a função de ativação da camada de saída, $\mathbf{W}_{ho}$ é a matriz de pesos definindo a conexão entre a camada escondida e a camada de saída e $\vec{b}_o$ é o vetor de viés na camada de saída (BELAY et al., 2023).

Segundo Belay et al. (2023), as RNN são treinadas via *Back Propagation Through Time* (BPTT), mas devido ao decaimento exponencial do gradiente, as RNN tendem a performar mal quando modelando comportamentos com dependência de longo prazo. As redes LSTM, foram introduzidas de forma a evitar este tipo do problema, com a implementação de *gate units*, que controlam se a informação será salva ou sobrescrita a cada instante de tempo (CHUNG et al., 2014).

Segundo Hundman et al. (2018), redes LSTM têm a capacidade de aprender a relação entre os valores de dados passados e atuais e representar essa relação na forma de pesos aprendidos. Hundman et al. (2018) adiciona que, quando treinada com dados normais, redes LSTM podem capturar e modelar o comportamento normal de um sistema.

LSTM também é capaz de lidar com séries temporais multivariadas sem a necessidade de redução de dimensionalidade ou o conhecimento de domínio da aplicação específica (HUNDMAN et al., 2018).

A arquitetura da LSTM consiste de uma célula de estado e três *gates* de controle (entrada, esquecer e saída) (BELAY et al., 2023). Belay et al. (2023) adiciona que a célula de estado é a unidade de memória da rede e carrega informação que pode ser armazenada, atualizada ou lida a partir de uma célula de estado anterior. Os *gates* de entrada e esquecer regulam a atualização/exclusão da memória de longo prazo retida na célula de estado, enquanto o *gate* de saída regula a saída do estado escondido atual (BELAY et al., 2023). De acordo com Belay et al. (2023), as operações internas de uma célula LSTM são descritas pelas equações a seguir:

$$\vec{i}[t] = \sigma(\mathbf{W}_{hi}\vec{h}[t-1] + \mathbf{W}_{xi}\vec{x}[t] + \vec{b}_i), \tag{3.6}$$

$$\vec{f}[t] = \sigma(\mathbf{W}_{hf}\vec{h}[t-1] + \mathbf{W}_{xf}\vec{x}[t] + \vec{b}_f), \tag{3.7}$$

$$\vec{o}[t] = \sigma(\mathbf{W}_{ho}\vec{h}[t-1] + \mathbf{W}_{xo}\vec{x}[t] + \vec{b}_o), \tag{3.8}$$

$$\tilde{C}[t] = \tanh(\mathbf{W}_{hc}\vec{h}[t-1] + \mathbf{W}_{xc}\vec{x}[t] + \vec{b}_c), \tag{3.9}$$

$$\vec{C}[t] = \vec{f}[t] \odot \vec{C}[t-1] + (1 - \vec{f}[t]) \odot \tilde{C}[t], \tag{3.10}$$

$$\vec{h}[t] = \vec{o}[t] \odot \tanh \vec{C}[t], \tag{3.11}$$

onde $\vec{i}[t]$, $\vec{f}[t]$ e $\vec{o}[t]$ representam os *gates* de entrada, esquecer e saída respectivamente, $\tilde{C}[t]$ é a célula de estado candidata, $\vec{C}[t]$ é a célula de estado, $\vec{h}[t]$ é o estado escondido e saída da célula, $\sigma(\cdot)$ é a função sigmoide, $\odot$ é o produto de Hadamard, $\mathbf{W}_\cdot$ é uma matriz de pesos e $\vec{b}_\cdot$ é o vetor de viés em cada *gate*.

Conforme Belay et al. (2023), assume-se que os erros de predição resultantes seguem uma distribuição Gaussiana multivariada. Essa distribuição é então utilizada para determinar a probabilidade de comportamento anômalo.

3.2.2 TranAD

Tuli, Casale and Jennings (2022) propõem uma abordagem baseada em redes neurais *transformers* chamada TranAD, apresentada na Figura 3.1. De acordo com tais autores, modelos *enconder-decoder* tradicionais tendem a não ser capazes de capturar tendências de curto prazo, e não conseguem detectar anomalias se os desvios são muito pequenos em relação ao comportamento normal. Em uma tentativa de evitar isso, o modelo TranAD utiliza um processo de treinamento inspirado em redes adversárias, que pode amplificar erros de reconstrução (TULI; CASALE; JENNINGS, 2022).

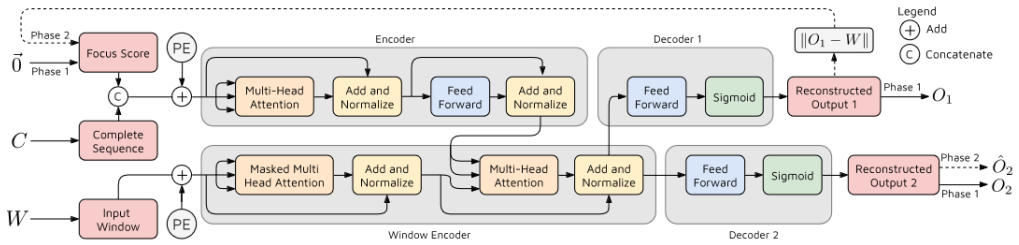


Figura 3.1 – Modelo TranAD. Fonte: (TULI; CASALE; JENNINGS, 2022).

No processo de treinamento adversário, inicialmente é feita uma inferência em duas fases: 1) Reconstrução de entrada; 2) Reconstrução focada de entrada (TULI; CASALE; JENNINGS, 2022). Essas fases estão apresentadas na Figura 3.1 representadas por "Phase 1"e "Phase 2".

Na primeira fase, o modelo *transformer* objetiva reconstruir a série temporal de entrada (designada na Figura 3.1 como W). O erro de reconstrução dessa primeira etapa é

usado como um *score* de foco, que indica pontos onde a segunda fase deveria focar (TULI; CASALE; JENNINGS, 2022). Na segunda fase, o *score* de foco gerado atua como uma prioridade para modificar os pesos da camada de atenção, dando maior ativação da rede neural para sub-sequências de entradas específicas, a fim de extrair tendências temporais de curto prazo (TULI; CASALE; JENNINGS, 2022).

A contribuição das redes adversárias provém do uso da função perda desse tipo de arquitetura, na saída do segundo *decoder*, conforme Figura 3.1 (TULI; CASALE; JENNINGS, 2022). Assim, o segundo *decoder* objetiva distinguir entre os dados originais de entrada (W) e a reconstrução obtida no primeiro *decoder*, maximizando a diferença entre a saída $\hat{O}_2$ e a entrada W, conforme Figura 3.1; enquanto que o primeiro *decoder* objetiva enganar o segundo tentando reconstruir a janela de entrada perfeitamente, assim minimizando a diferença entre a saída O_1 e a entrada W, conforme Figura 3.1 (TULI; CASALE; JENNINGS, 2022).

Ao final do processo de treinamento, estima-se um *score* de anomalia baseado no erro de reconstrução final (TULI; CASALE; JENNINGS, 2022).

3.3 Modelos compostos

Zhang et al. (2019) propõe um sistema para detecção de anomalias em séries temporais multivariadas de forma não-supervisionada, empregando redes LSTM convolucionais junto com redes *enconder-decoder*, chamado MSCRED. De acordo com Belay et al. (2023), o modelo MSCRED consegue caputar padrões temporais efetivamente.

Zhang et al. (2019), introduz o conceito das matrizes multi-escala de assinatura. Essas são utilizadas para caracterizar múltiplos níveis dos estados do sistema através de diferentes instantes de tempo (ZHANG et al., 2019).

Na sequência, tendo essas matrizes de assinatura, um *encoder* convolucional é empregado de maneira a codificar os padrões de correlação inter-sensor e uma rede LSTM convolucional baseada em atenção é desenvolvida para captar padrões temporais (ZHANG et al., 2019). Por fim, um *decoder* convolucional é utilizado para reconstruir a matriz de assinatura, e a matriz residual é utilizada para identificar anomalias (ZHANG et al., 2019). Ainda segundo Zhang et al. (2019), uma vez que o MSCRED seja exposto a dados sem a presença de anomalias, espera-se que o modelo não reconstrua bem a matriz de assinatura quando exposto a um período com anomalias.

Dada uma série temporal multivariada $\mathbf{X_{treino}} \in \mathbb{R}^{K \times M}$, com K variáveis e M

observações não anômalas, para representar as correlações entre diferentes pares de sensores, é construída uma matriz de assinatura $\mathcal{M}[t] \in \mathbb{R}^{M \times M}$, baseada no produto interno, par a par de séries temporais dos sensores. Sejam as séries temporais de dois sensores no segmento de tempo $t - w$ até t, $\vec{x}_i^w = (x_i[t-w], x_i[t-w-1], \ldots, x_i[t])$ e $\vec{x}_j^w = (x_j[t-w], x_j[t-w-1], \ldots, x_j[t])$ a sua correlação é calcula de acordo com a seguinte equação:

$$m_{ij}[t] = \frac{\sum_{\delta=0}^{w} x_i^{t-\delta}[t] x_j^{t-\delta}[t]}{\kappa}, \tag{3.12}$$

onde κ é um fator de re-escala ($\kappa = w$) (ZHANG et al., 2019). Segundo Zhang et al. (2019), a matriz de assinatura $\mathcal{M}[t]$ não só pode capturar as semelhanças de forma e as correlações de escala de valor entre duas séries temporais, mas também é robusta ao ruído de entrada, pois a turbulência em determinadas séries temporais tem pouco impacto nas matrizes de assinatura.

4 METODOLOGIA PROPOSTA

Nesta seção, são detalhadas as etapas da metodologia adotada no presente trabalho. Na Seção 4.1, é descrito o processo de marcação dos dados meteorológicos.

Na sequência, é detalhado como foi feita a preparação dos conjuntos de dados de treinamento e de teste na Seção 4.2. Por fim, na Seção 4.3 são expostos procedimentos de avaliação dos modelos. Na Figura 4.1 são ilustradas as etapas da metodologia.

4.1 Marcação de dados anômalos

Neste trabalho, foram utilizados dados de torres anemométricas públicas, como o *dataset* da *Technical University of Denmark* (DTU) (HANSEN; VASILJEVIC; SØRENSEN, 2021). Não se tem conhecimento da existência de conjuntos de dados marcados, ou seja, com períodos de sensores falhos devidamente identificados.

Dentre os dados de torres disponibilizados pela DTU, foi selecionada uma Torre situada na ilha de Kegnæs na Dinamarca (HANSEN; VASILJEVIC; SøRENSEN, 2021), devido à sua alta cobertura de dados, que mede desde 01/01/1991 até 31/12/2001. Esta torre está localizada em terreno plano em área costeira, possui uma altura de 23 metros e estava equipada com três sensores de velocidade, dois de direção, dois de temperatura, dois de umidade relativa do ar, um de pressão, entre outros.

Entende-se como ponto anômalo, toda medida que, durante a etapa de limpeza de dados é marcada para exclusão. Para a marcação dos períodos anômalos das séries temporais, definiu-se um Protocolo de Marcação de Medidas Anômalas (PMMA) para cada tipo de sensor, conforme Figura 4.2. Este protocolo foi construído com base em recomendações de especialistas da área, no que diz respeito à sensores de velocidade e direção. Decisões referentes aos sensores de temperatura, pressão e umidade relativa não

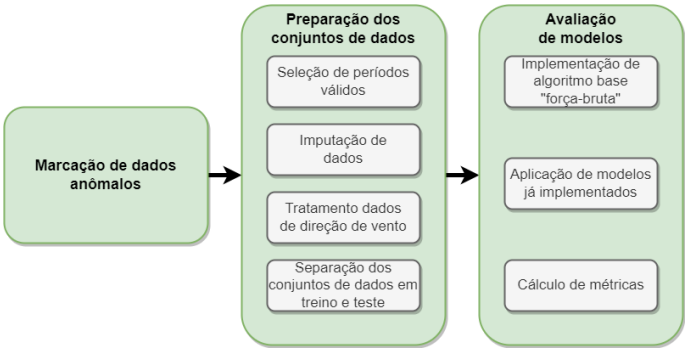


Figura 4.1 – Metodologia proposta. Fonte: o autor.

foram validados.

Os conjuntos de dados selecionados são séries temporais de dez minutos onde, a cada marca temporal, a medida representa uma média dos últimos dez minutos de medidas instantâneas. No entanto, a resolução de coleta das medidas instantâneas é desconhecida.

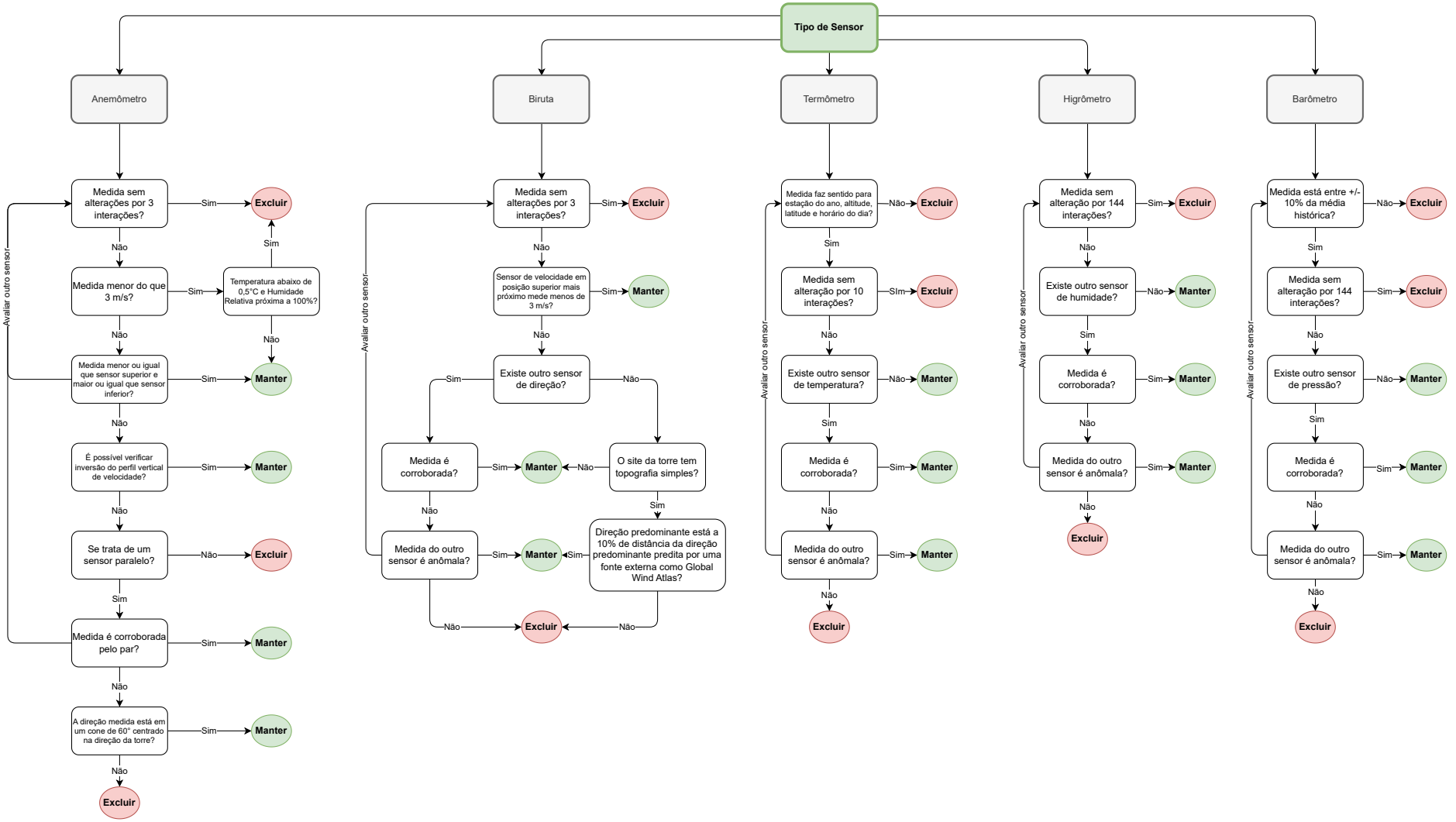


Figura 4.2 – Protocolo de exclusão de medidas anômalas. Fonte: o autor.

Para a marcação dos dados meteorológicos, foi utilizado o *software* proprietário desenvolvido pela empresa DNV chamado *WindFarmer: Analyst*, onde é possível marcar períodos das séries temporais para exclusão. Além de marcar o período para exclusão, é possível atribuir um motivo para a exclusão daquele período, para posterior verificação de terceiros.

Uma vez que os dados estão marcados, é possível exportar as séries temporais contendo somente os dados com medidas representativas da posição das torres anemométricas. Estas séries temporais "limpas" são então utilizadas para as demais etapas de uma certificação do recurso eólico e energético, conforme descrito na Seção 2.3. Para o presente trabalho, são exportadas as séries de classificação para posterior uso.

A seguir, serão descritas as etapas de marcação dos dados meteorológicos, conforme Protocolo definido na Figura 4.2. Cada tipo de sensor tem as suas especificidades quanto aos critérios avaliados na decisão de marcar uma medida para exclusão ou não, e as seções a seguir objetivam detalhar isso.

4.1.1 Anemômetro

Anemômetros são sensores que medem a velocidade do vento (CUSTÓDIO, 2013). Existem anemômetros de copo e anemômetros ultra-sônicos (MORTENSEN, 1994). Anemômetros de copo são os mais comuns no que diz respeito à medida de velocidade de vento (KRISTENSEN, 1999). Usualmente, anemômetros de copo são calibrados individualmente em um túnel de vento, e um certificado de calibração é emitido.

Uma das primeiras verificações a se fazer nas medidas de ventos é se o sensor não está medindo valores constantes. É incomum que um anemômetro meça o mesmo valor por mais do que três interações (30 minutos). Devido à característica variável do vento em uma escala de tempo pequena, esses períodos de medidas são marcados para exclusão. Isso é demarcado no primeiro balão do Protocolo para sensores do tipo Anemômetro na Figura 4.2.

Em regiões onde as temperaturas podem chegar a patamares abaixo de zero, efeitos de congelamento do anemômetro podem ocorrer. O congelamento pode acontecer após chuvas congelantes em altitudes baixas ou em sensores que possam estar expostos à passagem de nuvens congelantes (MAKKONEN; LEHTONEN; HELLE, 2001). Assim, é razoável pensar que, quando há a presença de sensores de temperatura e/ou juntamente com os anemômetros, pode-se verificar a hipótese de congelamento. No Protocolo defi-

nido na Figura 4.2, medidas de velocidade inferiores a $3ms^{-1}$, conjuntamente com medidas de temperatura abaixo de $0,5°C$ e umidade relativa próxima à 100% são marcados para exclusão.

Conforme descrito na Seção 2.2.1 e 2.2.2, o vento possui um perfil onde a velocidade do vento aumenta com o crescimento da altitude. Assim, quando um sensor inferior mede uma velocidade maior que um sensor superior, por mais de três interações (30 minutos), as medidas deste sensor são marcadas para exclusão. No entanto, efeitos térmicos, podem gerar uma inversão completa do perfil de vento, fazendo com que a velocidade diminua com o aumento da altitude e, neste caso, as medidas dos sensores não são marcadas para exclusão.

Quando há a montagem de anemômetros em paralelo, é possível corroborar as medidas entre os sensores à mesma altura. No entanto, quando o vento incide de uma direção paralela à direção torre-anemômetro, a torre estará influenciando o escoamento do vento diretamente a jusante (LOTFI, 2015), conforme pode ser verificado na Figura 4.3, e neste caso o anemômetro que fica à sombra não pode ser usado para corroborar a medida do sensor paralelo.

A fim de remover a influência da torre nas medições de sensores paralelos, na etapa de reconstrução de dados, conforme Figura 2.4, pode-se fazer uma média seletiva entre os anemômetros baseando-se na direção de incidência do vento. Quando o vento incide a partir de uma direção onde nenhum dos sensores é afetado pela torre, toma-se a média das velocidades medidas. Quando um dos sensores está afetado, toma-se a medida do sensor não-afetado. Como o comportamento de anemômetro afetado pela torre é removido em etapa posterior de uma análise do recurso élico, no PMMA definiu-se que, quando o vento incide de uma direção compreendida por um cone de $60°$ centrado na direção da anemômetro-torre, o registro não necessita ser marcado para exclusão.

4.1.2 Biruta

Birutas são os instrumentos utilizados para a medição da direção do vento (CUSTÓDIO, 2013). Existem alguns tipos de birutas, mas, para medidas de meteorológicas, os mais comuns são compostos de uma haste com um aerofólio na parte traseira, e uma ponteira que aponta para a direção do vento. Além disso, ela é montada sobre uma base que possibilita o giro facilmente com as mudanças de direção do vento.

Assim como para os anemômetros, é incomum que uma biruta meça uma mesma

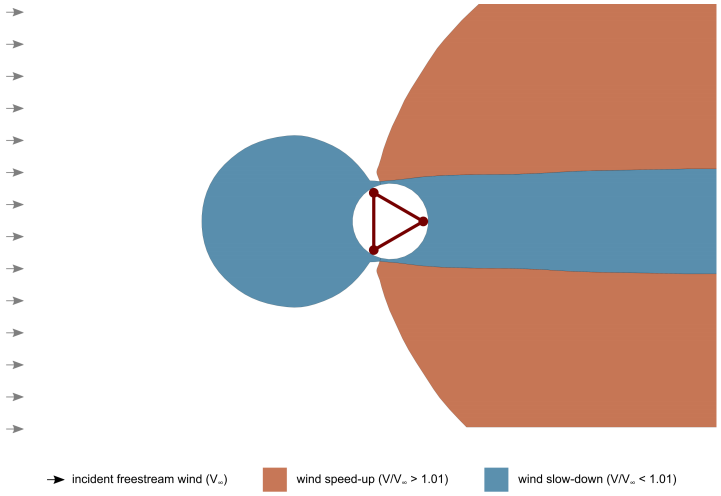


Figura 4.3 – Influência de uma torre treliçada no escoamento do vento. Fonte: (LOTFI, 2015).

direção do vento por muito mais do que três interações (30 minutos). Quando isso acontece, esse grupo de dados deve ser marcado como anômalo, conforme o primeiro balão do PMMA para sensores do tipo Biruta na Figura 4.2.

Quando expostas a velocidades baixas de vento, os anemômetros vão registrar velocidades que não são de interesse de um estudo eólico. Como podemos verificar na Seção 2.3.4, os aerogeradores só começam a produzir energia a partir de uma determinada velocidade de vento. Assim, no Protocolo, definiu-se que, quando a velocidade de vento é menor do que $3ms^{-1}$, não há necessidade de marcar qualquer comportamento da biruta como anômala.

Se existem outras birutas ao longo da torre, pode-se corroborar a direção medida. Vale ressaltar que, existem efeitos de meso-escala, como as brisas que podem resultar em uma torção do perfil de direção de vento em determinados horários do dia. O Protocolo não contempla este tipo de efeito.

Por fim, quando o *site* possui baixa complexidade topográfica, pode-se corroborar as direções de vento para um determinada região através de fontes públicas, como o *Global Wind Atlas* (BADGER et al., 2015).

4.1.3 Termômetro

Termômetros são os sensores responsáveis pela medição da temperatura (KATSAROS, 2001), e é comum que torres anemométricas tenham pelo menos um sensor de temperatura. Através dos sensores de temperatura, é possível corroborar as data e hora dos registros da série temporal, pois em um ciclo normal de temperatura, a temperatura

começa a aumentar com o nascer do sol e a diminuir com o poente; e registra temperaturas maiores em estações quentes do que em estações frias.

No primeiro balão do Protocolo para marcação de Termômetros da Figura 4.2, confirma-se se a medida faz sentido para a estação do ano, altitude e latitude, e horário do dia. Em latitudes distantes da linha do equador, há uma distinção maior entre estações frias e quentes no que diz respeito à temperatura e com isso é possível corroborar a medida. Por exemplo, é incomum que no inverno do Sul da Argentina durante a madrugada uma temperatura de 30° seja medida, assim como uma medida de 0° seja medida ao meio dia no Nordeste do Brasil.

Assim, como com os sensores previamente descritos, é incomum que a temperatura fique constante por muito tempo. Para o Protocolo definido neste trabalho, definiu-se que quando um sensor fica sem alteração por dez interações (1h e 40min), este período deve ser marcado para exclusão.

Se existem outros sensores, é possível corroborar a medida de temperatura. Quando instalados em diferentes alturas, é possível identificar efeitos de estratificação térmica, e deve-se ficar atento para que não sejam marcados para exclusão dados de fenômenos reais.

4.1.4 Higrômetro

Higrômetro é o sensor que mede a umidade relativa do ar (KATSAROS, 2001). Existem alguns tipos de higrômetros, mas a maioria deles funciona através da medição de vapor de água em um determinado volume de ar o qual é comparado com uma temperatura ou pressão conhecida.

É razoável pensar que medidas de umidade relativa não fiquem constantes por muito tempo, e a fim de marcar medidas constantes para exclusão, definiu-se no Protocolo de Higrômetros na Figura 4.2 que medidas constantes por 144 ou mais interações (24h) sejam marcadas para exclusão. Se existem outros sensores de umidade ao longo da torre, é possível corroborar as medidas.

4.1.5 Barômetro

Barômetros medem a pressão atmosférica local (KATSAROS, 2001). Há dois tipos principais: os barômetros de mercúrio e os barômetros aneroides. Barômetros de

mercúrio se utilizam de uma coluna de mercúrio para realizar medidas de pressão, enquanto os aneroides são compostos por um diafragma de metal sensível a variação da pressão.

As medidas de pressão atmosférica tendem a não ter grandes variações ao longo do tempo, e é razoável assumir que qualquer medida muito distante da média histórica seja anômala. No Protocolo de marcação de dados de Barômetros para exclusão na Figura 4.2, definiu-se que medidas que se distanciem 10% da média histórica, devem ser marcados para exclusão.

Assim como para os Higrômetros, é incomum que a pressão atmosférica se mantenha constante por um período maior do que 144 interações (24h). Nestes casos, esses períodos de medições constantes são marcados para exclusão.

Por fim, se existem sensores de pressão em diferentes alturas da torre é possível corroborar as medidas. Vale ressaltar que vão existir pequenas diferenças nos valores medidos a diferentes alturas dependendo da diferença de altura. Por exemplo, um barômetro medindo próximo ao solo irá medir uma pressão atmosférica levemente maior que um barômetro medindo à 150 metros de altura.

4.2 Preparação dos conjuntos de dados

Uma vez que os dados estão marcados, é necessário fazer o devido tratamento para utilizá-los como entrada para os modelos selecionados. Um primeiro problema que surge é a descontinuidade de dados medidos, seja de sensores individuais ou de toda a torre. A maioria dos modelos estatísticos ou de aprendizado não é robusta o suficiente para lidar com *gaps* de dados (JADHAV; PRAMOD; RAMANATHAN, 2019), e é comum o emprego de técnicas de imputação.

Como descrito na Seção 2.3.1, erros na velocidade do vento são elevados ao cubo no cálculo de energia, e com isso deve-se ter cautela ao utilizar técnicas de imputação de dados em séries temporais de velocidade de vento que serão utilizadas para estudos de certificação.

Período	Data de Início	Data de Fim	Tamanho
P_0	07/03/1991 09:45	11/06/1991 09:35	96 dias
P_1	10/10/1991 13:45	02/02/1992 23:55	115 dias
P_2	02/04/1992 13:05	14/10/1992 20:35	195 dias
P_3	17/10/1992 10:25	20/01/1993 15:25	95 dias
P_4	23/07/1997 08:45	25/11/1997 13:55	125 dias
P_5	24/06/1998 12:45	02/11/1998 11:05	131 dias
P_6	02/11/1998 12:05	25/10/1999 09:35	357 dias
P_7	08/12/1999 10:55	29/03/2000 10:55	112 dias
P_8	04/05/2000 09:15	28/12/2000 23:55	238 dias
P_9	07/02/2001 11:55	28/06/2001 11:05	141 dias
P_{10}	28/06/2001 12:45	07/12/2001 23:25	162 dias

Tabela 4.1 – Lista de períodos de dados válidos.

4.2.1 Imputação e seleção de períodos de dados válidos

A primeira etapa é identificar as lacunas na série temporal. Se um ou mais sensores estiverem ausentes em um determinado instante de tempo, todos os registros daquele instante serão removidos.

Uma vez identificados os *gaps*, são geradas informações sobre o tamanho dos períodos de dados válidos e dos períodos de *gap*. De modo a evitar a inserção de vieses nos dados, optou-se pela imputação de dados faltantes de apenas um ponto através da média entre a medida anterior ao *gap* e a medida posterior ao *gap*. Vale ressaltar que, se apenas um dos sensores possuía ausência de medida, apenas esse será imputado, e para os demais serão mantidos os valores medidos.

Com isso, tem-se mais períodos de dados válidos e mais extensos. Por fim, seleciona-se os períodos com uma cobertura mínima de três meses de dados válidos e obtém-se os períodos elencados na Tabela 4.1, que totalizam quatro anos e dez meses de dados. Foram selecionados períodos de no mínimo três meses de dados, de modo a ter uma boa quantidade de conjuntos de dados ao mesmo tempo em que se tenta ter uma boa cobertura temporal.

Durante a etapa de marcação dos dados da torre em análise, percebeu-se que as medidas de umidade relativa apresentavam valores anômalos em toda a sua extensão medindo 100% por períodos superiores a um dia. Desta forma, os dados dos sensores de umidade não foram considerados neste trabalho.

4.2.2 Tratamento dos dados de direção

Uma vez que tem-se os períodos de dados válidos, é necessário fazer os devidos tratamentos para utilizá-los como entrada dos modelos. No que diz respeito aos dados de direção medidos pelas Birutas, haverão saltos nas direções medidas, já que as medidas variam de $0°$ a $360°$.

Em situações onde o vento varia em torno dessas direções, olhando apenas para o ângulo parecerá que há grande variação, quando na verdade pode ser que a variação da direção seja mínima. Isso se deve à natureza descontínua, quando tomamos a direção por um ângulo e a falsa percepção de distância grande entre ângulo $0°$ e $360°$.

De modo a evitar este problema, e eventuais dificuldades ao utilizar os modelos, é aplicada uma função seno nas medições de direção. Este procedimento faz com que a percepção de grande distância entre $0°$ e $360°$ desapareça.

4.2.3 Separação dos conjuntos de dados em treino e teste

Após a devida seleção, imputação e tratamento dos dados, cada período de dados válidos é separado ao meio. A primeira parte do período é definida como conjunto de treinamento e a segunda como conjunto de teste.

Afim de estimar um índice de anomalia para cada conjunto, assumiu-se que, se ao menos um sensor possui medida anômalo, todas as medidas do instante de tempo são considerados anômalos. Desta forma, é possível estimar um valor de anomalias independente do número de sensores considerados e isso possibilita comparar os resultados com torres que possuam configurações diferentes.

Na Tabela 4.2 é possível notar que existe um desbalanço entre as classes "Anômalo" e "Não-anômalo". O índice de anomalia é definido conforme:

$$I_A = \frac{N_A}{N} \times 100\%, \tag{4.1}$$

onde N_A é o número de pontos anômalos e N é o número de pontos do conjunto de dados.

Período	Conjunto de Treino [%]	Conjunto de Teste [%]
P_0	2,53	1,06
P_1	0,17	0,26
P_2	0,82	0,00
P_3	0,00	0,31
P_4	0,22	0,54
P_5	1,29	0,32
P_6	1,21	0,72
P_7	2,00	4,90
P_8	1,82	1,61
P_9	0,86	0,94
P_{10}	1,36	0,39

Tabela 4.2 – Índice de Anomalia dos conjuntos de treino e teste.

4.3 Avaliação de modelos

A fim de avaliar modelos promissores para a detecção de anomalias em séries temporais multivariadas, foi selecionada a implementação MTAD *toolkit* (LIU et al., 2022), pois contempla desde modelos básicos, como PCA e iForest, a modelos de aprendizado profundo como LSTM e TranAD e modelos compostos como MSCRED.

O *toolikt* MTAD é implementado de forma a facilitar o *benchmarking* de novos conjuntos de dados. Uma vez que os conjuntos de teste e treino estejam preparados, basta selecionar o modelo desejado para detecção de anomalias. Ao rodar o algoritmo, os conjuntos de dados são normalizados e são calculados *scores* de anomalia associado a cada instante de tempo da série temporal.

Na sequência, é selecionado um *threshold* de anomalia, que definirá se um ponto é anômalo ou não baseado em seu *score*. O *toolkit* permite selecionar o *threshold* de forma a aumentar ou diminuir uma determinada métrica, como o *F1-Score* por exemplo. Assim, um ponto cujo *score* de anomalia é igual ou superior o *threshold* é considerado anômalo, e o que for abaixo é considerado não anômalo. No presente trabalho, o *threshold* de anomalia é selecionado de modo a maximizar o *F1-Score*.

Após isso, com base no *threshold* de anomalia, é feita a predição e o cálculo das métricas do modelo executado. Na Figura 4.4, podemos ver de forma simplificada o fluxo de operação do *toolkit* MTAD.

De modo a ter um caso base para comparação, foi implementado um algoritmo de força bruta, baseando-se no Protocolo definido na Figura 4.2. No entanto, algumas premissas do Protocolo, como verificação cíclica de sensores à mesma altura, ou verificação em fontes fora da série temporal não puderam ser implementadas.

Este algoritmo foi implementado em Python, e utiliza de bibliotecas como NumPy

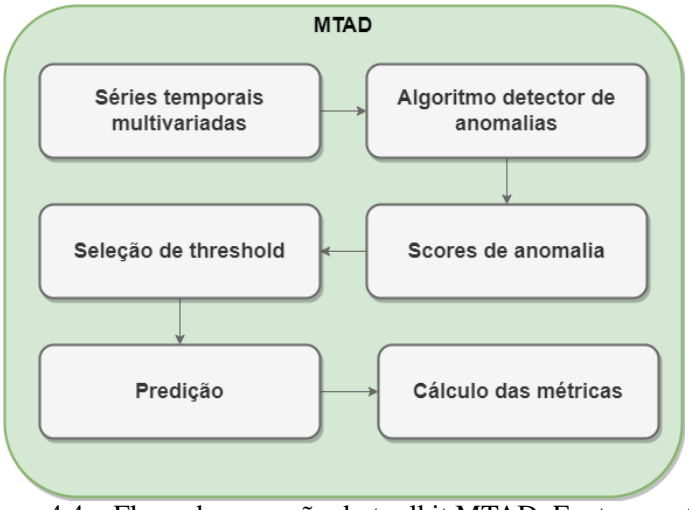


Figura 4.4 – Fluxo de operação do toolkit MTAD. Fonte: o autor.

e Pandas. Todos os códigos foram implementados e rodados na plataforma Google Colab, em sua versão Pro que dispõe de melhores recursos como CPUs mais eficientes, mais memória RAM variando de 12GB até 83GB e disponibilidade de GPUs de até 16GB.

Para possibilitar a comparação entre as técnicas, são avaliadas as métricas Precisão, Revocação e *F1-score*. A Precisão, definida na Equação (4.2), avalia o quanto um modelo está marcando dados como anômalo e eles de fato são anômalos. Já a Revocação, definida na Equação (4.3), avalia o quanto o modelo está marcando como não anômalo quando de fato é anômalo. Por fim, o *F1-score* é uma média harmônica entre a Precisão e a Revocação e é definida na Equação (4.4). Nas equações

$$\textit{Precisão} = \frac{TP}{TP + FP}, \tag{4.2}$$

$$\textit{Revocação} = \frac{TP}{TP + FN} \tag{4.3}$$

e

$$F_1 = \frac{2 \times \textit{Precisão} \times \textit{Revocação}}{\textit{Precisão} + \textit{Revocação}}, \tag{4.4}$$

TP são os verdadeiros positivos, FN são os falsos negativos e FP são os falsos positivos. Os valores resultante das Equações (4.2), (4.3) e (4.4) estão no intervalo $[0, 1]$.

De acordo com Su et al. (2019), na prática, observações de anomalias geralmente ocorrem continuamente formando segmentos contíguos. Com isso, Su et al. (2019) propõe uma etapa de pós-processamento dos rótulos preditos pelos modelos onde os rótulos entre pontos anômalos que inicialmente foram marcados como não-anômalos são ajustados para anômalos. Após esse ajuste, as métricas Precisão, Revocação e *F1-Score* são re-calculadas. Neste trabalho, ambas métricas são apresentadas, sendo que as que empre-

gam o pós-processamento serão identificadas pela palavra "Ajustado".

4.3.1 Configuração dos Modelos

Foram selecionados alguns modelos dentro das principais categorias descritas na Seção 3, para que se pudesse avaliar o desempenho na tarefa de identificar anomalias nas séries temporais de dados meteorológicos. A seguir são listados os modelos selecionados dentro de cada categoria:

- Técnicas Convencionas
  - PCA
  - *iForest*
- Técnicas baseadas em redes neurais profundas
  - LSTM
  - TranAD
- Modelos compostos
  - MSCRED

O *toolkit* MTAD possibilita diferentes configurações de parâmetros e hiper-parâmetros dos modelos avaliados, como taxa de aprendizado, número de camadas de redes neurais, entre outros. A seguir são especificadas as configurações de cada modelo.

4.3.1.1 PCA

Para este modelo, o *toolkit* não requer configurações de parâmetros ou hiper-parâmetros. Conforme comentado na Seção 3, os *scores* de anomalia são calculados com base na distância do ponto em relação ao eixo das componentes principais. Em específico para o MTAD, a função distância utilizada é a distância euclidiana.

4.3.1.2 Isolation Forest

Para este modelo é possível especificar o número de *iTrees* do *ensamble* a ser executado. Neste trabalho, foram utilizadas 10.000 *iTrees*.

Foram avaliados números pequenos de *iTrees* da ordem de 20 a 30 e percebeu-se que o aumento para valores maiores beneficiou o modelo. No entanto, uma análise completa da influência do número de *iTrees* não foi feita.

4.3.1.3 LSTM

Para o modelo LSTM, é possível configurar a taxa de aprendizado, o número de épocas de treinamento, o número de camadas LSTM em sequência e o número de células em cada camada. Neste trabalho, foram variados os hiper-parâmetros relacionados ao número de camadas e células, e percebeu-se que não houve benefício em aumentar o tamanho da rede. Em razão disso, foram configuradas três camadas com 64 células em cada.

Além disso, foram configuradas 1000 épocas de treinamento, com uma taxa de aprendizado de 0.001 e um mecanismo de parada antecipada, onde é verificado se a função perda não varia por cinco épocas de treinamento. Verificou-se que para todos os períodos válidos, o algoritmo parou antes das 1000 épocas, com essas configurações.

Por fim, é possível configurar o número de instantes de tempo de cada *batch* de treinamento e o tamanho da "janela deslizante" utilizada. Para o tamanho do *batch* foram utilizados 1008 instantes de tempo, que representam sete dias de medições e para a janela de tempo foi definido 144 instantes de tempo, que representa um dia de medições.

4.3.1.4 TranAD

Para o modelo TranAD, é possível configurar a taxa de aprendizado e o número de épocas de treinamento. Foi configurada uma taxa de aprendizado de 0.001 e 100 épocas de treinamento.

Foram configuradas tamanho de janela e *batch* com as mesmas configurações de treinamento da LSTM. O *toolkit* não implementa um mecanismo de parada antecipada, então os resultados apresentados são com base nos testes com 100 épocas.

4.3.1.5 MSCRED

Para este modelo, é possível configurar a taxa de aprendizado e número de épocas de treinamento. Investigações realizadas mostram que variar o número de épocas foi benéfico para o modelo.

Com uma taxa de aprendizado de 0.0001, o modelo apresentou melhora quando o

número de épocas foi de 10 para 30, mas não apresentou melhora quando foi de 30 para 200. O *toolkit* não implementa um mecanismo de parada antecipada, então os resultados apresentados são com base nos testes com 200 épocas.

Foram configuradas tamanho de janela e *batch* com as mesmas configurações de treinamento da LSTM.

5 RESULTADOS E DISCUSSÃO

Essa seção apresenta os resultados dos algoritmos PCA, *iForest*, LSTM, TranAD e MSCRED junto com o algoritmo força-bruta implementado, descritos na Seção 3, de acordo com as métricas Precisão, Revocação e *F1-Score*, revisitadas na Seção 4.1. As Tabelas 5.1, 5.2, 5.3, 5.4, 5.5 e 5.6 apresentam os resultados por período para os algoritmos listados acima. Em todas as tabelas, os melhores resultados por período são destacados em negrito. Os resultados são apresentados e discutidos em detalhes a seguir.

Inicialmente, foi avaliado o algoritmo de força bruta que implementa o Protocolo de marcação de dados para exclusão definido na Figura 4.2. Foram selecionados os conjuntos de dados de teste de cada período, conforme Tabela 4.1 e foram avaliadas as métricas definidas na Seção 4.3.

Na Tabela 5.1, pode-se verificar o desempenho do algoritmo para cada período avaliado. No entanto, para o algoritmo de força bruta, não foi realizado o pós-processamento mencionado na Seção 4.3.

É possível notar uma certa sensibilidade das métricas do algoritmo ao índice de anomalia, visto que os períodos P_1 e P_7 com o pior e melhor *F1-Score*, são os períodos com o menor e maior índice de anomalia. No caso do período P_2, o conjunto de testes não apresenta anomalias no período e devido a isso, não é possível computar as métricas selecionadas.

Além disso, é possível inferir que o algoritmo força-bruta esteja marcando como pontos anômalos medidas que passaram despercebidas na etapa de limpeza dos dados descrita na Seção 2.3.1. A baixa Precisão e baixa Revocação podem indicar a existência de muitos falsos positivos e muitos falso negativos, ou seja, o algoritmo prevê muitos pontos não-anômalos como anômalos e muitos pontos anômalos como não-anômalos.

Período	Precisão	Revocação	*F1-Score*	Índice de Anomalia [%]
P_0	0,280	0,192	0,228	1,056
P_1	0,011	0,091	0,020	0,265
P_2	-	-	-	0,000
P_3	0,000	0,000	0,000	0,306
P_4	0,100	0,490	0,166	0,543
P_5	0,127	0,633	0,211	0,318
P_6	0,091	0,297	0,139	0,720
P_7	0,378	0,633	**0,473**	4,898
P_8	0,329	0,290	0,308	1,606
P_9	0,196	0,337	0,248	0,936
P_{10}	0,259	0,630	0,367	0,393

Tabela 5.1 – Métricas do algoritmo de força bruta. O Índice de Anomalia do conjunto de teste já foi apresentado na Tabela 4.2 e está colocado nessa coluna para simplificação da análise.

Período	Precisão	Revocação	*F1-Score*	Prec. Ajustada	Rev. Ajustada	*F1-Score* Ajustada
P_0	0,018	0,945	0,035	0,029	1,000	0,057
P_1	0,027	0,409	0,050	0,067	0,545	0,120
P_2	-	-	-	-	-	-
P_3	0,006	0,857	0,012	0,009	1,000	0,019
P_4	0,006	0,878	0,013	0,009	0,898	0,017
P_5	0,011	1,000	0,021	0,014	0,667	0,027
P_6	0,009	0,957	0,017	0,012	0,708	0,024
P_7	0,203	0,628	0,307	0,449	0,965	**0,613**
P_8	0,029	0,471	0,054	0,037	0,431	0,069
P_9	0,010	0,979	0,020	0,013	1,000	0,025
P_{10}	0,008	1,000	0,016	0,009	1,000	0,019

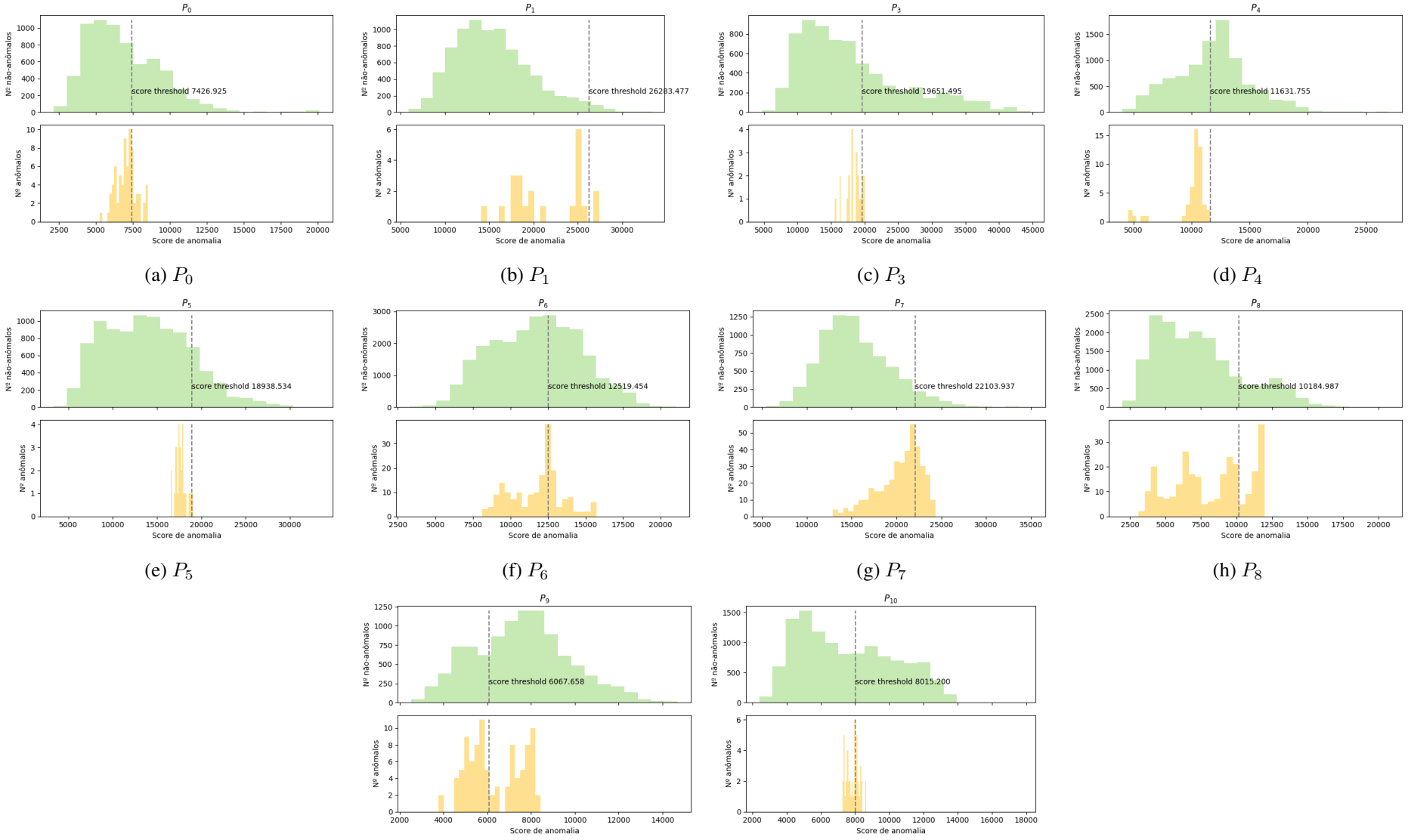
Tabela 5.2 – Métricas do algoritmo PCA.

Já considerando um algoritmo convencional como o PCA, percebe-se desempenho bastante pobre no que diz respeito às métricas avaliadas, conforme Tabela 5.2. Mesmo realizando o pós-processamento descrito por Su et al. (2019), o método não apresenta resultados satisfatórios para a maioria dos períodos de dados válidos. Os valores baixos de Precisão e altos de Revocação, indicam que existe um número grande de falsos positivos em relação ao número de verdadeiros positivos e um baixo número de falsos negativos.

A partir das de saída do *toolkit* MTAD, é possível construir um histograma dos *scores* de anomalia para pontos Não-anômalos e pontos Anômalos. Com essas informações, e o *threshold* de anomalia é possível visualizar como o algoritmo está performando na hora de atribuir os *scores* a cada ponto.

Na Figura 5.1 são apresentadas as distribuições de *score* de anomalia para o modelo PCA. Para a construção dessas distribuições, são utilizados os scores de anomalia propostos pela abordagem em questão, e o *threshold* apresentado é o ajustado pelo pós-processamento. Para separar a distribuição dos pontos anômalos dos não anômalos, foram utilizados os rótulos de teste. Idealmente haveriam três conjuntos de dados, sendo um para treinamento, um para validação e um para teste. No presente trabalho, dispõe-se apenas de conjuntos de treinamento e teste, e por esta razão são utilizados os rótulos de teste para seleção do *threshold*, assim como feito por Su et al. (2019).

É possível verificar que, na maioria dos períodos o pico das distribuições é sobreposta, e o *threshold* selecionado acaba ou classificando muitos dados normais como anômalos ou muitos dados anômalos como normais. Na Figura 5.1g, é possível notar que o algoritmo de PCA apresenta um bom desempenho em identificar anomalias para período P_7, em relação aos outros períodos, visto que o *threshold* quase que isola os picos das distribuições. Conforme descrito na Seção 4.3, o *threshold* de anomalia é selecionado de maneira a maximizar o *F1-Score* para o período avaliado.



(a) P_0 (b) P_1 (c) P_3 (d) P_4

(e) P_5 (f) P_6 (g) P_7 (h) P_8

(i) P_9 (j) P_{10}

Figura 5.1 – Distribuição de *score* de anomalias do algoritmo PCA. Fonte: o autor.

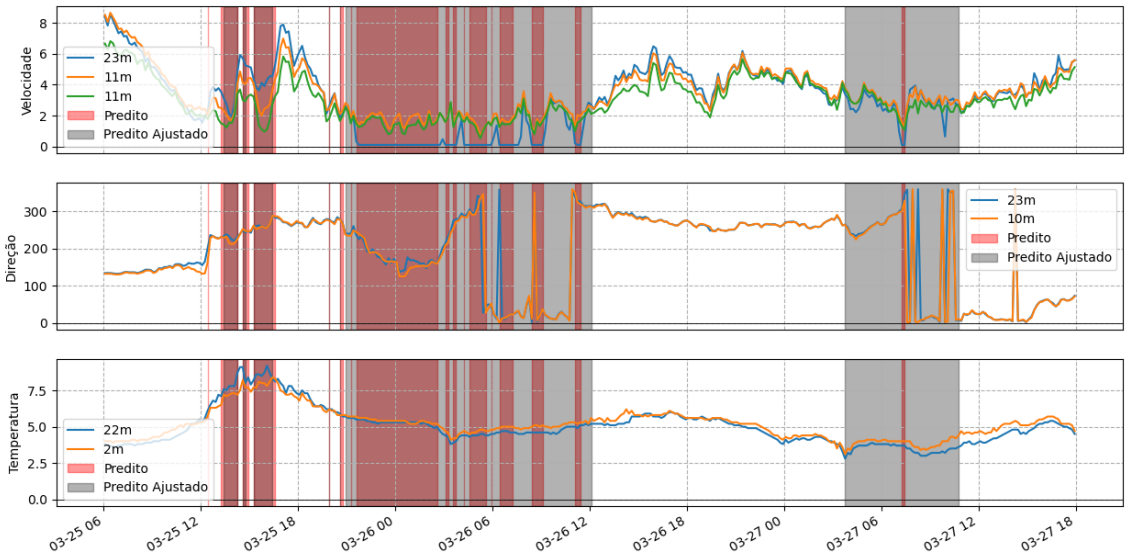


Figura 5.2 – Visualização do pós-processamento para métricas ajustadas, para um sub-período do período P_7 para o modelo *iForest*. Fonte: o autor.

.

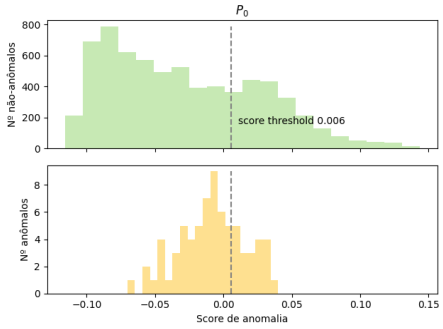
O desempenho do método *iForest* também não é tão bom, embora melhor que o do PCA, no que diz respeito à métrica *F1-Score* Ajustada. Na Tabela 5.3, é possível identificar mais uma vez o modelo acerta mais para o período P_7 do que para os demais.

Na Figura 5.2 onde é apresentado um recorte do período P_7, podemos ver o efeito que o pós-processamento tem sobre os pontos anômalos preditos pelo modelo *iForest*. É possível notar períodos preditos como anômalos, destaque em vermelho, têm uma certa dispersão e quando ajustados, passam a integrar um grupo contínuo de anomalias, destacado em cinza. As áreas sem destaque indicam períodos de dados normais, ou seja, sem a presença de anomalias preditas pelo modelo ou rotuladas.

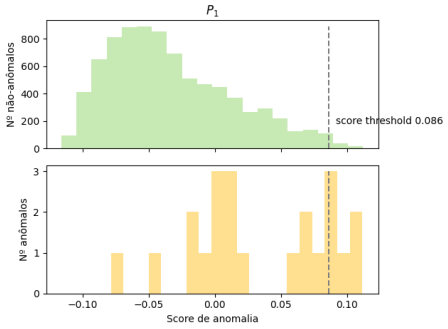
Na Figura 5.3, são apresentadas as distribuições de *score* de anomalia. Neste caso, é possível notar na Figura 5.3g que o *threshold* de anomalia isola a maior parte dos dados não anômalos na categoria correta, mais do que é observado para a técnica PCA e isso pode explicar uma métrica *F1-Score* ajustada maior.

Período	**Precisão**	**Revocação**	***F1-Score***	**Prec. Ajustada**	**Rev. Ajustada**	***F1-Score* Ajustada**
P_0	0,019	0,849	0,037	0,035	1,000	0,068
P_1	0,071	0,273	0,113	0,133	0,545	0,214
P_2	-	-	-	-	-	-
P_3	0,006	0,810	0,012	0,011	1,000	0,021
P_4	0,009	0,857	0,018	0,036	0,898	0,069
P_5	0,015	0,933	0,030	0,034	0,667	0,065
P_6	0,119	0,335	0,176	0,366	0,708	0,483
P_7	0,342	0,494	0,404	0,639	0,965	**0,769**
P_8	0,032	0,322	0,058	0,076	0,511	0,132
P_9	0,036	0,116	0,055	0,248	0,347	0,289
P_{10}	0,019	0,913	0,037	0,047	1,000	0,089

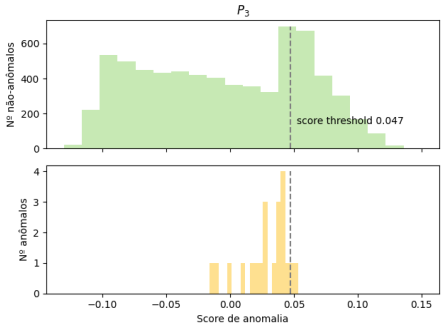
Tabela 5.3 – Métricas do algoritmo *iForest*.



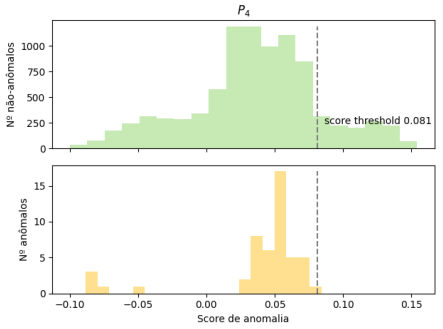
(a) P_0



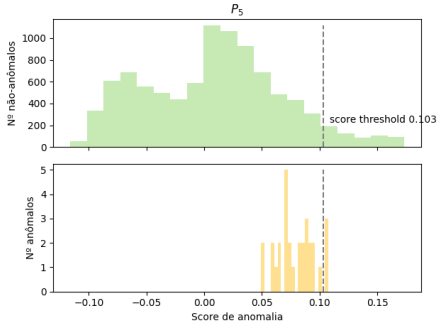
(b) P_1



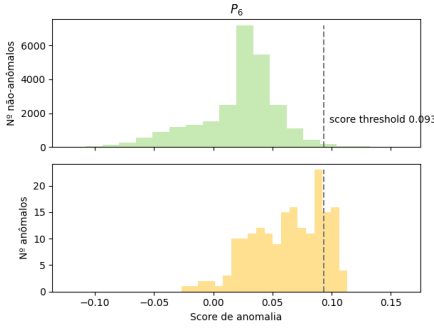
(c) P_3



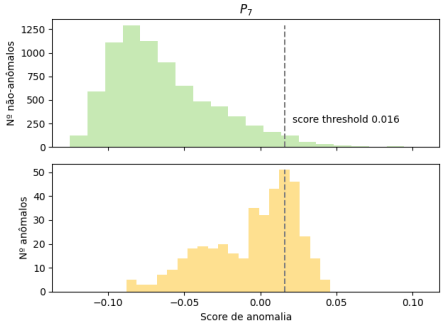
(d) P_4



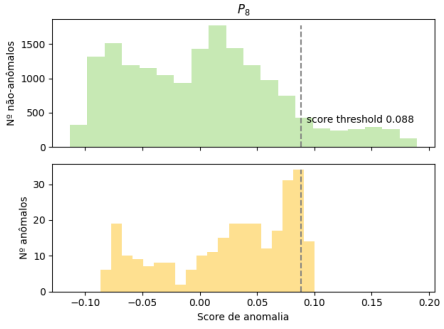
(e) P_5



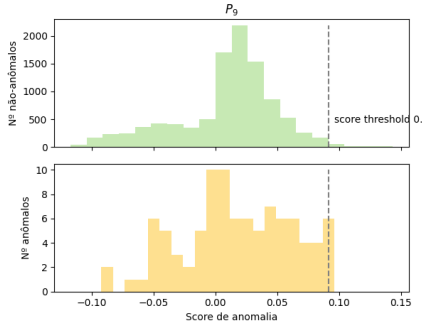
(f) P_6



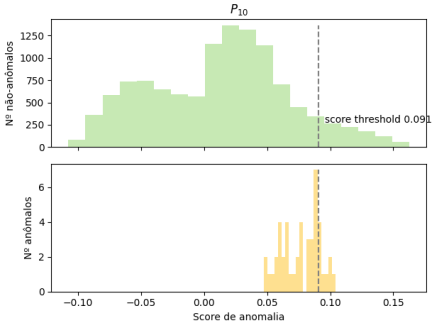
(g) P_7



(h) P_8



(i) P_9



(j) P_{10}

Figura 5.3 – Distribuição de *score* de anomalias do algoritmo *iForest*. Fonte: o autor.

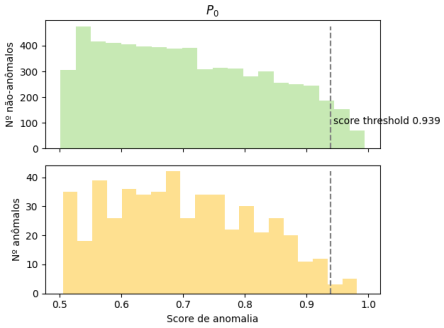
Período	Precisão	Revocação	*F1-Score*	Prec. Ajustada	Rev. Ajustada	*F1-Score* Ajustada
P_0	0,078	0,896	0,143	0,657	1,000	0,793
P_1	0,044	0,536	0,082	0,798	1,000	0,888
P_2	-	-	-	-	-	-
P_3	0,036	0,321	0,065	1,000	1,000	**1,000**
P_4	0,061	0,164	0,089	0,801	1,000	0,890
P_5	0,041	0,307	0,073	0,622	0,484	0,544
P_6	0,061	0,993	0,114	0,561	1,000	0,718
P_7	0,216	0,985	0,354	0,894	1,000	0,944
P_8	0,083	0,915	0,151	0,645	0,860	0,737
P_9	0,067	1,000	0,125	0,701	1,000	0,824
P_{10}	0,017	1,000	0,034	0,458	1,000	0,628

Tabela 5.4 – Métricas do algoritmo LSTM.

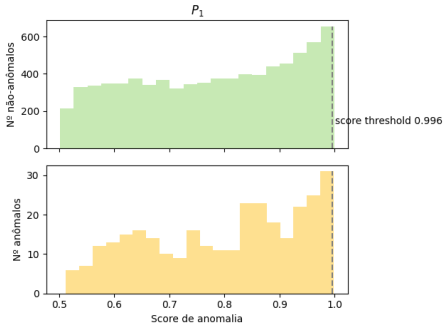
Já no caso do modelo LSTM, é possível notar um incremento significativo no desempenho em identificar anomalias ao comparar com as técnicas *iForest* e PCA. Em um primeiro momento, ao observar o *F1-Score* baixo sem ajuste levaria à uma conclusão de que o modelo não está performando bem, mas uma vez que é feito o pós-processamento proposto por Su et al. (2019), as métricas passam a indicar que o modelo está performando bem na tarefa de identificar anomalias, dado um *F1-Score* ajustado próximo a um.

É possível que o LSTM não esteja captando bem períodos anômalos contínuos e isso pode explicar as métricas pobres. Uma vez que o pós-processamento é aplicado, anomalias contíguas são consideradas contínuas dentro de um grupo de pontos anômalos e com isso, as métricas passam a exibir valores altos.

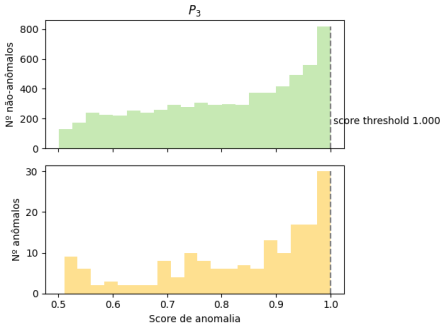
Na Figura 5.4, podemos observar que o modelo não consegue atribuir *scores* de anomalia de forma a separar eficientemente o que é anômalo do que não é anômalo e isso se reflete nas métricas não ajustas da Tabela 5.4. Apesar disso, supomos que o modelo acerta com eficiência alguns pontos anômalos em segmentos que de fato são anômalos, e o pós-processamento resolve essa inabilidade de identificar grupos contínuos de anomalia.



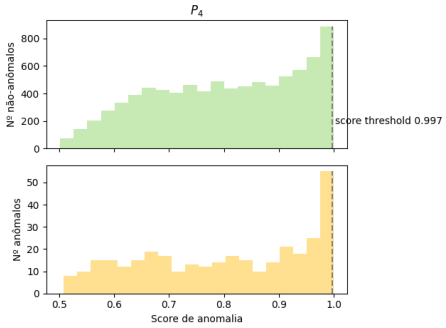
(a) P_0



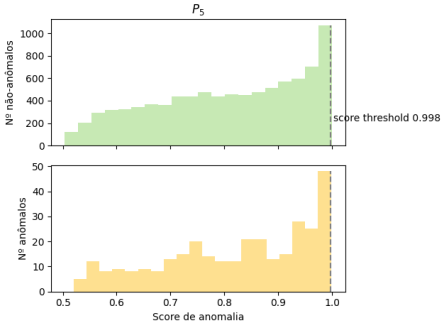
(b) P_1



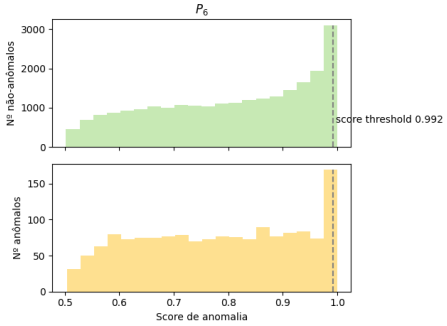
(c) P_3



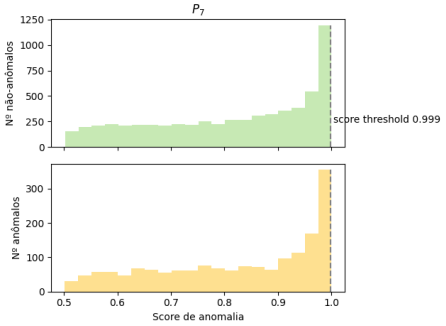
(d) P_4



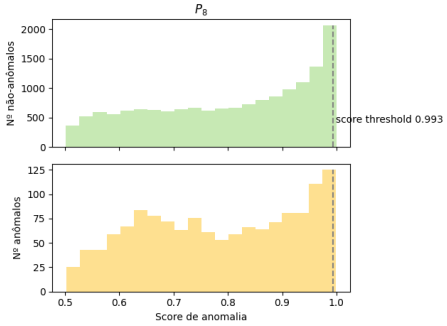
(e) P_5



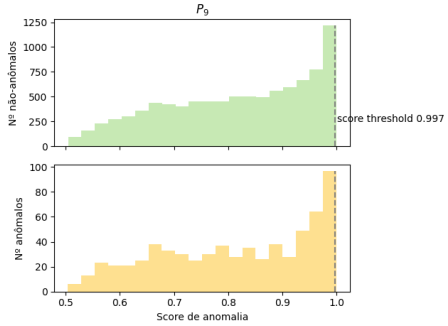
(f) P_6



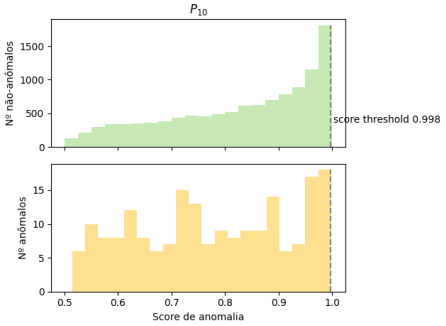
(g) P_7



(h) P_8



(i) P_9



(j) P_{10}

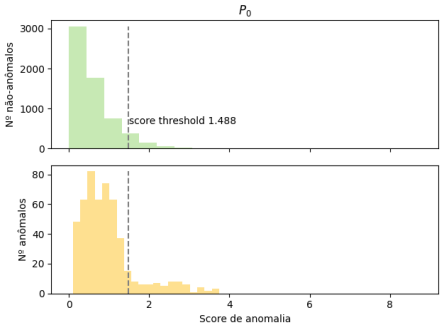
Figura 5.4 – Distribuição de *score* de anomalias do algoritmo LSTM. Fonte: o autor.

Período	Precisão	Revocação	*F1-Score*	Prec. Ajustada	Rev. Ajustada	*F1-Score* Ajustada
P_0	0,139	0,578	0,224	0,515	1,000	0,680
P_1	0,078	0,312	0,124	0,660	0,503	0,571
P_2	-	-	-	-	-	-
P_3	0,025	1,000	0,049	0,079	1,000	0,146
P_4	0,038	1,000	0,073	0,044	1,000	0,084
P_5	0,053	0,937	0,100	0,125	1,000	0,222
P_6	0,075	0,851	0,138	0,266	0,523	0,353
P_7	0,385	0,453	0,416	0,966	1,000	**0,983**
P_8	0,113	0,309	0,165	0,199	0,444	0,275
P_9	0,067	1,000	0,125	0,142	1,000	0,249
P_{10}	0,020	0,990	0,039	0,050	1,000	0,095

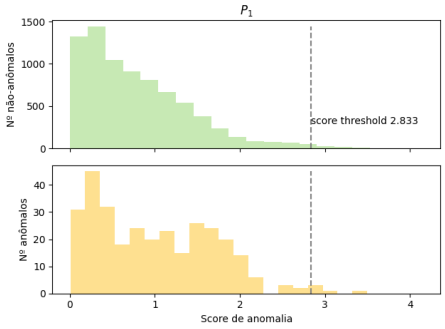
Tabela 5.5 – Métricas do algoritmo TranAD.

Já no modelo TranAD, é possível notar que não há um benefício nas métricas com o pós-processamento como para o LSTM, visto que os valores de *F1-Score* continuam baixos para a maioria dos períodos. Na Tabela 5.5, é possível identificar que apenas o período P_7 parece se beneficiar do pós-processamento, enquanto que os demais apresentam métricas ajustadas baixas.

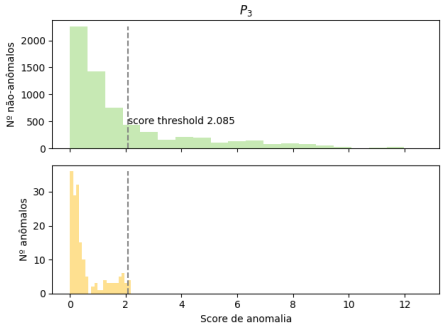
Na Figura 5.5, é possível verificar que, assim como para o LSTM, o modelo TranAD não consegue separar o que é anômalo do que não é anômalo de forma eficiente. Isso se reflete nas métricas não ajustas, e o fato de as métricas ajustadas não serem muito altas pode indicar que o modelo não acerta tanto os pontos anômalos em segmentos contínuos de anomalia.



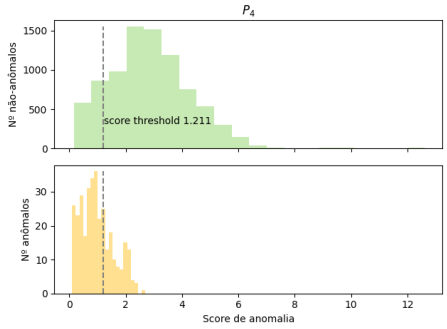
(a) P_0



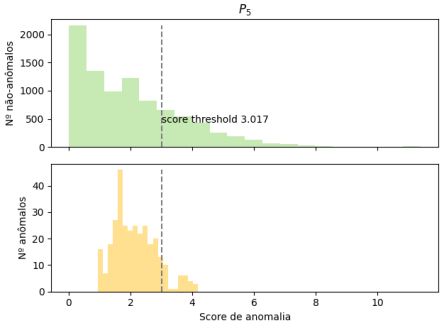
(b) P_1



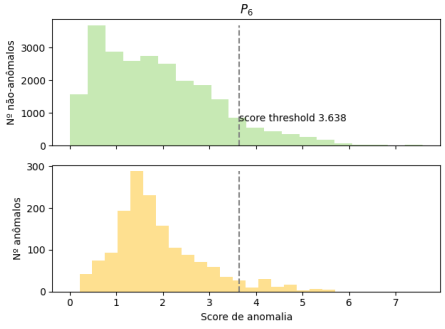
(c) P_3



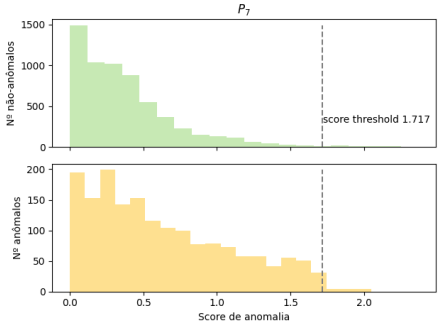
(d) P_4



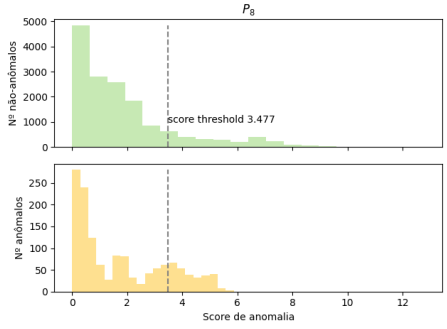
(e) P_5



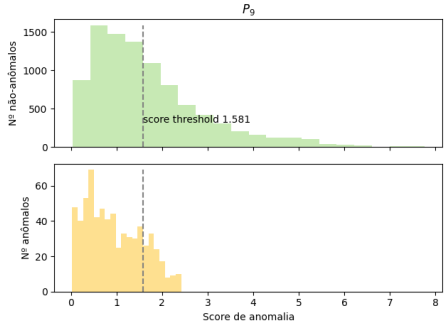
(f) P_6



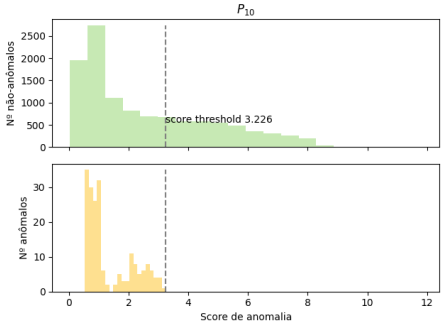
(g) P_7



(h) P_8



(i) P_9



(j) P_{10}

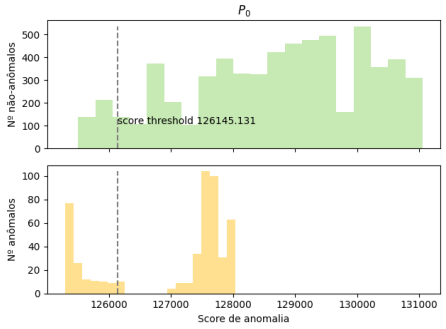
Figura 5.5 – Distribuição de *score* de anomalias do algoritmo TranAD. Fonte: o autor.

Período	Precisão	Revocação	*F1-Score*	Prec. Ajustada	Rev. Ajustada	*F1-Score* Ajustada
P_0	0,075	1,000	0,140	0,080	1,000	0,148
P_1	0,406	0,435	0,420	0,742	0,503	**0,600**
P_2	-	-	-	-	-	-
P_3	0,042	0,988	0,081	0,058	1,000	0,110
P_4	0,180	0,528	0,268	0,208	0,558	0,303
P_5	0,326	0,484	0,390	0,466	0,484	0,475
P_6	0,067	1,000	0,126	0,070	1,000	0,130
P_7	0,258	0,885	0,399	0,361	0,899	0,515
P_8	0,215	0,428	0,286	0,350	0,519	0,418
P_9	0,075	1,000	0,140	0,086	1,000	0,159
P_{10}	0,116	0,964	0,208	0,184	1,000	0,311

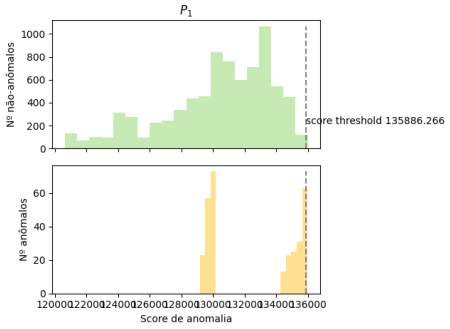
Tabela 5.6 – Métricas do algoritmo MSCRED.

Para o modelo MSCRED, o pós-processamento também parece não beneficiar tanto o modelo como foi no caso do LSTM. Na Tabela 5.6, podemos verificar que, para alguns períodos como P_1 e P_7, o modelo consegue ter um desempenho relativamente bom.

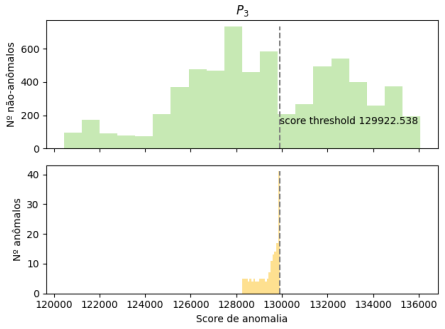
Observando a Figura 5.6, é possível verificar que, em alguns casos, o modelo consegue ter algum nível de separação entre as distribuições de pontos anômalos e não anômalos, a exemplo da Figura 5.6b. O fato de o período P_1, possuir a maior *F1-Score*, pode ser um indício da capacidade de o MSCRED de separar as distribuições para este período em específico.



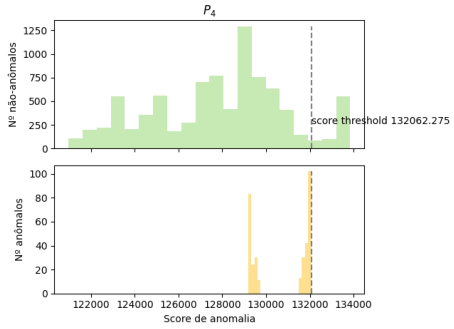
(a) P_0



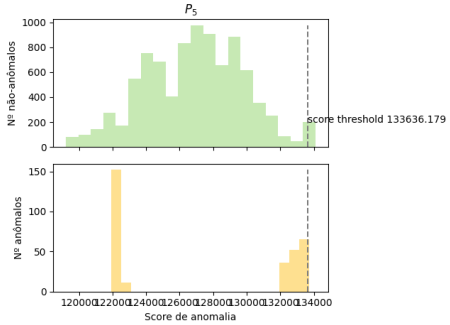
(b) P_1



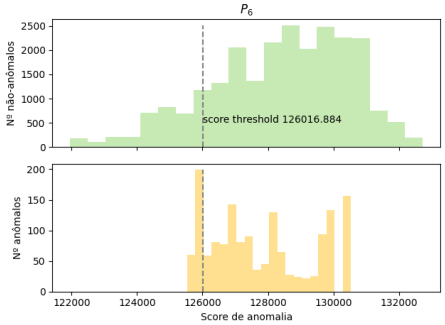
(c) P_3



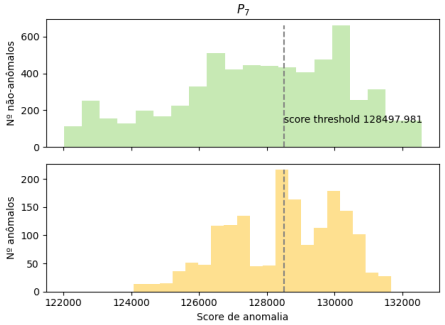
(d) P_4



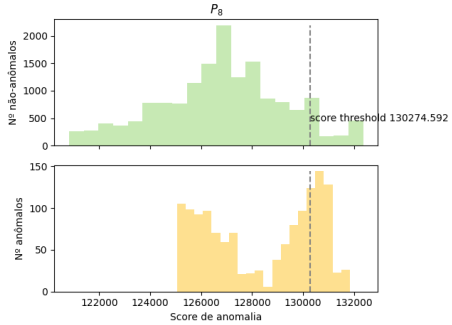
(e) P_5



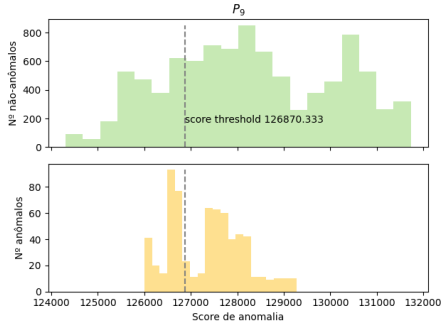
(f) P_6



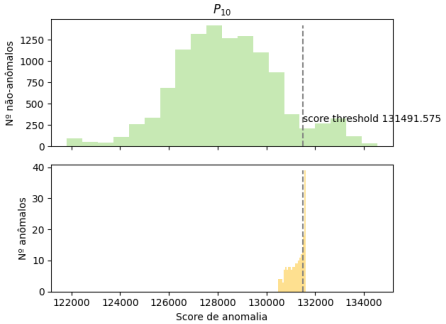
(g) P_7



(h) P_8



(i) P_9



(j) P_{10}

Figura 5.6 – Distribuição de *score* de anomalias do algoritmo MSCRED. Fonte: o autor.

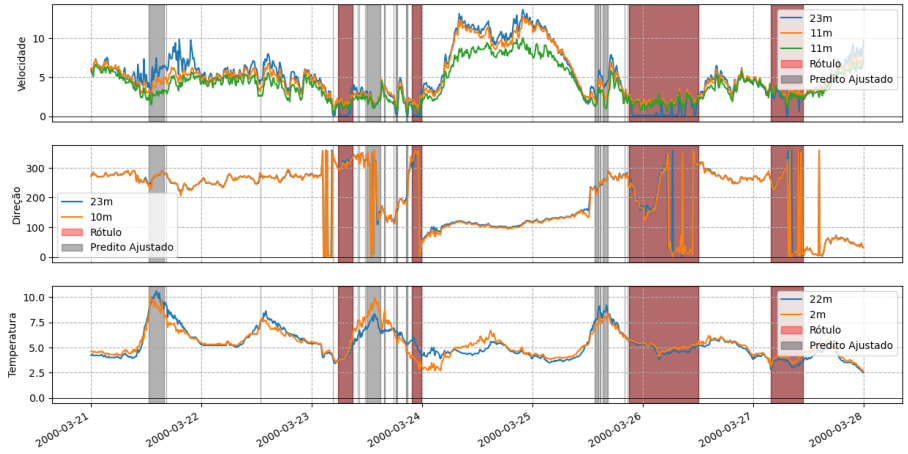


Figura 5.7 – Exemplos de Falsos Positivos no período P_7 para o modelo *iForest*. Fonte: o autor.

Os modelos avaliados, junto com o algoritmo de força-bruta exibem diferentes capacidades de identificar anomalias. Enquanto o algoritmo de força-bruta carece de capacidade de prever anomalias, modelos convencionais, como PCA e *iForest*, parecem conseguir identificar grupos contínuos de pontos anômalos, mas pecam ao classificar muitos dados considerados normais como anômalos. Na Figura 5.7, podemos ver exemplos onde o modelo *iForest* erroneamente marca os períodos, destacados em cinza, como anômalos. A sobreposição entre o destaque de períodos rotulados como anômalo e destaque de períodos previstos como anômalos possui cor marrom escura.

Já o modelo baseado em redes neurais LSTM apresenta bom potencial para identificar anomalias, desde que o devido pós-processamento seja aplicado. Isso se deve ao fato do modelo não conseguir capturar bem anomalias em sequência, fazendo com que as suas métricas sejam ruins, mas ao aplicar o pós-processamento é possível identificar visível melhora nas métricas.

Já modelos como TranAD e MSCRED, aparentam um desempenho superior às técnicas convencionais, mas inferior ao LSTM. Possivelmente, isso se deve ao fato desses modelos não conseguirem separar bem o que é anômalo do que não é, e mesmo com o devido pós-processamento, os resultados não melhoram tanto.

Em resumo, entende-se que o modelo que melhor performou na tarefa de identificar pontos anômalos e se apresenta como um modelo promissor para a análise de dados meteorológicos foi o LSTM. Na Figura 5.8, são sumarizados os resultados de cada modelo para cada período. Para cada modelo é calculada a média de cada métrica ajustada, excluindo-se o período P_2, e os gráficos são ordenados em ordem decrescente de média da métrica *F1-Score* ajustada, sendo a maior Precisão ajustada o critério de desempate.

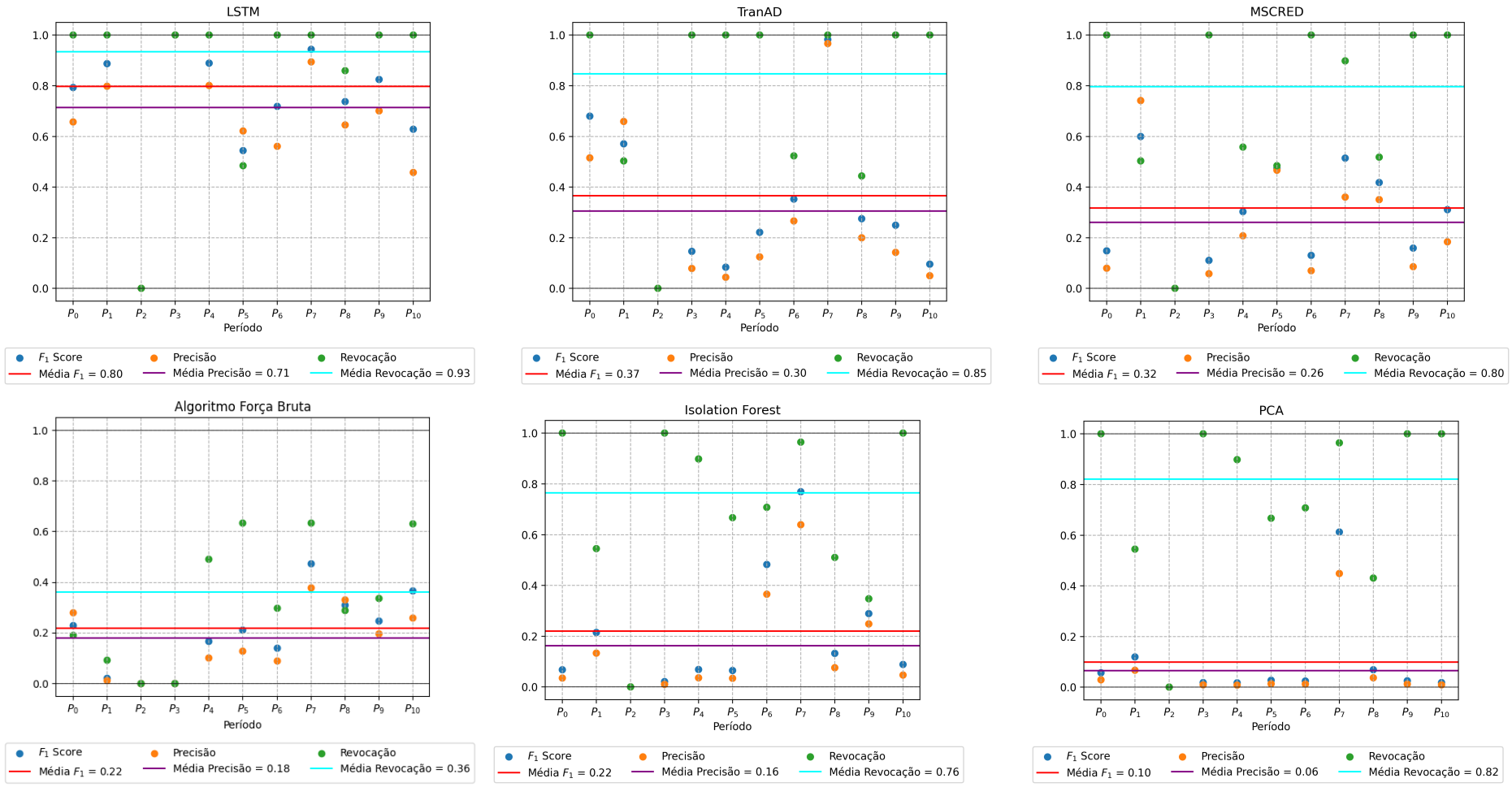


Figura 5.8 – Resumo das métricas ajustadas dos modelos e algoritmos avaliados. Fonte: o autor.

6 CONCLUSÕES

Este trabalho endereçou o problema de detecção de medidas anômalas em séries temporais multivariadas de torres meteorológicas, que são utilizadas para estudos eólicos pré-construtivos. Para isso, foram utilizados dados de uma torre situada na Dinamarca, que possui três sensores de velocidade, dois de direção, dois de temperatura, dois de umidade relativa e um de pressão com medições ao longo de dez anos. Foi proposto um Protocolo para marcação de medidas anômalas, que posteriormente foi implementado na forma de um algoritmo "força-bruta".

Foram avaliadas técnicas convencionais para detecção de anomalia, como PCA e *iForest* juntamente com um algoritmo força-bruta. Também foram avaliados modelos modernos baseados em redes neurais, como LSTM, TranAD e MSCRED.

Para comparar os modelos, foram utilizadas as métricas Precisão, Revocação e *F1-Score*. Além de apresentar as métricas com as saídas brutas dos modelos, foi implementada uma abordagem de pós-processamento, e com isso, também são apresentadas métricas "ajustadas". A abordagem que apresentou as melhores métricas ajustadas foi o LSTM, com *F1-Score* Ajustado médio de $0,8$.

Como o processo de rotular as séries temporais foi feita pelo próprio autor, é possível que hajam vieses. Ao propor um protocolo de marcação de medidas anômalas, espera-se que as decisões tomadas possibilitem que os rótulos possam ser verificados no futuro. Espera-se que em trabalhos futuros, a marcação dos dados seja corroborado por terceiro, de forma a mitigar eventuais erros ao rotular.

Outra limitação da análise feita é a inexistência de um conjunto de validação para o cálculo de *threshold* de anomalia, visto que utilizar o conjunto de teste pode introduzir um viés nas métricas. Salienta-se, entretanto, que todos os modelos foram ajustados a fim de maximizar seus *F1-scores* sob mesmo protocolo. Espera-se que, em trabalhos futuros, os modelos sejam avaliados com conjuntos de treino, validação e teste, e que o *threshold* possa ser calculado com base no conjunto de validação.

É possível que a escolha de parâmetros diferentes para os modelos influencie nos resultados. Espera-se que em trabalhos futuros sejam avaliadas de maneira mais profunda a variação dos parâmetros como número de camadas, tamanho de rede, taxa de aprendizado, etc.

Além disso, espera-se que o PMMA seja validado, refinado e melhorado, possibilitando assim a criação de conjuntos de dados públicos de medidas anômalas em torres

meteorológicas. Isso possibilitará que mais trabalhos relacionados possam ser realizados.

Como neste trabalho objetivou-se indicar períodos de atenção, onde pode haver medidas anômalas, não foram avaliadas anomalias na granularidade dos sensores. Espera-se que, em trabalhos futuros, seja possível avaliar os modelos considerando falhas de sensores individualmente.

Também espera-se que mais torres, em diferentes exposições de recurso eólico, sejam avaliadas. Além disso, espera-se que outros modelos promissores possam ser avaliados. Por fim, almeja-se que em trabalhos futuros sejam propostas *pipelines* de detecção de anomalias empregando diversos modelos, que consigam distinguir os vários tipos de anomalia, fazendo distinção entre falha de sensor, congelamento, degradação e etc.

REFERÊNCIAS

3DOTENERGY. 2023. Available from Internet: <https://www.linkedin.com/pulse/advantages-triangular-vs-squared-cross-sectioned-met-mast/>.

ABEEóLICA. 2021. Available from Internet: <https://abeeolica.org.br/wp-content/uploads/2022/07/424_ABEEOLICA_RELATORIO-ANUAL-2021_V3.pdf>.

ALEXIADIS, A. Global warming and human activity: A model for studying the potential instability of the carbon dioxide/temperature feedback mechanism. **Ecological Modelling**, Elsevier BV, v. 203, n. 3-4, p. 243–256, may 2007. Available from Internet: <https://doi.org/10.1016/j.ecolmodel.2006.11.020>.

ASTRUP, P.; LARSEN, S. **WAsP engineering flow model for wind over land and sea**. [S.l.: s.n.], 1999. (Denmark. Forskningscenter Risoe. Risoe-R, 1107(EN)). ISBN 87-550-2529-3.

BADGER, J. et al. **The Global Wind Atlas: An EUDP project carried out by DTU Wind Energy**. [S.l.: s.n.], 2015.

BARRIATTO, L. C. **Efeitos da estabilidade atmosférica na modelagem do escoamento para aplicações no setor de energia eólica**. Dissertation (Master), 2018. Available from Internet: <http://hdl.handle.net/10183/179414>.

BELAY, M. A. et al. Unsupervised anomaly detection for IoT-based multivariate time series: Existing solutions, performance analysis and future directions. **Sensors**, MDPI AG, v. 23, n. 5, p. 2844, mar. 2023. Available from Internet: <https://doi.org/10.3390/s23052844>.

BURTON, T. et al. **Wind Energy Handbook**. Wiley, 2011. Available from Internet: <https://doi.org/10.1002/9781119992714>.

CARTA, J. A.; VELÁZQUEZ, S.; CABRERA, P. A review of measure-correlate-predict (MCP) methods used to estimate long-term wind characteristics at a target site. **Renewable and Sustainable Energy Reviews**, Elsevier BV, v. 27, p. 362–400, nov. 2013. Available from Internet: <https://doi.org/10.1016/j.rser.2013.07.004>.

CHOI, K. et al. Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. **IEEE Access**, Institute of Electrical and Electronics Engineers (IEEE), v. 9, p. 120043–120065, 2021. Available from Internet: <https://doi.org/10.1109/access.2021.3107975>.

CHUNG, J. et al. **Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling**. arXiv, 2014. Available from Internet: <https://arxiv.org/abs/1412.3555>.

CUSTÓDIO, R. dos S. **Energia eólica para produção de energia elétrica**. [S.l.]: Synergia Editora, 2013.

DNV. **Energy Transition Outlook 2022**. 2022. Available from Internet: <https://www.dnv.com/energy-transition-outlook>.

ELDRIDGE, F. R. **Wind machines / Frank R. Eldridge**. 2d ed. ed. New York: Van Nostrand Reinhold Co, 1980. (MITRE energy resources and environment series). ISBN 0442261349.

EPE. 2021. Available from Internet: <https://www.epe.gov.br/pt/leiloes-de-energia/leiloes/instrucoes-para-cadastramento>.

FREIRE, L. S. **Teorias de Camada Limite Atmosférica: modelo de crescimento, fluxo de entranhamento e análise espectral**. Dissertation (Master), 2012. Available from Internet: <https://acervodigital.ufpr.br/bitstream/handle/1884/27998/R%20-%20D%20-%20LIVIA%20SOUZA%20FREIRE.pdf>.

GALLON, G. P. Análise comparativa entre o potencial eólico previsto e a energia produzida pelo complexo eólico rio do fogo. In: . [s.n.], 2015. Available from Internet: <http://hdl.handle.net/10183/127949>.

GOH, J. et al. A dataset to support research in the design of secure water treatment systems. In: ____. **Lecture Notes in Computer Science**. Springer International Publishing, 2017. p. 88–99. ISBN 9783319713687. Available from Internet: <http://dx.doi.org/10.1007/978-3-319-71368-7_8>.

HANSEN, K. S.; VASILJEVIC, N.; SØRENSEN, S. A. **Resource data**. Technical University of Denmark, 2021. Available from Internet: <https://data.dtu.dk/collections/Resource_data/5405286/4>.

HANSEN, K. S.; VASILJEVIC, N.; SøRENSEN, S. A. **Resource data from the Kegnes mast**. Technical University of Denmark, 2021. Available from Internet: <https://data.dtu.dk/articles/dataset/Resource_data_from_the_Kegnes_mast/14135618>.

HOEGH-GULDBERG, O. et al. **Impacts of 1.5°C Global Warming on Natural and Human Systems**. [S.l.]: World Meteorological Organization Technical Document, 2018.

HUNDMAN, K. et al. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In: **Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining**. ACM, 2018. (KDD '18). Available from Internet: <http://dx.doi.org/10.1145/3219819.3219845>.

JADHAV, A.; PRAMOD, D.; RAMANATHAN, K. Comparison of performance of data imputation methods for numeric dataset. **Applied Artificial Intelligence**, Informa UK Limited, v. 33, n. 10, p. 913–933, jul. 2019. ISSN 1087-6545. Available from Internet: <http://dx.doi.org/10.1080/08839514.2019.1637138>.

JOLLIFFE, I. T. **Principal Component Analysis**. Springer New York, 1986. ISSN 0172-7397. ISBN 9781475719048. Available from Internet: <http://dx.doi.org/10.1007/978-1-4757-1904-8>.

KATSAROS, K. B. Sensors for mean meteorology. Academic Press, 2001.

KRISTENSEN, L. The perennial cup anemometer. **Wind Energy**, Wiley, v. 2, n. 1, p. 59–75, jan. 1999. ISSN 1099-1824. Available from Internet: <http://dx.doi.org/10.1002/(SICI)1099-1824(199901/03)2:1<59::AID-WE18>3.0.CO;2-R>.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation forest. In: **2008 Eighth IEEE International Conference on Data Mining**. IEEE, 2008. Available from Internet: <http://dx.doi.org/10.1109/ICDM.2008.17>.

LIU, J. et al. **MTAD: Tools and Benchmarks for Multivariate Time Series Anomaly Detection**. 2022. <https://github.com/OpsPAI/MTAD>.

LOTFI, M. Atmospheric wind flow distortion effects of meteorological masts. Unpublished, 2015. Available from Internet: <http://rgdoi.net/10.13140/RG.2.2.14794.62409>.

MAKKONEN, L.; LEHTONEN, P.; HELLE, L. Anemometry in icing conditions. **Journal of Atmospheric and Oceanic Technology**, American Meteorological Society, v. 18, n. 9, p. 1457–1469, sep. 2001. ISSN 1520-0426. Available from Internet: <http://dx.doi.org/10.1175/1520-0426(2001)018<1457:AIIC>2.0.CO;2>.

MALHOTRA, P. et al. Long short term memory networks for anomaly detection in time series. In: **Esann**. [S.l.: s.n.], 2015. v. 2015, p. 89.

MORTENSEN, N. Wind measurements for wind energy applications. a review. In: ELLIOT, G. (Ed.). **Wind energy conversion 1994**. [S.l.]: Mechanical Engineering Publications Limited, 1994. p. 353–360. 16th British Wind Energy Association Conference ; Conference date: 15-06-1994 Through 17-06-1994.

NFAOUI, H. Wind energy potential. In: **Comprehensive Renewable Energy**. Elsevier, 2012. p. 73–92. Available from Internet: <https://doi.org/10.1016/b978-0-08-087872-0.00204-3>.

NIK, V. M.; PERERA, A. The importance of developing climate-resilient pathways for energy transition and climate change adaptation. **One Earth**, Elsevier BV, v. 3, n. 4, p. 423–424, oct. 2020. Available from Internet: <https://doi.org/10.1016/j.oneear.2020.09.013>.

NOAA. **Global Atmospheric Circulations**. 2023. Available from Internet: <https://www.noaa.gov/jetstream/global/global-atmospheric-circulations>.

PALUTIKOF, J.; GUO, X.; HALLIDAY, J. The reconstruction of long wind speed records in the uk. In: **Proceedings of the Thirteenth British Wind Energy Association Conference**. [S.l.: s.n.], 1991. p. 275–280.

PANG, G. et al. Deep learning for anomaly detection. **ACM Computing Surveys**, Association for Computing Machinery (ACM), v. 54, n. 2, p. 1–38, mar. 2021. Available from Internet: <https://doi.org/10.1145/3439950>.

RODRIGUEZ, S. **Applied Computational Fluid Dynamics and Turbulence Modeling**. Springer International Publishing, 2019. Available from Internet: <https://doi.org/10.1007/978-3-030-28691-0>.

SHYU, M.-L. et al. A novel anomaly detection scheme based on principal component classifier. In: . [S.l.: s.n.], 2003.

STAUDEMEYER, R. C.; MORRIS, E. R. **Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks**. arXiv, 2019. Available from Internet: <https://arxiv.org/abs/1909.09586>.

SU, Y. et al. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In: **Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining**. ACM, 2019. (KDD '19). Available from Internet: <http://dx.doi.org/10.1145/3292500.3330672>.

SUCEVIC, N.; DJURISIC, Z. Influence of atmospheric stability variation on uncertainties of wind farm production estimation. In: . [S.l.: s.n.], 2012.

SUN, W.-Y.; SUN, O. M. Bernoulli equation and flow over a mountain. **Geoscience Letters**, Springer Science and Business Media LLC, v. 2, n. 1, jun. 2015. Available from Internet: <https://doi.org/10.1186/s40562-015-0024-1>.

TULI, S.; CASALE, G.; JENNINGS, N. R. **TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data**. arXiv, 2022. Available from Internet: <https://arxiv.org/abs/2201.07284>.

UNFCCC, U. N. F. C. on C. C. **Acordo de Paris**. 2015. Available from Internet: <https://unfccc.int/process-and-meetings/the-paris-agreement>.

VEERS, P. et al. Grand challenges in the science of wind energy. **Science**, v. 366, n. 6464, p. eaau2027, 2019. Available from Internet: <https://www.science.org/doi/abs/10.1126/science.aau2027>.

ZHANG, C. et al. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. **Proceedings of the AAAI Conference on Artificial Intelligence**, Association for the Advancement of Artificial Intelligence (AAAI), v. 33, n. 01, p. 1409–1416, jul. 2019. ISSN 2159-5399. Available from Internet: <http://dx.doi.org/10.1609/aaai.v33i01.33011409>.